Studies in Big Data 9

Aboul Ella Hassanien · Ahmad Taher Azar
Vaclav Snasel · Janusz Kacprzyk
Jemal H. Abawajy *Editors*

Big Data in Complex Systems

Challenges and Opportunities

Springer

Studies in Big Data

Volume 9

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence incl. neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/11970

Aboul Ella Hassanien · Ahmad Taher Azar
Vaclav Snasel · Janusz Kacprzyk
Jemal H. Abawajy
Editors

Big Data in Complex Systems

Challenges and Opportunities

Springer

Editors
Aboul Ella Hassanien
Cairo University
Cairo
Egypt

Ahmad Taher Azar
Faculty of Computers and Information
Benha University
Benha
Egypt

Vaclav Snasel
Faculty of Elec. Eng. & Comp. Sci.
Department of Computer Science
VSB-Technical University of Ostrava
Ostrava-Poruba
Czech Republic

Janusz Kacprzyk
Polish Academy of Sciences
Warsaw
Poland

Jemal H. Abawajy
School of Information Technology
Deakin University
Victoria
Australia

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-319-11055-4 ISBN 978-3-319-11056-1 (eBook)
DOI 10.1007/978-3-319-11056-1

Library of Congress Control Number: 2014949168

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Big data refers to large and complex massive amounts of data sets that it becomes difficult to process and analyze using traditional data processing technology. Over the past few years there has been an exponential growth in the rate of available data sets obtained from complex systems, ranging from the interconnection of millions of users in social media data, cheminformatics, hydroinformatics to the information contained in the complex biological data sets. This taking and opened new challenges and opportunities to researcher and scientists on how to acquisition, Recording, store and manipulate this huge amount of data sets and how to develop new tools, mining, study, and visualize the massive amount data sets and what insight can we learn from systems that were previously not understood due to the lack of information. All these aspect, coming from multiple disciples under the theme of big data and their features.

The ultimate objectives of this volume are to provide challenges and Opportunities to the research communities with an updated, in-depth material on the application of Big data in complex systems in order to finding solutions to the challenges and problems facing big data sets applications. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business.

The material of this book can be useful to advanced undergraduate and graduate students. Also, researchers and practitioners in the field of big data may benefit from it. Each chapter in the book openes with a chapter abstract and key terms list. The material is organized into seventeen chapters. These chapters are organized

along the lines of problem description, related works, and analysis of the results. Comparisons are provided whenever feasible. Each chapter ends with a conclusion and a list of references which is by no means exhaustive.

As the editors, we hope that the chapters in this book will stimulate further research in the field of big data. We hope that this book, covering so many different aspects, will be of value for all readers.

The contents of this book are derived from the works of many great scientists, scholars, and researchers, all of whom are deeply appreciated. We would like to thank the reviewers for their valuable comments and suggestions, which contribute to enriching this book. Special thanks go to our publisher, Springer, especially for the tireless work of the series editor of Big data sets sereies, Dr. Thomas Ditzinger.

December 2014

Aboul Ella Hassanien, SRGE, Egypt
Ahmad Taher Azar, Egypt
Vaclav Snasel, Czech Republic
Janusz Kacprzyk, Poland
Jemal H. Abawajy, Australia

Contents

Cloud Computing Infrastructure for Massive Data: A Gigantic Task Ahead

Renu Vashist

Abstract. Today, in the era of computer we collect and store data from innumerable sources and some of these are Internet transactions, social media, mobile devices and automated sensors. From all of these sources massive or big data is generated and gathered for finding the useful patterns. The amount of data is growing at the enormous rate, the analyst forecast that the expected global big data storage to grow at the rate of 31.87% over the period 2012-2016, thus the storage must be highly scalable as well as flexible so the entire system doesn't need to be brought down to increase storage. In order to store and access the massive data the storage hardware and network infrastructure is required.

Cloud computing can be viewed as one of the most viable technology for handling the big data and providing the infrastructure as services and these services should be uninterrupted. This computing is one of the cost effective technique for storage and analysis of big data.

Cloud computing and Massive data are the two rapidly evolving technologies in the modern day business applications. Lot of hope and optimism are surrounding around these technologies because analysis of massive or big data provides better insight into the data that may create competitive advantage and generates data related innovations having tremendous potential to revive the business bottom lines. Tradition ICT (information and communication) technology is inadequate and ill-equipped to handle terabytes or petabytes of data whereas cloud computing promises to hold unlimited, on-demand, elastic computing and data storage resources without huge upfront investments that is otherwise required when setting up traditional data centers. These two technologies are on converging paths and the combinations of the two technologies are proving powerful when it comes to perform analytics. At the same time, cloud computing platforms provide massive

Renu Vashist
Faculty of Computer Science,
Shri Mata Vaishno Devi University Katra, (J & K), India
e-mail: vashist.renu@gmail.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_1

scalability, 99.999% reliability, high performance, and specifiable configurability. These capabilities are provided at relatively low cost compared to dedicated infrastructures.

There is an element of over enthusiasm and unrealistic expectations with regard to the use and future of these technologies. This chapter draws attention towards the challenges and risks involved in the use and implementation of these naive technologies. Downtime, data privacy and security, scarcity of big data analysts, validity and accuracy of the emerged data pattern and many more such issues need to be carefully examined before switching from legacy data storage infrastructure to the cloud storage. The chapter elucidates the possible tradeoffs between storing the data using legacy infrastructure and the cloud. It is emphasizes that cautious and selective use of big data and cloud technologies is advisable till these technologies matures.

Keywords: Cloud Computing, Big Data, Storage Infrastructure, Downtime.

1 Introduction

The growing demands of today's business, government, defense, surveillance agencies, aerospace, research, development and entertainment sector has generated multitude of data. Intensified business competition and never ending customer's demands have pushed the frontiers of technological innovations to the new boundaries. Expanded realm of new technologies has generated the big data on the one hand and cloud computing on the other. Data over the size of terabytes or petabytes is referred to as big data. Traditional storage infrastructure is not capable of storing and analyzing such massive data. Cloud computing can be viewed as one of the most viable technology that is available to us for handling big data. The data generated through social media sites such as Facebook, Twitter and YouTube are unstructured or big data. Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis (Villars et al., 2011). Big data has big potential and many useful patterns may be found by processing this data which may help in enhancing various business benefits. The challenges associated with big data are also big like volume (Terabytes, Exabytes), variety (Structured, Unstructured) velocity (continuously changing) and validity of data i.e. the pattern found by the analysis of data can be trusted or not (Singh, 2012). Data are no longer restricted to structured database records but include unstructured data having no standard formatting (Coronel et al., 2013).

Cloud computing refers to the delivery of computing services on demand through internet like any other utility services such as telephony, electricity, water and gas supply (Agrawal, 2010). Consumers of the utility services in turn have to pay according to their usage. Likewise this computing is on-demand network access to computing resources which are often provided by an outside entity and require little management effort by the business (IOS Press, 2011). Cloud

computing is emerging in the mainstream as a powerful and important force of change in the way that the information can be managed and consumed to provide services (Prince, 2011).

The big data technology mainly deals with three major issues which are storage, processing and cost associated with it. Cloud computing may be one of the most efficient solution that is cost effective for storing big data and at the same time providing the scalability and flexibility. The two cloud services that is Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) has ability to store and analyze more data at lower costs. The major advantage of PaaS is that it gives companies a very flexible way to increase or decrease storage capacity as needed by their business. IaaS technology ups processing capabilities by rapidly deploying additional computing nodes. This kind of flexibility allow resources to be deployed rapidly as needed, cloud computing puts big data within the reach of companies that could never afford the high costs associated with buying sufficient hardware capacity to store and analyze large data sets (Ahuja and Moore, 2013 a).

In this chapter, we examine the issues in big data analysis in cloud computing. The chapter is organized as followed: Section 2 reviews related work, Section 3 gives the overview of cloud computing, Section 4 describes the big data, Section 5 describe cloud computing and big data as compelling combination, Section 6 provides challenges and obstacle in handling big Data using cloud computing. Section 7 describes the discussions and Section 8 concludes the chapter.

2 Related Work

Two of the hottest IT trends today are the move to cloud computing and the emergence of big data as a key initiative for leveraging information. For some enterprises, both of these trends are converging, as they try to manage and analyze big data in their cloud deployments. Various researches with respect to the interaction between big data and cloud suggest that the dominant sentiment among developers is that big data is a natural component of the cloud (Han, 2012). Companies are increasingly using cloud deployments to address big data and analytics needs. Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds (Intel, 2013).

(Chadwick and Fatema, 2012) provides a policy based authorization infrastructure that a cloud provider can run as an infrastructure service for its users. It will protect the privacy of user's data by allowing the users to set their own privacy policies, and then enforcing them so that no unauthorized access is allowed to their data.

(Fernado et al., 2013) provides an extensive survey of mobile cloud computing research, while highlighting the specific concerns in mobile cloud computing.

(Basmadjian et al., 2012) study the case of private cloud computing environments from the perspective of energy saving incentives. The proposed approach can also be applied to any computing style.

Big data analysis can also be described as knowledge discovery from data (Sims, 2009). Knowledge discovery is a method where new knowledge is derived from a data set. More accurately, knowledge discovery is a process where different practices of managing and analyzing data are used to extract this new knowledge (Begoli, 2012). For considering big data, techniques and approaches used for storing and analysis of data needs to be reevaluated. Legacy infrastructures do not support massive data due to the inability to compute big data and scalability it requires. Other challenges associated with big data are presence of structured, unstructured data and variety of data. One approach to this problem is addressed by NoSQL databases. NoSQL databases are characteristically non-relational and typically do not provide SQL for data manipulation. NoSQL describes a class of databases that include: graph, document and key-value stores. The NoSQL database are designed with the aim to provide high scalability (Ahuja and Mani, 2013; Grolinger et al., 2013). Further a new class of database known as NewSQL databases has developed that follow the relational model but either distributes the data or transaction processing across nodes in a cluster to achieve comparable scalability (Pokorny, 2011). There are several factors that need to be looked around before switching to cloud for big data management. The two major issues are security and privacy of data that resides in the cloud (Agrawal, 2012). Storing big data using cloud computing provide flexibility, scalability and cost effective but even in cloud computing, big data analysis is not without its problems. Careful consideration must be given to the cloud architecture and the techniques for distributing these data intensive tasks across the cloud (Ji, 2012).

3 Overview of Cloud Computing

The symbolic representation of internet in the form of cloud in network diagram can be seen as the emergence of word ‘cloud’. Cloud means internet and cloud computing means services provided through internet. Every body across the globe is talking about this technology but till date it doesn’t have unanimous definition, terminology, concepts and much more clarification is needed over that. Two major institutions have significantly contributed in clearing the fog, National Institute of standards and technology and cloud security alliance. They both agree to a definition of cloud that “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”. (NIST, 2009).Cloud computing refers to the use of computers which access Internet locations for computing power, storage and applications, with no need for the individual access points to maintain any of the infrastructure. Examples of cloud services include online file storage, social networking sites,

webmail, and online business applications. Cloud computing is positioning itself as a new emerging platform for delivering information infrastructures and resources as IT services. Customers (enterprises or individuals) can then provision and deploy these services in a pay-as-you-go fashion and in a convenient way while saving huge capital investment in their own IT infrastructures (Chen and Wang, 2011). Due to the vast diversity in the available Cloud services, from the customer's point of view, it has become difficult to decide whose services they should use and what is the basis for their selection (Garg et al., 2013). Given the number of cloud services that are now available across different cloud providers, issues relating to the costs of individual services and resources besides ranking these services come to the fore (Fox, 2013).

The cloud computing model allows access to information and computer resources from anywhere, anytime where a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. Despite increasing usage of mobile computing, exploiting its full potential is difficult due to its inherent problems such as resource scarcity, frequent disconnections, and mobility (Fernado et al., 2013).

This cloud model is composed of four essential characteristics**,** three service models, and four deployment models.

3.1 Essential Characteristics of Cloud Computing

Cloud computing has a variety of characteristics, among which the most useful ones are: (Dialogic, 2010)

- **Shared Infrastructure.** A virtualized software model is used which enables the sharing of physical services, storage, and networking capabilities. Regardless of deployment model whether it be a public cloud or private cloud the cloud infrastructure is shared across a number of users.
- **Dynamic Provisioning.** According to the current demand requirement automatic services are provided. This is done automatically using software automation, enabling the expansion and contraction of service capability, as needed. This dynamic scaling needs to be done while maintaining high levels of reliability and security.
- **Network Access.** Capabilities are available over the network and a continuous internet connection is required for a broad range of devices such as PCs, laptops, and mobile devices, using standards-based APIs (for example, ones based on HTTP). Deployments of services in the cloud include everything from using business applications to the latest application on the newest smart phones.
- **Managed Metering.** Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service. Uses metering for managing and optimizing the service and to provide reporting and billing information. In this way, consumers are billed

for services according to how much they have actually used during the billing period. In short, cloud computing allows for the sharing and scalable deployment of services, as needed, from almost any location, and for which the customer can be billed based on actual usage.

In short, cloud computing allows for the sharing and scalable deployment of services, as needed, from almost any location, and for which the customer can be billed based on actual usage.

3.2 Service Models

After establishing the cloud the services are provided based on the business requirement. The cloud computing service models are Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Storage as a service. In Software as a Service model, a pre-made application, along with any required software, operating system, hardware, and network are provided. In PaaS, an operating system, hardware, and network are provided, and the customer installs or develops its own software and applications. The IaaS model provides just the hardware and network; the customer installs or develops its own operating systems, software and applications.

Software as a Service (SaaS): Software as a service provides businesses with applications that are stored and run on virtual servers in the cloud (Cole, 2012). A SaaS provider provides the consumer the access to application and resources. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. In this type of cloud services the customer has the least control over the cloud.

Platform as a Service (PaaS): The PaaS services are one level above the SaaS services. There are a wide number of alternatives for businesses using the cloud for PaaS (Géczy et al., 2012). The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. Other advantages of using PaaS include lowering risks by using pretested technologies, promoting shared services, improving software security, and lowering skill requirements needed for new systems development (Jackson, 2012).

Infrastructure as a Service (IaaS): is a cloud computing model based on the principle that the entire infrastructure is deployed in an on-demand model.

This almost always takes the form of a virtualized infrastructure and infrastructure services that enables the customer to deploy virtual machines as components that are managed through a console. The physical resources such as servers, storage, and network are maintained by the cloud provider while the infrastructure deployed on top of those components is managed by the user. It is important to mention here that the user of IaaS is always a team comprised of several IT experts in the required infrastructure components. The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications and possibly limited control of select networking components (e.g., host firewalls).

IaaS is often considered utility computing because it treats compute resources much like utilities (such as electricity, telephony) are treated. When the demand for capacity increases, more computing resources are provided by the provider (Rouse, 2010). As demand for capacity decreases, the amount of computing resources available decreases appropriately. This enables the "on-demand" as well as the "pay-per-use" properties of cloud architecture. Infrastructure as a service is the cloud computing model receiving the most attention from the market, with an expectation of 25% of enterprises planning to adopt a service provider for IaaS (Ahuja and Mani , 2012), 2009).Fig 1 provide the overview of cloud computing and the three service models.

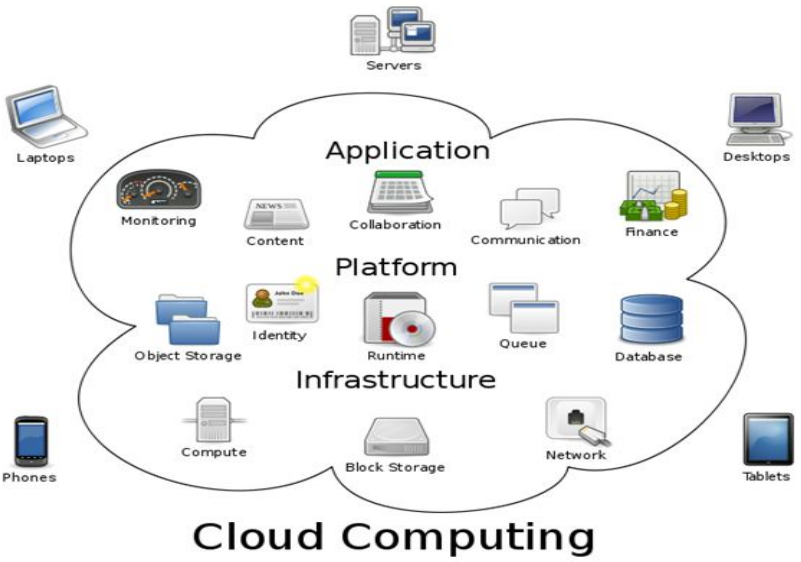

Fig. 1 Overview of Cloud Computing (Source: Created by Sam Johnston Wikimedia Commons)

Storage as a Service: These services are commonly known as StaaS it facilitates cloud applications to scale beyond their limited servers. StaaS allows users to store their data at remote disks and access them anytime from any place using internet. Cloud storage systems are expected to meet several rigorous requirements for maintaining users' data and information, including high availability, reliability, performance, replication and data consistency; but because of the conflicting nature of these requirements, no one system implements all of them together.

3.3 Deployment Models

Deploying cloud computing can differ depending on requirements, and the following four deployment models have been identified, each with specific characteristics that support the needs of the services and users of the clouds in particular ways

Private Cloud
The Private clouds are basically owned by the single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. The private cloud is a pool of computing resources delivered as a standardized set of services that are specified, architected, and controlled by a particular enterprise. The path to a private cloud is often driven by the need to maintain control of the service delivery environment because of application maturity, performance requirements, industry or government regulatory controls, or business differentiation reasons (Chadwick et al., 2013). Functionalities are not directly exposed to the customer it is similar to the SaaS from customer point of view. Example eBay.

For example, banks and governments have data security issues that may preclude the use of currently available public cloud services. Private cloud options include:

- **Self-hosted Private Cloud:** A Self-hosted Private Cloud provides the benefit of architectural and operational control, utilizes the existing investment in people and equipment, and provides a dedicated on-premise environment that is internally designed, hosted, and managed
- **Hosted Private Cloud:** A Hosted Private Cloud is a dedicated environment that is internally designed, externally hosted, and externally managed. It blends the benefits of controlling the service and architectural design with the benefits of datacenter outsourcing.
- **Private Cloud Appliance:** A Private Cloud Appliance is a dedicated environment that procured from a vendor is designed by that vendor with provider/market driven features and architectural control, is internally hosted, and externally or internally managed. It blends the benefits of using predefined functional architecture, lower deployment risk with the benefits of internal security and control.

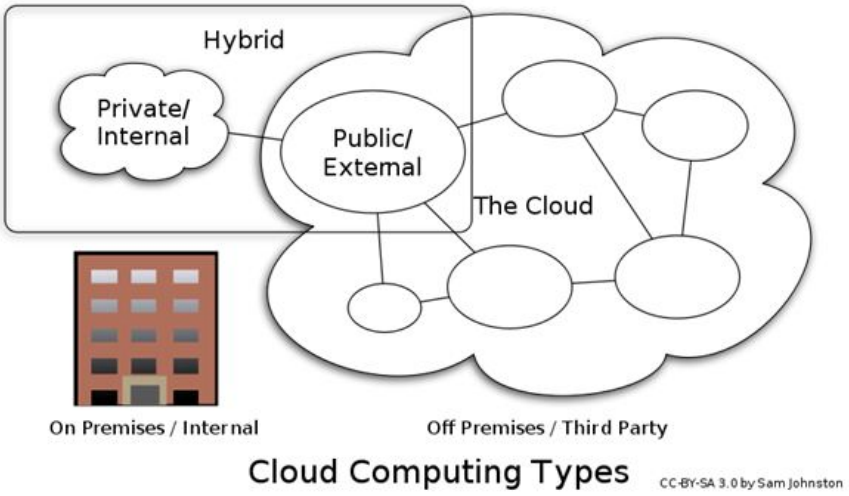

Fig. 2 Private, Public and Hybrid Cloud Computing

Public Cloud

The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider. The Public Cloud is a pool of computing services delivered over the Internet. It is offered by a vendor, who typically uses a "pay as you go" or "metered service" model (Armbrust et al., 2010). Public Cloud Computing has the following potential advantages: you only pay for resources you consume; you gain agility through quick deployment; there is rapid capacity scaling; and all services are delivered with consistent availability, resiliency, security, and manageability. A public cloud is considered to be an external cloud (Aslam et al., 2010). Example Amazon, Google Apps.

Public Cloud options include:

- **Shared Public Cloud:** The Shared Public Cloud provides the benefit of rapid implementation, massive scalability, and low cost of entry. It is delivered in a shared physical infrastructure where the architecture, customization, and degree of security are designed and managed by the provider according to market-driven specifications
- **Dedicated Public Cloud:** The Dedicated Public Cloud provides functionality similar to a Shared Public Cloud except that it is delivered on a dedicated physical infrastructure. Security, performance, and sometimes customization are better in the Dedicated Public Cloud than in the Shared Public Cloud. Its architecture and service levels are defined by the provider and the cost may be higher than that of the Shared Public Cloud, depending on the volume.

Community Cloud
If several organizations have similar requirements and seek to share infrastructure to realize the benefits of cloud computing, then a community cloud can be established. This is a more expensive option as compared to public cloud as the costs are spread over fewer users as compared to a public cloud. However, this option may offer a higher level of privacy, security and/or policy compliance.

Hybrid Cloud
The Hybrid cloud consist of a mixed employment of private and public cloud infrastructures so as to achieve a maximum of cost reduction through outsourcing while maintaining the desired degree of control over e.g. sensitive data by employing local private clouds. There are not many hybrid clouds actually in use today, though initial initiatives such as the one by IBM and Juniper already introduce base technologies for their realization (Aslam et al., 2010).

Some users may only be interested in cloud computing if they can create a private cloud which if shared at all, is only between locations for a company or corporation. Some groups feel the idea of cloud computing is just too insecure. In particular, financial institutions and large corporations do not want to relinquish control to the cloud, because they don't believe there are enough safeguards to protect information. Private clouds don't share the elasticity and, often, there is multiple site redundancy found in the public cloud. As an adjunct to a hybrid cloud, they allow privacy and security of information, while still saving on infrastructure with the utilization of the public cloud, but information moved between the two could still be compromised.

3.4 Cloud Storage Infrastructure

Effortless data storage "in the cloud" is gaining popularity for personal, enterprise and institutional data backups and synchronization as well as for highly scalable access from software applications running on attached compute servers (Spillner et al., 2013). Cloud storage infrastructure is a combination of both hardware equipments like servers, routers, and computer network and software component such as operating system and virtualization softwares. However, when compared to a traditional or legacy storage infrastructure it differs in terms of accessibility of files which under the cloud model is accessed through network which is usually built on an object-based storage platform. Access to object-based storage is done through a Web services application programming interface (API) based on the Simple Object Access Protocol (SOAP). An organization must ensure some essential necessities such as secure multi-tenancy, autonomic computing, storage efficiency, scalability, a utility computing chargeback system, and integrated data protection before embarking on cloud storage.

Data storage is one of the major use of cloud computing. With the help of cloud storage organizations may control their rising storage cost. In tradition storage the data is stored on dedicated servers whereas in cloud storage, data is stored on multiple third-party servers. The user sees a virtual server when data is stored and it

appears to the user as if the data is stored in a particular place with a specific name but that place doesn't exist in reality. It's just a virtual space which is created out of the cloud and the user data is stored on any one or more of the computers used to create the cloud. The actual storage location is changing frequently from day to day or even minute to minute, because the cloud dynamically manages available storage space using specific algorithms. Even though the location is virtual, but the user feels it as a "static" location and can manage his storage space as if it were connected to his own PC. Cost and security are the two main advantages associated with cloud storage. The cost advantage in the cloud system is achieved through economies of scale by means of large scale sharing of few virtual resources rather than dedicated resources connected to personal computer. Cloud storage gives due weightage to security aspect also as multiple data back ups at multiple locations eliminates the danger of accidental data erosion or hardware crashes. Since multiple copies of data are stored at multiple machines if one machine goes offline or crashes, data is still available to the user through other machines.

It is not beneficial for some small organizations to maintain an in-house cloud storage infrastructure due to the cost involved in it. Such organization, can contract with a cloud storage service provider for the equipment used to support cloud operations. This model is known as Infrastructure-as-a-Service (IaaS), where the service provider owns the equipment (storage, hardware, servers and networking components) and the client typically pays on a per-use basis. The selection of the appropriate cloud deployment model whether it be public, private and hybrid depend on the requirement of the user and the key to success is creating an appropriate server, network and storage infrastructure in which all resources can be efficiently utilized and shared. Because all data reside on same storage systems, data storage becomes even more crucial in a shared infrastructure model. Business needs driving the adoption of cloud technology typically include (NetApp, 2009):

- Pay as you use
- Always on
- Data security and privacy
- Self service
- Instant deliver and capacity elasticity

These businesses needs translate directly to the following infrastructure requirements

- Secure multi-tenancy
- Service automation and management
- Data mobility
- Storage efficiency
- Integrated data protection.

3.5 Cloud Storage Infrastructure Requirements

The data is growing at the immense rate and the combination of technology trends such as virtualization with the increased economic pressures, exploding growth of unstructured data and regulatory environments that are requiring enterprises to keep data for longer periods of time, it is easy to see the need for a trustworthy and appropriate storage infrastructure. Storage infrastructure is the backbone of every business. Whether a cloud is public or private, the key to success is creating a storage infrastructure in which all resources can be efficiently utilized and shared. Because all data resides on the storage systems, data storage becomes even more crucial in a shared infrastructure model (Promise, 2010). Most important cloud infrastructure requirement are as follows

1) Elasticity: Cloud storage must be elastic so that it can quickly adjust with underlying infrastructure according to changing requirement of the customer demands and comply with service level agreements.

2) Automatic: Cloud storage must have the ability to be automated so that policies can be leveraged to make underlying infrastructure changes such as placing user and content management in different storage tiers and geographic locations quickly and without human intervention.

3) Scalability: Cloud storage needs to scale quickly up and down according to the requirement of customer. This is one of the most important requirements that make cloud so popular.

4) Data Security: Security is one of the major concerns of the cloud users. As different users store more of their own data in a cloud, they want to ensure that their private data is not accessible to other users who are not authorized to see it. If this is the case than the user can have a private clouds because security is assumed to be tightly controlled in case of private cloud. But in case of public clouds, data should either be stored on a partition of a shared storage system, or cloud storage providers must establish multi-tenancy policies to allow multiple business units or separate companies to securely share the same storage hardware.

5) Performance: Cloud storage infrastructure must provide fast and robust data recovery as an essential element of a cloud service.

6) Reliability: As more and more users are depending on the services offered by a cloud, reliability becomes increasingly important. Various users of the cloud storage want to make sure that their data is reliably backed up for disaster recovery purposes and cloud should be able to continue to run in the presence of hardware and software failures.

7) Operational Efficiency: Operational efficiency is a key to successful business enterprise which can be ensured by better management of storage capacities and cost benefit. Both these features should be the integral part of the cloud storage.

8) Data Retrieval: Once the data is stored on the cloud it can be easily accessed from anywhere at anytime where the network connection is available. Ease of access to data in the cloud is critical in enabling seamless integration of cloud storage into existing enterprise workflows and to minimize the learning curve for cloud storage adoption.

9) Latency: Cloud storage model are not suitable for all applications especially for real time applications. It is important to measure and test network latency before committing to a migration. Virtual machines can introduce additional latency through the time-sharing nature of the underlying hardware and unanticipated sharing and reallocation of machines can significantly affect run times.

Storage is the most important component of IT Infrastructure. Unfortunately, it is almost always managed as a scarce resource because it is relatively expensive and the consequences of running out of storage capacity can be severe. Nobody wants to take the responsibility of storage manager thus the storage management suffers from slow provisioning practices.

4 Big Data or Massive Data

Big data or Massive data is emerging as a new keyword in all businesses from last one year or so. Big data is a term that can be applied to some very specific characteristics in terms of scale and analysis of data. Big Data (juniper networks, 2012) refers to the collection and subsequent analysis of any significantly large collection of unstructured data (data over the petabyte) that may contain hidden insights or intelligence. Data are no longer restricted to structured database records but include unstructured data that is data having no standard formatting (Coronel et al., 2013). When analyzed properly, big data can deliver new business insights, open new markets, and create competitive advantages. According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it" (Edd Dumbill, 2012).

Big data is on one hand very large amount of unstructured data while on the other hand is dependent on rapid analytics, whose answer needs to be provided in seconds. Big Data requires huge amounts of storage space. While the price of storage continued to decline, the resources needed to leverage big data can still pose financial difficulties for small to medium sized businesses. A typical big data storage and analysis infrastructure will be based on clustered network-attached storage (Oracle, 2012).

The data is growing at the enormous rate and the growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 exabytes of data were created and stored in 2005. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015. The expected growth of data in 2020 by IBM is 35,000 exabytes. (IBM, IDC, 2013). The Data growth over year is shown in fig 3.

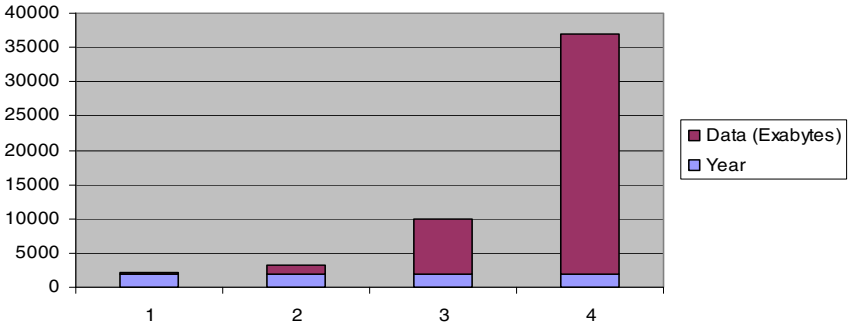

Fig. 3 Data growth over years

4.1 Characteristics of Big Data

Big data consist of traditional enterprise data, Machine data and social data. Examples of which are Facebook, Google or Amazon, which analyze user status. These datasets are large because the data is no longer traditional structured data, but data from many new sources, including e-mail, social media, and Internet-accessible sensors (Manyika et al., 2011) The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44 times between 2009 and 2020. But while it's often the most visible parameter, volume of data is not the only characteristic that matters. In fact, there are five key characteristics that define big data are **volume**, **velocity, variety**, **value and veracity**. These are known as the five V's of massive data (Yuri, 2013). The three major attribute of the data are shown in fig 4.

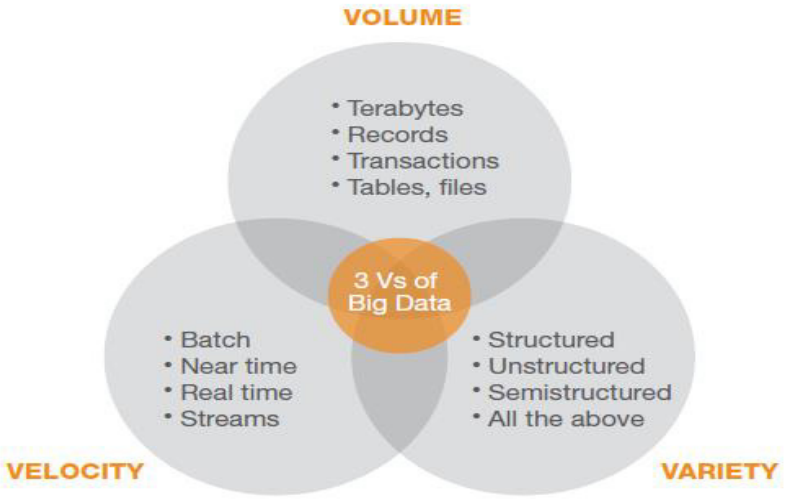

Fig. 4 Three V's of Big Data

Volume is used to define the data but the volume of data is a relative term. Small and medium size organizations refer gigabytes or terabytes of data as Big Data whereas big global enterprises consider petabytes and exabytes as big data. Most of the companies now days are storing the data, which may be medical data, financial Market data, social media data or any other kind of data. Organizations which have gigabytes of data today may have exabytes of data in near future. Since data is collected from **variety** of sources such as Biological and medical, facial research, Human psychology and behavior research and History, archeology and artifact. Due to variety of sources this data may be structured, unstructured and semi structured or combination of these. The **velocity** of the data means how frequently the data arrives and is stored, and how quickly it can be retrieved. The term velocity refers to the data in motion the speed at which the data is moving. Data such as financial market, movies, and ad agencies should travel very fast for proper rendering. Various aspects of big data are shown in fig 5.

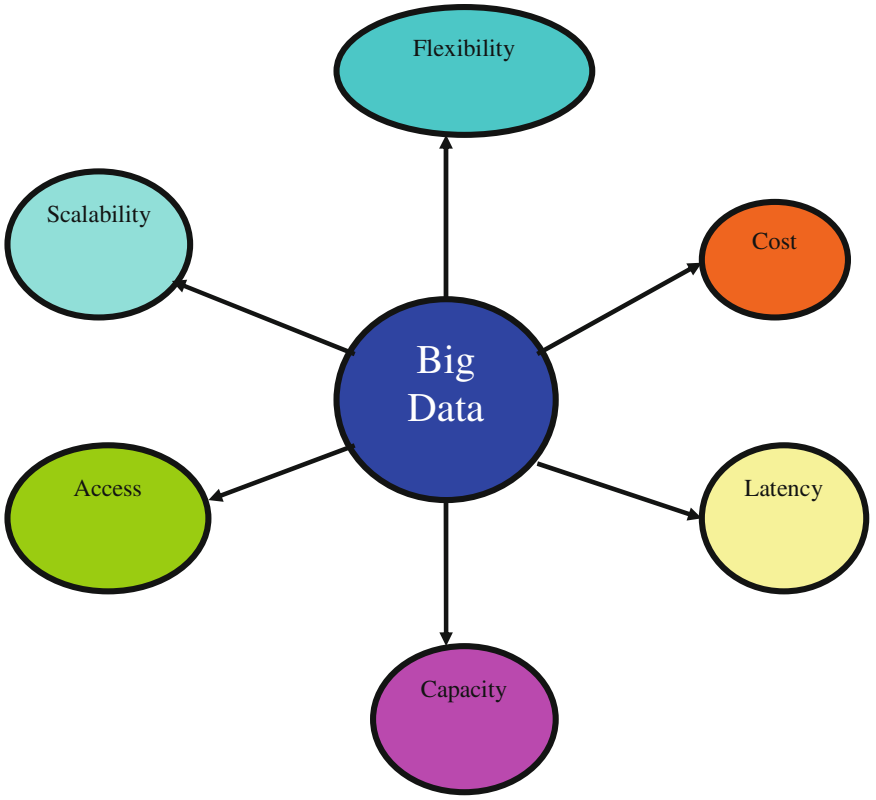

Fig. 5 Various aspect of Big Data

4.2 Massive Data has Major Impact on Infrastructure

A highly scalable infrastructure is required for handling big data unlike large data sets that have historically been stored and analyzed, often through data warehousing, big data is made up of discretely small, incremental data elements with real-time additions or modifications. It does not work well in traditional, online transaction processing (OLTP) data stores or with traditional SQL analysis tools. Big data requires a flat, horizontally scalable database, often with unique query tools that work in real time with actual data. Table 1 compares big data with traditional data.

Table 1 Comparison of big data with traditional data

Parameters	**Traditional Data**	**Big Data**
Type of Data	Structured	Unstructured
Volume of Data	Terabytes	Petabytes and Exabytes
Architecture	Centralized	Distributed
Relationship between Data	Known	Complex

For handling the new high-volume, high-velocity, high-variety sources of data and to integrate them with the pre-existing enterprise data organizations must evolve their infrastructures accordingly for analyzing big data. When big data is distilled and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation all of which can have a significant impact on the bottom line (Oracle, 2013). Analyzing big data is done using a programming paradigm called MapReduce (Eaton, et al., 2012). In the MapReduce paradigm, a query is made and data are mapped to find key values considered to relate to the query; the results are then reduced to a dataset answering the query (Zhang, et al., 2012).

The data is growing at enormous rate and traditional file system can't support big data. For handling big data the storage must be highly scalable and flexible so the entire system doesn't need to be brought down to increase storage. Institution must provide a proper infrastructure for handling five v's of massive data. For implementation of big data the primary requirement are software component and hardware component among which hardware refers to for infrastructure and analytics. Big data infrastructure components are Hadoop (Hadoop Project, 2009; Dai, 2013) and cloud computing infrastructure services for data centric applications. Hadoop is the big data management software infrastructure used to distribute, catalog, manage, and query data across multiple, horizontally scaled server nodes. This is a framework for processing, storing, and analyzing massive amounts of distributed unstructured data. This Distributed File system was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop is an open source data management framework that has become widely deployed for massive parallel computation and distributed file systems in a cloud environment. The infrastructure is the foundation of big data technology

stack. Big data infrastructure includes management interfaces, actual servers (physical or virtual), storage facilities, networking, and possibly back up systems. Storage is the most important infrastructure requirement and storage systems are also becoming more flexible and are being designed in a scale-out fashion, enabling the scaling of system performance and capacity (Fairfield, 2014). A recent Data Center Knowledge report explained that big data has begun having such a far-reaching impact on infrastructure that it is guiding the development of broad infrastructure strategies in the network and other segments of the data center (Marciano, 2013). However, the clearest and most substantial impact is in storage, where big data is leading to new challenges in terms of both scale and performance.

These are main points about big data which must be noticed

- Big data, if not managed properly the sheer volume of unstructured data that's generated each year within an enterprise can be costly in terms of storage.
- It is not always easy to locate information from unstructured data.
- The underlying cost of the infrastructure to power the analysis has fallen dramatically, making it economic to mine the information.
- Big Data has the potential to provide new forms of competitive advantage for organizations.
- Using in-house servers for storing big data can be very costly.

At root the key requirement of big data storage are that it can handle very large amounts of data and keep scaling to keep up with growth, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools. The infrastructure needed to deal with high volumes of high velocity data coming from real-time systems needs to be set up so that the data can be processed and eventually understood. This is a challenging task because the data isn't simply coming from transactional systems; it can include tweets, Facebook updates, sensor data, music, video, WebPages etc. Finally, the definition of today's data might be different than tomorrow's data.

Big-Data infrastructure companies, such as Cloudera, HortonWorks, MapR, 10Gen, and Basho offer software and services to help corporations create the right environments for the storage, management, and analysis of their big data. This infrastructure is essential for deriving information from the vast data stores that are being collected today. Setting up the infrastructure used to be a difficult task, but these and related companies are providing the software and expertise to get things running relatively quickly.

4.3 The Impact of Big Data on Markets in Coming Years

As the Big Data technology matures and users begin to explore more strategic business benefits, the potential of Big Data's impact on data management and business analytics initiatives will grow significantly. According to IDC, the Big Data technology and service market was about US$4.8 billion in 2011(IDC, 2011)).

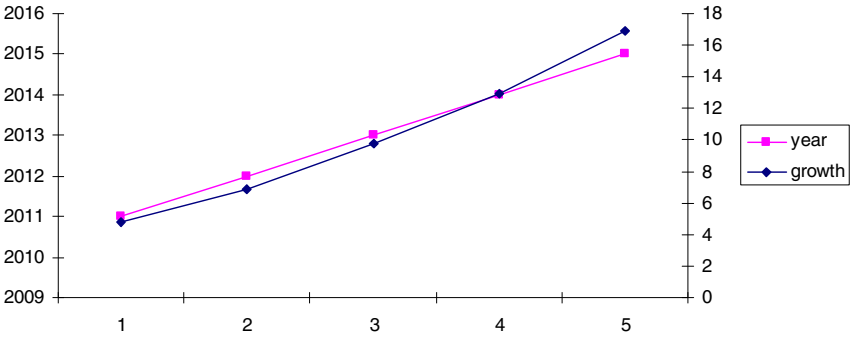

Fig. 6 Big Data Market Projection

The market is projected to grow at a compound annual growth rate (CAGR) of 37.2% between 2011 and 2015. By 2015, the market size is expected to be US$16.9 billion.

It is important to note that 42% of IT leaders have already invested in big data technology or plan to do so in the next 12 months. But irony is that most organizations have immature big data strategies. Businesses are becoming aware that big data initiatives are critical because they have identified obvious or potential business opportunities that cannot be met with traditional data sources and technologies. In addition, media hype is often backed with rousing use cases. By 2015, 20% of Global 1000 organizations will have established a strategic focus on "information infrastructure" equal to that of application management (Gartner Report, 2013).

5 Cloud Computing and Big Data: A Compelling Combination

From the last few years, cloud computing has been one of the most-talked about technology. But now a day big data is also coming on strong. Big Data refers to the tools, processes, and procedures that allow an organization to create, manipulate, and manage very large data sets and storage facilities (Knapp, 2013). By combining these two upcoming technologies we may get the opportunity to save money, improve end-user satisfaction and use more of your data to its fullest extent. This past January, National Institute of Standards and Technology (NIST, 2009) as well as other government agencies, industry, and academia, got together to discuss the critical intersection of big data and the cloud. Although government agencies have been slower to adopt new technologies in the past, the event underscored the fact that the public sector is leading and in some cases creating big data innovation and adoption. A recent survey conducted by GigaSpaces found that 80 percent of those IT executives who think big data processing is important

are considering moving their big data analytics to one or more cloud delivery models (Gardner, 2012).

Big Data and Cloud Computing are two technologies which are on converging paths and the combination of these two technologies are proving powerful when used to perform analytics and storing. It is no surprise that the rise of Big Data has coincided with the rapid adoption of Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) technologies. PaaS lets firms scale their capacity on demand and reduce costs while IaaS allows the rapid deployment of additional computing nodes when required. Together, additional compute and storage capacity can be added to almost instantaneously. The flexibility of cloud computing allows resources to be deployed as needed. As a result, firms avoid the tremendous expense of buying hardware capacity they'll need only occasionally. Cloud computing promises on demand, scalable, pay-as-you-go compute and storage capacity. Compared to an in-house datacenter, the cloud eliminates large upfront IT investments, lets businesses easily scale out infrastructure, while paying only for the capacity they use. It's no wonder cloud adoption is accelerating – the amount of data stored in Amazon Web Services (AWS) S3 cloud storage has jumped from 262 billion objects in 2010 to over 1 trillion objects at the end of the first second of 2012. Using cloud infrastructure to analyze big data makes sense because (Intel 2013).

Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure. Only large and midsized data centers have the in-house resources to support distributed computing models. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data in-house, while augmenting internal resources with public cloud services. This hybrid cloud option enables companies to use on-demand storage space and computing power via public cloud services for certain analytics initiatives (for example, short-term projects), and provide added capacity and scale as needed.

Big data may mix internal and external sources. Most of the enterprises often prefer to keep their sensitive data in-house, but the big data that companies owns may be stored externally using cloud. Some of the organizations are already using cloud technology and others are also switching to it. Sensitive data may be stored on private cloud and public cloud can be used for storing big data. Data can be analyzes externally from the public cloud or from private cloud depending on the requirement of enterprise.

Data services are needed to extract value from big data. For extracting the valid information from the data the focus should be on the analytics. It is also required that analytics is also provided as services supported by internal private cloud, a public cloud, or a hybrid model.

With the help of cloud computing scalable analytical solution may be found for big data. Cloud computing offer efficiency and flexibility for accessing data. Organizations can use cloud infrastructure depending on their requirements such as cost, security, scalability and data interoperability. A private cloud infrastructure is used to mitigate risk and to gain the control over data and public cloud

infrastructure is used to increase the scalability. A hybrid cloud infrastructure may be implemented to use the services and resources of both the private and the public cloud. By analyzing the big data using cloud based strategy the cost can be optimized. Major reasons of using cloud computing for big data implementation are hardware cost reduction and processing cost reduction.

5.1 Optimizing Current Infrastructure for Handling Big Data

For handling the volume, velocity, veracity and variety of big data the important component is the underlying infrastructure. Many business organizations are still dependent on legacy infrastructure for storing big data which are not capable for handling many real time operations. These firms need to replace the outdated legacy system and be more competitive and receptive to their own big data needs. In reality getting rid off legacy infrastructure is a very painful process. The time and expense required to handle such a process means the value of the switch must far outweigh the risks. Instead of totally removing the legacy infrastructure there is a need to optimize the current infrastructure.

For handling this issue many organizations have implemented software-as-a-service (SaaS) applications that are accessible via the Internet. With these type of solutions businesses can collect and store data remote service and without the need to worry about overloading their existing infrastructure. Open source software which allows companies to simply plug their algorithms and trading policies into the system, leaving it to handle their increasingly demanding processing and data analysis tasks can be used for addressing infrastructure concerns other than SaaS.

Today, however, more and more businesses believe that big data analysis is giving momentum to their business. Hence they are adopting SaaS and open source software solutions ultimately leaving their legacy infrastructure behind. A recent Data Center Knowledge report explained that big data has begun having such a far-reaching impact on infrastructure that it is guiding the development of broad infrastructure strategies in the network and other segments of the data center. However, the clearest and most substantial impact is in storage, where big data is leading to new challenges in terms of both scale and performance.

Cloud computing has become a viable, mainstream solution for data processing, storage and distribution, but moving large amounts of data in and out of the cloud presented an insurmountable challenge for organizations with terabytes of digital content.

6 Challenges and Obstacle in Handling Big Data Using Cloud Computing

Using Cloud computing for big data is a daunting task and continues to pose new challenges to those business organizations who decide to switch to cloud computing. Since the big data deals with the dataset measuring in tens of terabytes, therefore it has to rely on traditional means for moving big data to cloud as

moving big data to and fro from the cloud and moving the data within the cloud may compromise data security and confidentiality.

Managing big data using cloud computing is though cost effective, agile and scalable but involves some tradeoffs like possible downtime, data security, herd instinct syndrome, correct assessment of data collection, cost, validity of patterns. It's not an easy ride and there is a gigantic task ahead to store, process and analyze big data using cloud computing. Before moving to big data using cloud following points should be taken care of

- **Possible Downtime:** Internet is the backbone of cloud computing. If there is some problem in the backbone whole system shattered down immediately. For accessing your data you must have fast internet connection. Even with fast and reliable internet connection we have poor performance because of latency. For cloud computing just like video conferencing the requirement is as little latency as possible. Even with minimum latency there is possible downtime. If internet is down we can't access our data which is at cloud. The most reliable cloud computing service providers suffer server outages now and again. This could be a great loss to the enterprise in term of cost. At such times the in-house storage gives advantages.
- **Herd instinct syndrome:** The major problem related with the big data is that most of the organizations do not understand whether there is an actual need for big data or not. It is often seen that companies after companies are riding the bandwagon of 'Big data and cloud computing' without doing any homework. A minimum amount of preparation is required before switching to these new technologies because big data is getting bigger day by day and thereby necessitating a correct assessment regarding the volume and nature of data to be collected. This exercise is similar to separating wheat from chaff! Provisioning the correct amount of the cloud resources is the key to ensure that any big data project achieve the impressive returns on its investments.
- **Unavailability of Query Language:** There is no specific query language for big data. When moving toward big data we are giving up a very powerful query language i.e. SQL and at the same time compromising the consistency and accuracy. It is important to understand that if the relational database using SQL is serving the purpose effectively then what is the need to switch to big data (After all it is not the next generation of database technology). Big data is unstructured data which scale up our analysis and has a limited query capability.
- **Lack of Analyst:** One of the major emerging concerns is the **lack of analysts** who have the expertise to handle big data for finding useful patterns using cloud computing. It is estimated that nearly 70% business entities do not have the necessary skills to understand the opportunities and challenges of big data, even though they acknowledge its importance for the survival of the business. More than two third believe their job profile has changed because of the evolution of big data in their organization. Business experts have emphasized that more can be earned by using simple or traditional technology on small but relevant data rather than wasting money, effort and time on big data and cloud computing which is like digging through a mountain of information with fancy tools.

- **Identification of Right Dataset:** Till date most of the enterprise feels ill equipped to handle big data and some who are competent to handle this data are struggling to identify the right data set. Some of the enterprise are launching major project for merely capturing raw web data and converting it into structured usable information ready for analysis. Take smaller step toward big data and don't jump directly on big data. It is advisable that the transformation towards big data and cloud computing should be a gradual process rather than a sudden long jump.
- **Proactive Approach:** Careful planning is required about the quantum, nature and usage of data so that long term data requirement may be identified well in advance. Scale of big data and cloud computing may be calibrated according to such medium or long term plans. How much data is required by a particular enterprise in coming years, as big data is growing exponentially petabytes over petabytes so you must have resources to scale up your data storage as require using cloud. The enterprise may have resources for storing data today but plan for future well in advance. For this there is a need for making strategies today and how the existing infrastructure can store the volumes of data in the future. There is no need of immediately switching big data to cloud; do it but gradually.
- **Security Risks:** In order for cloud computing is to adopt universally, security is the most important concern (Mohammed, 2011). Security is one of the major concerns of the enterprise which are using big data and cloud computing. The thought of storing company's data on internet make most of the people insecure and uncomfortable which is obvious when it comes to the sensitive data. There are so many security issues which need to be settled before moving big data to cloud. Cloud adoption by businesses has been limited because of the problem of moving their data into and out of the cloud.
- **Data Latency:** Presently, real time data has low latency. The cloud does not currently offer the performance necessary to process real-time data without introducing latency that would make the results too "stale" (by a millisecond or two) to be useful. In the coming years it may be possible that technologies may evolve that can accommodate these ultra low-latency use cases but till date we are not well equipped.
- **Identification of Inactive Data:** The top challenges in handling the big data is the growth of data. Data is growing day by day. The enterprise which is capable of handling data today may not be able to handle the data tomorrow. The most important thing about data is to identify the active data. The ironic thing about the data is that most of the enterprise data are inactive (about 70%) and is no longer used by the end user. For example, the typical data access profile for corporate data follows a pattern where data is used most often in the days and weeks after it is created and then is used less frequently thereafter.

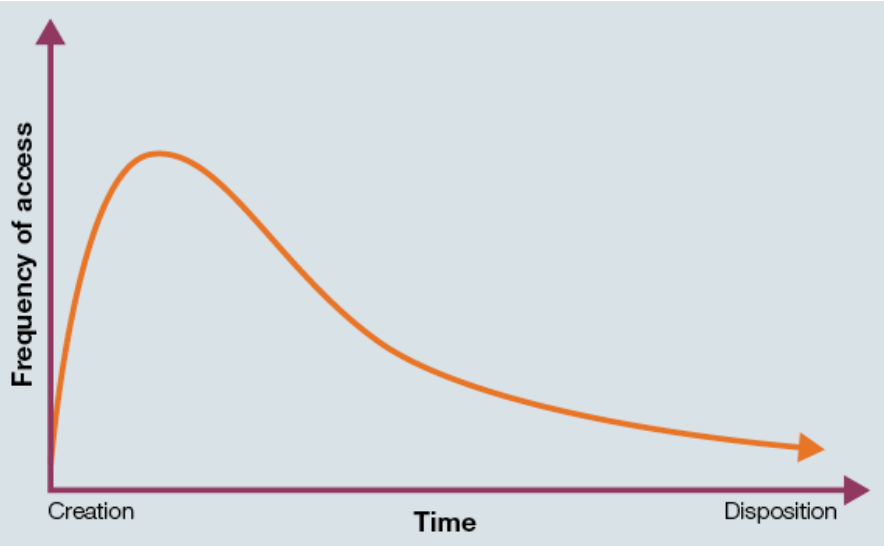

Fig. 7 A data lifecycle profile (Source: IBM Corporation)

Different applications have different lifecycle profile. There are some applications which keeps data active for several months such as banking applications on the other hand data in emails will be active for few days later on this data becomes inactive and sometimes, is of no use. In many companies inactive data takes up 70% or more of the total storage capacity, which means that storage capacity constraints, which are the root cause of slow storage management, are impacted severely by inactive data that is no longer being used. This inactive data needs to be identified for storage optimization and efforts required to store big data. If we are using cloud for storing inactive data then we are wasting our money. It is utmost required that we identify inactive data and remove the inactive data as soon as possible.

- **Cost:** Cost is one of the other major issues which need to be address properly. At first glance, a cloud computing application for storing big data may appear to be a lot cheaper than a particular software and hardware installed for storing and analysis of big data. But it should be ensured that the cloud application has all the features that the software has and if it doesn't, some features may be missing which is important to us. Cost savings of cloud computing primarily occur when a business first starts using it. SaaS (Software as a Service) applications will have lower total cost of ownership for the first two years because these applications do not require large capital investment for licenses or support infrastructure. After that, the on-premises option can become the cost-savings winner from an accounting perspective as the capital assets involved depreciate.

- **Validity of Patterns:** The validity of the patterns found after the analysis of big data is another important factor. If the patterns found after analysis are not at all valid then the whole exercise of collecting, storing and analysis of data go in vain which involves effort, time and money.

7 Discussions

Big Data, just like Cloud Computing, has become a popular phrase to describe technology and practices that have been in use for many years. Ever-increasing storage capacity and falling storage costs along with vast improvements in data analysis, however, have made big data available to a variety of new firms and industries. Scientific researchers, financial analysts and pharmaceutical firms have long used incredibly large datasets to answer incredibly complex questions. Large datasets, especially when analyzed in tandem with other information, can reveal patterns and relationships that would otherwise remain hidden.

Every organization wants to convert big data into business values without understanding the technological architecture and infrastructure. The big data projects may fail because the organization want to draw too much and too soon. For achieving their business goals every organization must first learn how to handle big data and challenges associated with big data. Cloud computing can be a possible solution as it provides a solution that is cost efficient while meeting the need of rapid scalability an important feature when dealing with big data. Using cloud computing for big data storage and analysis is not without problems. There are various problems such as downtime, Herd instinct syndrome, and unavailability of query language, lack of analyst, Identification of right dataset, security risks, cost and many more. These issues need to be addressed properly before switching big data to cloud.

8 Conclusion

Over the last one decade cloud computing and derivative technologies have emerged and developed. Like any other technology its growth and fate depends on its need and suitability for various purposes. Cloud computing may not be termed as a revolutionary technology but another offshoot of ever-growing internet based gamut of technologies. On the other hand big data is also emerging as a new keyword in all the businesses. Data generated through social media sites such as Facebook, Twitter and You tube is termed as big data or unstructured data. Big Data is becoming a new way for exploring and discovering interesting, valuable patterns from the data. The volume of data is constantly increasing and enterprises which are capable of handling data today may not be able to handle data tomorrow. Big data is comparatively younger technology which is marking its footprints on the landscape of web based technologies. However, cloud computing is the natural platform for storing big data but there are several trade offs to use both

these technologies in unison. Cloud enables big data processing for enterprises of all sizes by relieving a number of problems, but there is still complexity in extracting the business value from a sea of data. Many big projects are failed due to the lack of understanding of problem associated with big data and cloud computing.

It has been the Endeavour of this chapter to emphasis the point that any attempt to switch to the cloud computing from legacy platform should be well researched, cautious and gradual. The chapter has invited readers attention towards such trade offs like herd instinct syndrome, unavailability of query language, lack of analyst, identification of right dataset and many more. Future of these technologies is promising provided these challenges are successfully addressed and overcome.

References

Agrawal, D., Das, S., Abbadi, A.E.: Big data and cloud computing: New wine or just new bottles? PVLDB 3(2), 1647–1648 (2010)

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Widom, J.: Challenges and Opportunities with Big Data – A community white paper developed by leading researchers across the United States (2012), http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (retrieved)

Ahuja, S.P., Moore, B.: State of Big Data Analysis in the Cloud. Network and Communication Technologies 2(1), 62–68 (2013)

Ahuja, S.P., Mani, S.: Empirical Performance Analysis of HPC Bench-marks Across Variations of Cloud Computing. International Journal of Cloud Applications and Computing (IJCAC) 3(1), 13–26 (2013)

Ahuja, S.P., Mani, S.: Availability of Services in the Era of Cloud Computing. Journal of Network and Communication Technologies (NCT) 1(1), 97–102 (2012)

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Zaharia, M.: A view of cloud computing. Communications of the ACM 53(4), 50–58 (2010), doi:10.1145/1721654.1721672

Aslam, U., Ullah, I., Ansara, S.: Open source private cloud computing. Interdisciplinary Journal of Contemporary Research in Business 2(7), 399–407 (2010)

Basmadjian, R., De Meer, H., Lent, R., Giuliani, G.: Cloud Computing and Its Interest in Saving Energy: the Use Case of a Private Cloud. Journal of Cloud Computing: Advances, Systems and Applications 1(5) (2012), doi:10.1186/2192-113X-1-5.

Begoli, E., Horey, J.: Design Principles for Effective Knowledge Discovery from Big Data. In: 2012 Joint Working IEEE/IFIP Conference on Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), pp. 215–218 (2012), http://dx.doi.org/10.1109/WICSA-ECSA.212.32

Chadwick, D.W., Casenove, M., Siu, K.: My private cloud – granting federated access to cloud resources. Journal of Cloud Computing: Advances, Systems and Applications 2(3) (2013), doi:10.1186/2192-113X-2-3

Chadwick, D.W., Fatema, K.: A privacy preserving authorizations system for the cloud. Journal of Computer and System Sciences 78(5), 1359–1373 (2012)

Chen, J., Wang, L.: Cloud Computing. Journal of Computer and System Sciences 78(5), 1279 (2011)

Cole, B.: Looking at business size, budget when choosing between SaaS and hosted ERP. E-guide: Evaluating SaaS vs. on premise for ERP systems (2012), http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf (retrieved)

Coronel, C., Morris, S., Rob, P.: Database Systems: Design, Implementation, and Management, 10th edn. Cengage Learning, Boston (2013)

Dai, W., Bassiouni, M.: An improved task assignment scheme for Hadoop running in the clouds. Journal of Cloud Computing: Advances, Systems and Applications 2, 23 (2013), doi:10.1186/2192-113X-2-23

Dialogic, Introduction to Cloud Computing (2010), http://www.dialogic.com/~/media/products/docs/whitepapers/12023-cloud-computing-wp.pdf

Eaton, Deroos, Deutsch, Lapis, Zikopoulos: Understanding big data: Ana-lytics for enterprise class Hadoop and streaming data. McGraw-Hill, New York (2012)

Edd, D.: What is big data (2012), http://radar.oreilly.com/2012/01/what-is-big-data.html

Fairfield, J., Shtein, H.: Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. Journal of Mass Media Ethics: Exploring Questions of Media Morality 29(1), 38–51 (2014)

Fernado, N., Loke, S., Rahayu, W.: Mobile cloud computing: A survey. Future Generation Computer Systems 29(1), 84–106 (2013)

Fox, G.: Recent work in utility and cloud computing. Future Generation Computer System 29(4), 986–987 (2013)

Gardner, D.: GigaSpaces Survey Shows Need for Tools for Fast Big Data, Strong Interest in Big Data in Cloud. ZDNet Briefings (2012), http://Di-rect.zdnet.com/gigaspaces-survey-showsneed-for-tools-for-fast-big-data-strong-interest-in-big-data-incloud-7000008581/

Garg, S.K., Versteeg, S., Buygga, R.: A framework for ranking of cloud computing services. Future Generation Computer System 29(4), 1012–1023 (2013)

Gartner, Top 10 Strategic Technology Trends For 2014 (2013), http://www.forbes.com/sites/peterhigh/2013/10/14/gartner-top-10-strategic-technology-trends-for-2014/

Géczy, P., Izumi, N., Hasida, K.: Cloud sourcing: Managing cloud adoption. Global Journal of Business Research 6(2), 57–70 (2012)

Grolinger, K., Higashino, W.A., Tiwari, A., Capretz, M.: Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing: Advances, Systems and Applications 2, 22 (2013)

Hadoop Project (2009), http://hadoop.apache.org/core/

Han, Q., Abdullah, G.: Research on Mobile Cloud Computing: Review, Trend and Perspectives. In: Proceedings of the Second International Conference on Digital Information and Communication Technology and its Ap-plications (DICTAP), pp. 195–202. IEEE (2012)

IBM. Data growth and standards (2013), http://www.ibm.com/developerworks/xml/library/x-datagrowth/index.html?ca=drs

IDC. Digital Universe Study: Extracting Value from Chaos (2013), http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

IDC. Worldwide Big Data Technology and Services 2012-2015 Forecast (2011), http://www.idc.com/getdoc.jsp?containerId=233485

Intel, Big Data in the Cloud: Converging Technology (2013), http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf.Intel.com

IOS Press. Guidelines on security and privacy in public cloud computing. Journal of E-Governance 34, 149–151 (2011), doi:10.3233/GOV-2011-0271

Jackson, K.L.: Platform-as-a-service: The game changer (2012), http://www.forbes.com/sites/kevinjackson/2012/01/25/platform-as-a-service-the-game-changer/ (retrieved)

Ji, C., Li, Y., Qiu, W., Awada, U., Li, K.: Big Data Processing in Cloud Computing Environments. In: 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), pp. 17–23. IEEE (2012)

Juniper, Introduction to Big Data: Infrastructure and Networking Consideration (2012), http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf

Knapp, M.M.: Big Data. Journal of Electronic Resources in Medical Libraries 10(4), 215–222 (2013)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011), http://www.mckinsey.com/InsightsMGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation (retrieved)

Marciano, R.J., Allen, R.C., Hou, C., Lach, P.R.: Big Historical Data" Feature Extraction. Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives 9(1), 69–80 (2013)

Mohammed, D.: Security in Cloud Computing: An Analysis of Key Drivers and Constraints. Information Security Journal: A Global Perspective 20(3), 123–127 (2011)

NIST, Working Definition of Cloud Computing v15 (2009), http://csrc.nist.gov/groups/SNS/Cloudcomputing/

Netapp, Storage Infrastructure for Cloud Computing (2009), http://delimiter.com.au/wp-content/uploads/2010/10/Storage-infrastructure-for-cloud-computing-NetApp.pdf

Oracle, Oracle platform as a service (2012), http://www.oracle.com/us/technologies/cloud/oracle-platform-as-a-service-408171.html (retrieved)

Oracle, Oracle: Big Data for the Enterprise (2013), http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf (retrieved)

Pokorny, J.: NoSQL databases: a step to database scalability in web environ-ment. In: Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services (iiWAS 2011), pp. 278–283. ACM, New York (2011), http://doi.acm.org/10.1145/2095536.2095583 (retrieved)

Prince, J.D.: Introduction to Cloud Computing. Journal of Electronic Resources in Medical Libraries 8(4), 449–458 (2011)

Promise, Cloud Computing and Trusted Storage (2010), http://firstweb.promise.com/product/cloud/PROMISETechnologyCloudWhitePaper.pdf

Rouse, M.: Infrastructure as a Service (2010b), http://searchcloudcomputing.techtarget.com/definition/Infrastructure-as-a-Service-IaaS (retrieved)

Sims, K.: IBM Blue Cloud Initiative Advances Enterprise Cloud Computing (2009), http://www-03.ibm.com/press/us/en/pressrelease/26642.wss

Singh, S., Singh, N.: Big Data analytics. In: International Conference on Communication, Information & Computing Technology (ICCICT), pp. 1–4 (2012), http://dx.doi.org/10.1109/ICCICT.2012.6398180

Spillner, J., Muller, J., Schill, A.: Creating optimal cloud storage systems. Future Generation Computer Systems 29(4), 1062–1072 (2013)

Villars, R.L., Olofson, C.W., Eastwood, M.: Big data: What it is and why you should care. IDC White Pape. IDC, Framingham (2011)

Yuri, D.: Addressing Big Data Issues in the Scientific Data Infrastructure (2013), https://tnc2013.terena.org/includes/tnc2013/documents/bigdata-nren.pdf

Zhang, Y., Gao, Q., Gao, L., Wang, C.: iMapReduce: A Distributed Computing Framework for Iterative Computation. Journal of Grid Computing 10(1), 47–68 (2012)

Big Data Movement: A Challenge in Data Processing

Jaroslav Pokorný, Petr Škoda, Ivan Zelinka, David Bednárek, Filip Zavoral, Martin Kruliš, and Petr Šaloun

Abstract. This chapter discusses modern methods of data processing, especially data parallelization and data processing by bio-inspired methods. The synthesis of novel methods is performed by selected evolutionary algorithms and demonstrated on the astrophysical data sets. Such approach is now characteristic for so called Big Data and Big Analytics. First, we describe some new database architectures that support Big Data storage and processing. We also discuss selected Big Data issues, specifically the data sources, characteristics, processing, and analysis. Particular interest is devoted to parallelism in the service of data processing and we discuss this topic in detail. We show how new technologies encourage programmers to consider parallel processing not only in a distributive way (horizontal scaling), but also within each server (vertical scaling). The chapter also intensively discusses interdisciplinary intersection between astrophysics and computer science, which has been denoted astroinformatics, including a variety of data sources and examples. The last part of the chapter is devoted to selected bio-inspired methods and their application on simple model synthesis from

Jaroslav Pokorný · David Bednárek · Filip Zavoral · Martin Kruliš
Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University, Malostranské nám. 25, 118 00 Praha 1, Czech Republic
e-mail: {bednarek,krulis,pokorny}@ksi.mff.cuni.cz

Ivan Zelinka · Petr Šaloun
Department of Computer Science, Faculty of Electrical Engineering and Computer Science
VŠB-TUO, 17. listopadu 15 , 708 33 Ostrava-Poruba, Czech Republic
e-mail: {petr.saloun,ivan.zelinka}@vsb.vcz

Petr Škoda
Astronomical Institute of the Academy of Sciences,
Fričova 298, Ondřejov, Czech Republic
e-mail: skoda@sunstel.asu.cas.cz

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_2

astrophysical Big Data collections. We suggest a method how new algorithms can be synthesized by bio-inspired approach and demonstrate its application on an astronomy Big Data collection. The usability of these algorithms along with general remarks on the limits of computing are discussed at the conclusion of this chapter.

Keywords: Big Data, Big Analytics, Parallel processing, Astroinformatics, Bio-inspired methods.

1 Introduction

Usually we talk about the Big Data when the dataset size is beyond the ability of the current software to collect, process, retrieve and manage this data. McKinsey, a leading research firm, describes in (Manyika et al., 2011) Big Data more functionally, as large pools of unstructured and structured data that can be captured, communicated, aggregated, stored, and analyzed which are now becoming part of every sector and function of the global economy. It seems that from a user's point view just Big Analytics is the most important aspect of Big Data computing. Unfortunately, large datasets are expressed in different formats, for instance, relational, XML, textual, multimedia or RDF, which may cause difficulties in its processing, e.g. by data mining algorithms. Also, increasing either data volume in a repository or the number of users of this repository requires more feasible solution of scaling in such dynamic environments than it is offered by traditional database architectures.

Users have a number of options how to approach the problems associated with Big Data. For storing and processing large datasets they can use traditional parallel database systems, Hadoop technologies, key-value datastores (so called NoSQL databases), and also so called NewSQL databases.

NoSQL databases are a relatively new type of databases which is becoming more and more popular mostly among web companies today. Clearly, Big Analytics is done also on big amounts of transaction data as extension of methods used usually in technology of data warehouses (DW). But DW technology was always focused on structured data in comparison to much richer variability of Big Data as it is understood today. Consequently, analytical processing of Big Data requires not only new database architectures but also new methods for analysing the data. We follow up the work (Pokorny, 2013) on NoSQL databases and focus in more extent on challenges coming with Big Data, particularly in Big Analytics context. We relate principles of NoSQL databases and Hadoop technologies with Big Data problems and show some alternatives in this area.

In addition, as modern science created a number of large datasets, storing and organizing the data themselves became an important problem. Although the principal requirements placed on a scientific database are similar to other database applications, there are also significant differences that often cause that standard database architectures are not applicable. Till now, the parallel capabilities and the extensibility of relational database systems were successfully used in a number of

computationally intensive analytical or text-processing applications. Unfortunately these database systems may fail to achieve expected performance in scientific tasks for various reasons like invalid cost estimation, skewed data distribution, or poor cache performance. Discussions initiated by researchers have shown advantages of specialized databases architectures for stream data processing, data warehouses, text processing, business intelligence applications, and also for scientific data.

There is a typical situation in many branches of contemporary scientific activities that there are incredibly huge amounts of data, in which the searched answers are hidden. As an example we can use astronomy and astrophysics, where the amount of data is doubled roughly each nine months (Szalay and Gray, 2001; Quinn et al., 2004). It is obvious, that the old classical methods of data processing are not usable, and to successfully solve problems whose dynamics is "hidden" in the data, new progressive methods of data mining and data processing are needed. And not only the astronomy needs them.

The research in almost all natural sciences is facing today the data avalanche represented by an exponential growth of information produced by big digital detectors, sensor networks and large-scale multi-dimensional computer simulations stored in the worldwide network of distributed archives. The effective retrieval of a scientific knowledge from petabyte-scale databases requires the qualitatively new kind of scientific discipline called *e-Science*, allowing the global collaboration of virtual communities sharing the enormous resources and power of supercomputing grids (Zhang et al., 2008; Zhao et al., 2008). As the data volumes have been growing faster than computer technology can cope with, a qualitatively new research methodology called Data Intensive Science or X-informatics is required, based on an advanced statistics and data mining methods, as well as on a new approach to sharing huge databases in a seamless way by global research communities. This approach, sometimes presented as a Fourth Paradigm (Hey et al., 2010) of contemporary science, promises new scientific discoveries as a result of understanding hidden dependencies and finding rare outliers in common statistical patterns extracted by machine learning methods from petascale data archives.

The implementation of X-informatics in astronomy, i.e. *Astroinformatics*, is a new emerging discipline, integrating computer science, advanced statistics, and astrophysics to yield new discoveries and better understanding of nature of astronomical objects. It has been fully benefitting from the long-term skill of astronomy of building well-documented astronomical catalogues and automatically processed telescope and satellite data archives. The astronomical Virtual Observatory project plays a key role in this effort, being the global infrastructure of federated astronomical archives, web-based services, and powerful client tools supported by supercomputer grids and clusters. It is driven by strict standards describing all astronomical resources worldwide, enabling the standardized discovery and access to these collections as well as advanced visualization and analysis of large data sets. Only sophisticated algorithms and computer technology can successfully handle such data flood. Thus a rich set of data processing methods has been developed since today together with increasing power of computational hardware.

One of possible unconventional methods that can be used in Big Data processing, even in parallel form, are evolutionary algorithms. They mimic working principles from natural evolution by employing a population–based approach, labelling each individual of the population with a fitness and including elements of random, albeit the random is directed through a selection process. They are suitable for optimizing systems where the functional relationship between the independent input variables and the output (objective function) of a system is not explicitly known. Using stochastic optimization algorithms such as Genetic Algorithms, Simulated Annealing and Differential Evolution, a system is confronted with a random input vector and its response is measured. This response is then used by the algorithm to tune the input vector in such a way that the system produces the desired output or target value in an iterative process.

Most engineering problems can be defined as optimization problems, e.g. the finding of an optimal trajectory for a robot arm, the optimal thickness of steel in pressure vessels, the optimal set of parameters for controllers, optimal relations or fuzzy sets in fuzzy models, etc. Solutions to such problems are usually difficult to find as their parameters usually include variables of different types, such as floating point or integer variables. Evolutionary algorithms, such as the Genetic Algorithms, Particle Swarm, Ant Colony Optimization, Scatter Search, Differential Evolution, etc., have been successfully used in the past for these engineering problems, because they can offer solutions to almost any problem in a simplified manner: they are able to handle optimizing tasks with mixed variables, including the appropriate constraints, and they do not rely on the existence of derivatives or auxiliary information about the system, e.g. its transfer function.

This chapter is going to be focused on modern methods of data processing, its parallelization with attention to the bio-inspired methods with explanation how selected evolutionary algorithms can do synthesis of models and novel algorithms, on the astrophysical databases. We also give a brief overview on usability areas of the algorithms and end with some general remarks of the limits of computing.

Section 2 describes two types of database architectures: the traditional universal one going through the whole history of databases and Hadoop-like implementing MapReduce (M/R) framework. In Section 2.1 we present a short overview of NoSQL technology, particularly data models and architectures. Section 2.3 is devoted to Big Data problems, i.e. their sources, characteristics, processing, and analysing. Section 3 is focused on parallelism in the service of data processing and is in fact continuation of the Section 2. In Section 4 we introduce astroinformatics as a possible source of the Big Data. Section 5 discusses possibilities of selected bio-inspired methods and their use on model synthesis and some suggestion (based on already done experiments) on how new algorithms can be synthesized by bio-inspired methods to be used for Big Data processing. Section 6 summarizes the chapter and emphasizes the role of Big Analytics in datasets produced in e-Science.

2 Data Processing in the World of Big Data

2.1 Database Architectures

In (Härder and Reuter, 1983) the authors described today already classical universal DBMS architecture based on a mapping model consisting from five abstraction layers. In the most general version the architecture is encapsulated together with use of the SQL language in the L1 layer. Layers L2-L5 include record-oriented data structures and a navigational approach, records and access path management, propagation control, and file management. The same model can be used in the case of distributed databases for every node of a network together with a connection layer responsible for communication, adaptation, or mediation services. Also a typical shared-nothing parallel relational DBMS can be described by this architecture. Layers L1-L4 are present usually at each machine in a cluster. A typical property of the universal architecture is that its users can see only the outermost (SQL) layer.

In practice, the number of layers in the universal architecture is often reduced to the well-known three-layer model.

In any case, associated database technologies both centralized and distributed were found not well-suited for web-scale data management. Perhaps the most important problem is a hard scalability of traditional DBMSs in web environment. A vertical scaling (called also *scale-up*), i.e. investments into new and expensive big servers, was replaced by database partitioning across multiple cheap machines added dynamically to a network. Such so called *horizontal scaling* (also *scale-out*) can apparently ensure scalability in a more effective and cheaper way. Data is distributed horizontally in the network that means, e.g. into groups of rows in the case of tabular data, but a vertical "shredding" or a combination of both styles are being used as well. Horizontal data distribution enables to divide computation into concurrently processed tasks.

Horizontal scaling is typical for so called NoSQL databases. Recall that NoSQL means "not only SQL" or "no SQL at all", that makes this collection of databases very diverse. NoSQL solutions starting in development from late 1990's provide simpler scalability and improved performance relative to traditional relational databases. Popularly said, the notion of NoSQL is used for non-relational, distributed data stores that often do not attempt to provide ACID guarantees. Particularly, these products are appropriate for storing semi-structured and unstructured data. NoSQL databases are also often used for storing Big Data.

When talking about Big Data, the other well-known term is Hadoop. The open source software Hadoop is based on the framework MapReduce (Dean and Ghemawat, 2008) developed in Google for data processing and the Hadoop Distributed File System (HDFS). On the top of HDFS there is e.g. the NoSQL database HBase. Hadoop has quickly become a standard in industry as a highly scalable data-intensive MapReduce platform. A typical layered architecture of Hadoop software often used in NoSQL environment has a three-layer model but it looks a little differently (see Table 1 based on (Borkar et al., 2012)).

Table 1 The three-layered Hadoop software stack

Level of abstraction	Data processing
L5	HiveQL/PigLatin/Jaql
L2-L4	Hadoop MapReduce Dataflow Layer: M/R jobs; Get/Put ops; HBase Key-Value Store
L1	Hadoop Distributed File System

One remarkable difference of the Hadoop software stack from the universal DBMS architecture is that we can access data by three different sets of tools in particular layers. The middle layer Hadoop MapReduce system server serves for batch analytics. HBase is available as a key-value layer, i.e. NoSQL database. Finally, high-level languages HiveQL (Facebook), PigLatin (Yahoo!), and Jaql (IBM) are for some users at the outermost layer at disposal. The use of declarative languages reduces code size by orders of magnitude and enables distributed or parallel execution. HDFS, Google File System (Ghemawat et al., 2003), and other belong among *distributed file systems*. Solely HDFS is good for sequential data access, while HBase provides random real-time read/write access to data.

Complexity of tasks for data processing in such alternative database architectures is minimized using programming languages like MapReduce, occurring especially in context of NoSQL databases. Obviously the approach is not easily realizable for arbitrary algorithm and arbitrary programming language. MapReduce inspired by functional programming enables to implement, e.g. multiplication of sparse matrices by a vector in a natural way, but also optimized relational joins for some analytical queries and queries involving paths in large graphs (Rajaraman et al., 2013). On the other hand, computing in such languages does not enable effective implementation of joins in general.

2.2 NoSQL Databases

The category of NoSQL databases described in the list of well-maintained and structured Web site[1] includes at least 150 products. Some of these projects are more mature than others, but each of them is trying to solve similar problems. Some of the products are in-memory databases. In-memory means data is stored in computer memory to make access to it faster. The fastest NoSQL data stores available today – Redis and Memcached belong to this category.

Another list of various opened and closed source NoSQL databases can be found in (Cattell, 2010). A very detailed explanation of NoSQL databases is presented in the work (Strauch, 2011). Some significant representatives of NoSQL databases are discussed in (Pokorny, 2013).

[1] `http://nosql-database.org/` (retrieved on 30.5.2014).

Data Models Used in NoSQL Databases

What is principal in classical approaches to databases – a (logical) data model – is in approaches to NoSQL databases described rather intuitively, without any formal fundamentals. The NoSQL terminology is also very diverse and a difference between conceptual and database view of data is mostly blurred.

Most simple NoSQL databases called *key-value stores* (or *big hash tables*) contain a set of couples (key, value). A key uniquely identifies a value (typically string, but also a pointer, where the value is stored). The value for a given key (or row-ID) can be a collection of couples composed from a name and a value attached to this name. The sequence of (name, value) couples are contained as a BLOB, usually in the format of a string. This means that data access operations, typically CRUD operations (create, read, update, and delete), have only a key as the address argument. The approach key-value reminds simple abstractions as file systems or hash tables, which enables efficient lookups. However, it is essential here, that couples (name, value) can be of different types. In terms of relational data model – they may not "come" from the same table. Though very efficient and scalable, the disadvantage of too a simple data model can be essential for such databases. On the other hand, NULL values are not necessary, since in all cases these databases are schema-less.

In a more complex case, NoSQL database stores combinations of couples (name, value) collected into collections, i.e. rows addressed by a key. Then we talk about *column NoSQL databases*. New columns can be added to these collections. There is a further level of structure (e.g. in DBMS CASSANDRA) called *super-columns*, where a column contains nested (sub)columns. Data access is improved by using column names in CRUD operations. Notice that column NoSQL databases have nothing to do with column-store DBMSs, e.g. MonetDB, which store each column of a relational table in a separate table.

The most general models are called (rather inconveniently) *document-oriented NoSQL databases*. DBMS MongoDB is a well-known representative of this category. The JSON (Java Script Object Notation) format is usually used to representation of such data structures. JSON is a binary and typed data model which supports the data types list, map, date, Boolean as well as numbers of different precisions. Document stores allow arbitrarily complex documents, i.e. subdocuments within subdocuments, lists with documents, etc., whereas column stores only allow a fixed format, e.g. strict one-level or two-level dictionaries. On a model level, they contain a set of couples (key, document).

We can observe that all these data models are in principle key-valued. The three categories considered distinguish mainly in possibilities of aggregation of (key, value) couples and accessing the values.

To the broad sense, NoSQL are of more categories, e.g. object-oriented, XML, and graph databases. Particularly, graph databases play a crucial role in social network area.

Architectures of NoSQL Databases

NoSQL databases optimize processing massive amounts of data, while imposing a relatively weak consistency model (typically row-level locks). They use lightweight coordination mechanisms, which allow them to scale while keeping a partially centralized point of control. Typically, a support of scalability works against transactionality. In fact, due to the Brewer's theorem (Brewer, 2012), assuming a partitioning tolerance in our unreliable web systems, mostly availability is prioritized over a consistency.

In the Big Data landscape NoSQL databases dominate rather for operational capabilities, i.e. interactive workloads where data is primarily captured and stored. Analytical Big Data workloads, on the other hand, tend to be addressed by parallel database systems and MapReduce. Nevertheless, NoSQL are also becoming as a dominant for Big Analytics, but with a lot disadvantages, e.g. a heavy computational model, low-level information about data processed, etc. Exclusions occur, e.g. Apache Mahout implemented on top of Hadoop brings to bear automated machine learning to finding hidden trends and otherwise unexpected or unconsidered ideas.

A new approach to database architectures supporting a combination of operational and analytical technologies can be found, e.g. in (Vinayak et al., 2012). Their ASTERIX system is fully parallel, it is able to store, access, index, query, analyze, and publish very large quantities of semi-structured data. Its architecture (Table 2) is similar to that one in Table 1, but with own Hyracks layer in the bottom to manage data-parallel computations, the Algebrics algebra layer in the middle, and the topmost ASTERIX system layer – a parallel information management system. Pregelix with Pregel API supports processing big graphs. The architecture includes also a Hadoop compatibility layer.

Table 2 Asterix software stack

<table>
<tr><th>Level of abstraction</th><th colspan="4">Data processing</th></tr>
<tr><td>L5</td><td colspan="4">Asterix QL</td></tr>
<tr><td rowspan="2">L2-L4
algebraic
approach</td><td colspan="4">HiveQL, Piglet</td></tr>
<tr><td>ASTERIX
DBMS</td><td>Other HLL
Compilers</td><td colspan="2">M/R Jobs Pregel jobs</td></tr>
<tr><td colspan="4">Algebrics
Algebra Layer | Hadoop M/R
Compatibility | Pregelix | Hayrack
jobs</td></tr>
<tr><td>L1</td><td colspan="4">Hyracks Data-parallel Platform</td></tr>
</table>

2.3 Big Data

A specification of Big Data by some concrete numbers is offered, e.g. in (Morgan, 2012). Big Data is a system that has to collect more than 100TB of data annually, it has to grow at least 60% per year and it has to be deployed on scale-out architectures.

Besides traditional enterprise data (e.g. customer information from CRM systems, transactional ERP data, web store transactions), sources of Big Data include also sensor data, digital data streams (e.g. video, audio, RFID data) or e-Science data coming from astronomy, biology, chemistry, and neuroscience, for example. In this context, related data driven applications are mentioned. As example we can use market research, smarter health-care systems, environment applications, posts to social media sites, water management, energy management, traffic management, but also astronomical data analysis.

Web plays a prominent role towards shifting to the Big Data paradigm. The textual web content, a typical example of Big Text – is a source that people want to easily consult and search. Challenges in this area include, e.g. document summarization, personalized search, and recommender systems. The social structures formed over the web, mainly represented by the online social networking applications such as Facebook, LinkedIn or Twitter, contribute intensively to Big Data. Typically, the interactions of the users within a social networking platform form graph structures, leading to the notion of Big Graph.

In general, Big Data comes from four main contexts:

- large data collections in traditional DW or databases,
- enterprise data of large, non-web-based companies,
- data from large web companies, including large unstructured data and graph data,
- data from e-Science.

In any case, a typical feature of Big Data is the absence of a schema characterization, which makes difficulties when we want to integrate structured and unstructured datasets.

Big Data Characteristics

Big Data embodies data characteristics created by our digitized world:

Volume	data at scale - size from TB to PB and more. Too much volume is a storage issue, but too much data is also a Big Analytics issue.
Velocity	data in motion – analysis of streaming data, structured record creation, and availability for access and delivery. Velocity means both how quickly data is being produced and how quickly the data must be processed to meet demand.
Variety	data in many formats/media types – structured, unstructured, semi-structured, text, media.
Veracity	uncertainty/quality – managing the reliability and predictability of inherently imprecise data.

The first three *V* have been introduced in (Gartner, Inc., 2011), the *V* from Veracity has been added by Dwaine Snow in his blog Dwaine Snow's Thoughts on Databases and Data Management in 2012. Both Variety and Velocity are actually working against the Veracity of the data. They decrease the ability to cleanse the data before analysing it and making decisions. A fifth *V* occurred in (Gamble and Goble, 2011) belongs to

Value worthwhile and valuable data for business.

Data value vision includes creating social and economic added value based on the intelligent use, management and re-uses of data sources with a view to increase business intelligence (BI) and efficiency of both private and business sectors as well as to support new business opportunities. Then, a major new trend in information processing is the trading of original and enriched data, effectively creating an information economy.

Sometimes another *V* is considered:

Visualization visual representations and insights for decision-making (e.g. word clouds, maps, clustergrams, history flows, spatial information flows, and infographics).

Big Data Processing

A general observation is that as data is becoming more and more complex, also its analysis is becoming increasingly complex. To exploit this new resource, we need to scale up and scale out both infrastructures and standard techniques. Big Data and high performance computing (HPC) are playing essential roles in attacking the most important problems in this context. We may distinguish between the HPC and the Big Data in terms of combinatorial complexity of storing and the complexity of addressing the data. In the first case, the problem is not related to the amount of data but to the combinatorial structure of the problem. In the second case, the problem is in the scalability by linearization rather than in the parallelism.

Big Data processing involves interactive processing and decision support processing of data-at-rest and real-time processing of data-in-motion. The latter is usually performed by Data Stream Management Systems. Hadoop based on MapReduce paradigm is appropriate rather for decision support. However, MapReduce is still very simple technique compared to those used in the area of distributed databases. MapReduce is well suited for applications which analyse elements of a large dataset independently; however, applications whose data access patterns are more complex must be built using several invocations of the Map and Reduce steps. The performance of such design is dependent on the overall strategy as well as the nature and quality of the intermediate data representation and storage. For example, e-Science applications involve complex computations which pose new challenges to MapReduce systems. As scientific data is often skewed and the runtime complexity of the reducer task is typically high, the resulted data processing M/R jobs may be not too effective. Similarly, MapReduce is not appropriate for ad hoc analyses but rather for organized data processing. On the contrary, NoSQL systems serve rather for interactive data serving environments.

Big Analytics is about turning information into knowledge using a combination of existing and new approaches. Related technologies include

- data management (uncertainty, query processing under near real-time constraints, information extraction),
- programming models,
- machine learning and statistical methods,

- systems architectures,
- information visualization.

A highly scalable platform supporting these technologies is called a *Big Data Management System* (BDMS) in (Vinayak et al., 2012). The ASTERIX mentioned in Section 2.2 belongs to the BDMS category.

Big Analytics
Big Data is often mentioned only in context with BI. But not only BI developers also scientists analyse large collections of data. A challenge for computer specialists or data scientists is to provide these people with tools that can efficiently perform complex analytics that take into account the special nature of such data and their intended tasks. It is important to emphasise that Big Analytics involves not only the analysis and modelling phase. For example, noisy context, heterogeneity, and interpretation of results are necessary to be taken into account.

Besides these rather classical themes of mining Big Data, other interesting issues have appeared in last years, e.g. entity resolution and subjectivity analysis. The latter includes Sentiment Analysis and Opinion Mining as topics using Information Retrieval and web data analysis. A particular problem is finding sentiment-based contradictions at a large scale and to characterize (e.g. in terms of demographics) and explain (e.g. in terms of news events) the identified contradictions.

3 Parallelism in the Service of Data Processing

Current development of CPU architectures clearly exhibits a significant trend towards parallelism. We can also observe an emergence of new computational platforms that are parallel in nature, such as GPGPUs or Xeon Phi accelerator cards. These new technologies force programmers to consider parallel processing not only in a distributive way (horizontal scaling), but also within each server (vertical scaling). The parallelism is getting involved on many levels and if the current trends hold, it will play even more significant role in the future.

3.1 Performance Evaluation

To determine the benefits of the parallelism, we need to address the issue of performance evaluation. The theoretical approach, which operates with well-established time complexities of algorithms, is not quite satisfactory in this case. On the other hand, the naïve approach of measuring the real execution time is highly dependent on various hardware factors and it suffers from significant errors of measurement. Unfortunately, the real running time is the only practical thing we can measure with acceptable relevance. In the light of these facts, we will provide most of the results as the parallel speedup, which is computed as

$$Speedup = t_{serial}/t_{parallel} \quad (1)$$

where t_{serial} is the real time required by the best serial version of the algorithm and $t_{parallel}$ is the real time taken by the parallel version. Both versions are executed on the same data, thus solving exactly the same problem.

The speedup is always provided along with the number of cores (threads, devices, etc.) used for the parallel version of the algorithm. To achieve sufficient precision, the times should be measured several times and outlining results should be dismissed.

3.2 Scalability and Amdahl's Law

We usually measure the speedup in several different settings, when the parallel implementation utilizes a different number of cores. These tests are designed to assess the scalability of the algorithm. In other words, how many computational units can be efficiently utilized, or how well is the problem parallelizable. In an optimal case, the speedup is equal to the number of computational units used (i.e., 2× on dual-core, 4× on quad-core, etc.) and we denote this case the *linear speedup*. The scalability also helps us predict how the application will perform in the future as each new generation of CPUs, GPUs, or other parallel devices has more cores than the previous generation.

The scalability of an algorithm can also be determined by measuring the ratio of its serial and parallel parts. If we identify the sizes of these parts, we can use the Amdahl's Law (Amdahl, 1967)

$$SN = \frac{1}{\left((1-P)+\frac{P}{N}\right)} \tag{2}$$

to estimate the speedup in advance. The SN denotes speedup of the algorithm when N computational units are used and the P is the relative size of the parallel part of the algorithm. The speedup estimation becomes particularly interesting when the N tends to the infinity:

$$lim\, SN = lim\, (1/(1-P) + P/N) = 1/(1-P) \tag{3}$$

The behaviour in infinity often helps us understand, what happens when an algorithm is moved from multi-core CPUs with tens of cores to a multi-GPU system with thousands of cores, for instance. When the serial part of the algorithm takes 5% of total work, we will never be able to achieve greater speedup than 20×, no matter how many cores we can employ. In such case, we can observe 3.48× speedup on a quad-core CPU (which looks adequate); however, we may achieve only 19.3× speedup on a 512-core GPU card. Therefore, one of our main objectives is to reduce the serial parts as much as possible even at the cost of using an algorithm with suboptimal time complexity.

3.3 Task and Data Parallelism

There are attempts to improve the education of parallel programming (Hwu et al., 2008), inspired by the success of design patterns in programming. Design patterns in general allow dissipation of knowledge acquired by experts; in the case of parallel programming, the approach tries to identify patterns used in successful parallel applications, to generalize them, and to publish them in a compact form of design pattern (Keutzer and Mattson, 2008).

A number of design patterns or strategies were proposed and widely used in parallel computing. In the area of data processing, the following approaches are of special interest:

- *task parallelism* – the problem is decomposed into reasonably sized parts called tasks, which may run concurrently with no or very little interference,
- *data parallelism* – the problem data is decomposed into reasonably sized blocks which may be processed in parallel by the same routine,
- *pipeline* – the problem is expressed as a chain of producers and consumers which may be executed concurrently.

In the task parallelism, the program is not viewed as process divided into several threads. Instead, it is seen as a set of many small tasks (Khan et al., 1999). A task encapsulates a fragment of data together with a routine that is to be executed on the data. In task-parallel systems, the execution of tasks is usually handled by a task scheduler. The scheduler maintains a set of tasks to be executed and a pool of execution threads, while its main purpose is to dispatch the task to the threads. At any given time, a thread can either be executing a task or be idle. If it is idle, the task scheduler finds a suitable task in the task pool and starts the execution of the task on the idle thread.

Having one central task scheduler would create a bottleneck that would reduce parallelism and scalability. This problem is usually solved by task-stealing (Bednárek et al., 2012), where a separate task queue is assigned to each thread in the thread pool. Thus, each thread has its own scheduler and the schedulers interact only when a thread's queue is empty – in this case, the idle thread steals tasks from another thread's queue.

Carefully designed scheduler may improve the use of CPU cache hierarchy. When a piece of data is transferred from one task to another, the scheduler will use the same CPU to execute the tasks, keeping the data hot in the cache.

On the other hand, there is an overhead associated with the work of the task scheduler, namely maintaining the task and thread pool and the execution of tasks – each task takes some time to set up before the actual code is executed and also some time to clean up after the execution. Thus, the designers of any task-parallel system must carefully choose the task size to balance between the scheduling overhead and scalability.

The data parallelism and pipeline parallelism are usually expressed in the terms of task parallelism with only minor adjustments to the task scheduler. On the other

hand, the data parallelism may natively utilize other CPU features, such as vectorization, or even other types of hardware like GPUs. The GPU architecture is designed to process large amounts of data by the same procedure, thus it is particularly suitable for these parallel problems.

3.4 Programming Environment

The programming environment plays also important role in the design of parallel applications. Different types of hardware and different problems require different solutions. Let us focus on the most important types of problems and technologies that solve them.

OpenMP[2] and the Intel Threading Building Blocks[3] are the most popular technologies used in the domain of task parallelism and CPU parallel programming. They provide language constructs and libraries, which can be easily adopted by the programmer to express various types of parallel problems. Both technologies implement sophisticated scheduler, which is optimized for multi-core CPUs.

In order to employ the raw computational power of GPUs, NVIDIA implemented proprietary framework called CUDA[4]. It allows the programmer to design generic (i.e., not only graphic related) code that is executed in parallel on the GPU hardware. AMD implemented their own solution, however, CUDA become much more popular in parallel processing and high performance computing.

With the boom of special parallel devices, a new standard for parallel computing API called OpenCL[5] was created. OpenCL standardizes host runtime function for detecting and operating parallel devices as well as language for designing pieces of code that can be compiled and executed on these devices. At present time, all major GPU developers as well as developers of new parallel devices, such as Intel Xeon Phi, implement their version of OpenCL API to allow programmers easily use their hardware and migrate their code between devices.

Even with these specialized libraries and frameworks, parallel programming still remains to be much more difficult than traditional (serial) programming. In the specific domain of (big) data processing, the problem at hand may be often simplified using well-known paradigms. One of these paradigms came from streaming systems. We can process the data as continuous, yet finite, streams of records. These data flows are processed by stages, which may run concurrently to each other, but their internal code is strictly serial. This way the hard work of concurrent scheduling is performed by the streaming system and the programmer writes serial code only. An example of such system designed primarily for parallel databases is Bobox (Bednárek et al., 2012).

[2] `http://www.openmp.org` (retrieved on 30.5.2014).
[3] `http://threadingbuildingblocks.org` (retrieved on 30.5.2014).
[4] `http://docs.nvidia.com/cuda` (retrieved on 30.5.2014).
[5] `http://www.khronos.org/opencl` (retrieved on 30.5.2014).

3.5 Programming Languages and Code Optimization

Code optimization by compilers includes a handful of key steps augmented by dozens of minor transformations. In modern compilers, the most important steps include procedure integration, vectorization, and scheduling. The term vectorization denotes the attempt to find regularly repeated patterns which may be evaluated in parallel – either by SIMD instructions (fine-grained parallelism) or by multiple cores (coarse-grained parallelism). Scheduling improves instruction-level parallelism (ILP) by permuting the order of instructions so that the execution units in a core are not delayed by dependencies between instructions. Both transformations require sufficient amount of code to act on; therefore, procedure integration is a necessary prerequisite.

These compiler technologies were developed in 1970's and 1980's when the first super-computers appeared. Later on, the relevant hardware technology (SIMD instructions, super-scalar architecture, multi-processors, etc.) descended from the astronomical-price category to consumer devices; consequently, every performance-oriented system must now contain a compiler capable of the optimization steps mentioned above.

Compiler-based optimization is limited by the ability of the compiler to prove equivalence between the original and the transformed codes. In particular, the extent of optimization depends on the ability of the compiler to precisely detect aliases and to analyze dependencies. Since deciding many questions about the behaviour of a program is, in general, algorithmically intractable, compilers are able to optimize only in the cases where the equivalence is obvious from the local structure of the code.

Thus, although the compiler technologies are generally applicable to any procedural programming language, the reachable extent of optimization depends on the ability to analyze the code which in turn depends on the programming language and the programming style.

Unfortunately, modern programming languages as well as modern programming methodologies work worse in this sense than the older languages and styles. In particular, object-oriented programming introduced extensive use of pointers or other means of indirect access to objects which, in many cases, hinders the ability of compilers to analyze and optimize the code.

Consequently, the available compilers of modern procedural languages like Java or C# still lack the automatic parallelization features known from FORTRAN or C. While most C++ compilers share the optimizing back-end with their C siblings, extensive use of C++ features often cripples their optimizers. As a result, the FORTRAN and C languages, although now considered archaic, still dominate in the high-performance community. The reign of these languages even extends to new hardware as many GPU frameworks including CUDA and OpenCL are presented primarily in C.

While FORTRAN and C have enough features (and enormous libraries) to handle numerical calculations, their use becomes difficult when more sophisticated data structures are required. Although C is essentially as strong as C++,

software-engineering concerns favorize C++ for its better encapsulation and type-checking features. Furthermore, generic programming becomes a must in software development and the C++ language is still an uncontested leader in this category. On the other hand, C++ is quite difficult to learn and it already became a minority compared to Java and C#.

Any Big Data project contains performance-critical code by definition. In most cases, it forces the developers to use C or C++ (e.g., the MonetDB as well as the core of the Hadoop framework are implemented in C while the MongoDB in C++). This in turn makes participating in such a project difficult as it requires expertise in C or C++. Domain experts participating in a project often have a level of knowledge in programming; however, they usually lack the software-engineering education necessary to handle the development of big projects. Thus, many Big Data projects struggle not only with the amount and complexity of the data but also with the amount and complexity of the code (for instance, the Belle II project).

4 Big Data Avalanche in Astronomy

Astronomy has always fascinated humans. It can be considered the oldest of all natural sciences, dating back to ancient times. Many of the most important breakthroughs in our understanding of the world have been rooted in astronomical discoveries. Like other sciences, it is currently going through a period of explosive growth with continuous discoveries of new phenomena and even of new types of objects with unusual physical parameters.

The modern instrumentation of large telescopes like large mosaics of CCD chips, massive multi-object spectrographs with thousands of optical fibers, as well as fast radio correlators mixing inputs of tens of antennas have been producing Terabytes of raw data per night and for their reduction grids of supercomputers are needed.

Astronomy and science in general are being transformed by this exponentially growing abundance of data, which provides an unprecedented scope and opportunity for discovery. This data comes either in the form of heterogeneous, sometimes very complex data sets, or in the form of massive data streams, generated by powerful instruments or sensor networks.

Today, typical scientific databases are already tens of Terabytes in size containing catalogue information about hundreds of millions of objects and millions of spectra. For example the Sloan Digital Sky Survey (SDSS)[6] contained in its 10th release in half of year 2013 catalogue of 470 million objects and total amount of 3.3 million spectra (Ahn et al., 2013). The world largest multi-object spectrograph of LAMOST telescope (Zhao et al., 2012) is acquiring 4000 spectra per single exposure and its archive is holding in its first release DR1[7] more than 2.2 million spectra of celestial objects.

[6] `http://www.sdss.org` (retrieved on 30.5.2014).
[7] `http://data.lamost.org/dr1/?locale=en` (retrieved on 30.5.2014).

The biggest data collections are, however, produced by the sky surveys. Technology advancements in CCD chips mosaics enabled fast collection of large field of view images of high resolution, which results in detailed multicolor deep maps of the sky. Massive digital sky surveys from medium-sized telescopes are becoming the dominant source of data in astronomy, with typical raw data volumes measured in hundreds of Terabytes and even Petabytes.

The important field of Time-Domain Astronomy, looking for time variability of astronomical objects, outbursts of novae and supernovae or transitions of distant exoplanets, has been transforming into the new generation of synoptic surveys, essentially a digital panoramic cinematography of the universe, with typical data rates of ~ 0.1 TB/night. The good examples are e.g. the Palomar Quest, current archive size ~20 TB (Djorgovski et al., 2008), Catalina Rapid Transient Survey, approximately 40 TB (Drake et al., 2009) or Panoramic Survey Telescope-Rapid Response System (Pan-STARRS)[8] (Kaiser et al., 2010) which is expected to produce a scientific database of more than 100 TB within several years (Kaiser, 2007).

Much larger survey projects are now under development, most notably the Large Synoptic Survey Telescope and space missions, like Gaia and EUCLID, with estimated multi-petabyte volumes and data rates larger than 10 TB per night over many years.

The Large Synoptic Survey Telescope (LSST) (Szalay et al., 2002), which will become operational in 2019, will yield 30 TB of raw data every night requiring for the reduction of data the processing power about 400 TFLOPs.

The telescope will be located on the El Peñón peak of Cerro Pachón, a 2682 meters high mountain in northern Chile. The 3.2-gigapixel camera will be the largest digital camera ever constructed. LSST will take pairs of 15-second exposures for each field with diameter about 3.5 degrees every 20 seconds, separated by a 2-second readout. This "movie" will open an entirely new window on the universe. LSST will produce on average 15 TB of data per night, yielding an uncompressed data set of 200 PB. The camera is expected to take over 200,000 pictures (1.28 PB uncompressed) per year, far more than can be reviewed by humans.

The LSST open database will rival the largest databases in existence today in size and complexity, presenting challenging opportunities for research into database architecture and data mining in addition to ground-based astronomy and physics. Given the size and nature of the data archive, it is clear LSST will be pushing the limits of current technology to fulfill its goals. This is especially true in terms of data release image processing (detecting and characterizing sources on petabytes of images), and to support demanding data-mining efforts on the resulting petascale database.

The milestone of European space astronomy, the Gaia space mission (Perryman, 2005), which has just begun with the launch of the satellite in December 2013, is supported by the largest digital camera (consisting of 106 CCDs of 4500 x 1966 pixels) ever built for a space mission. Acquiring the telemetry data at a

[8] `http://pan-starrs.ifa.hawaii.edu` (retrieved on 30.5.2014).

mean rate of about 5 Mbit/s it is expected to produce final archive about 1 PB through 5 years of exploration.

The EUCLID space mission, approved by the European Space Agency, is scheduled for launch by the end of 2019 and directed to study the accelerating expansion of the universe (Mellier et al., 2011). Methodologies for contextual exploitation of data from different archives (which is a must for an effective exploitation of these new data) further multiply data volume measures and the related management capability requirements.

Massive data streams started to be produced also in radio astronomy by multi-antenna arrays like ALMA[9] or Low Frequency Array (LOFAR)[10]. The LOFAR reduction pipeline has to process data streams of 3.2 Gbits/s from each of the 48 stations producing after final processing the typical image "data-cube" about 100 TB within 24 hours. The final archive aims at producing petabytes per year (van Haarlem et al., 2013).

The current growth of astronomical data in large archives has been rising exponentially with the doubling constant less than 6-9 months. It is much steeper than the famous Moore's law of technology advances, which predicts the doubling time of computer resources about 18 month (Szalay and Gray, 2001; Quinn et al., 2004).

Many other cutting-edge instruments that are beginning to come online, as well as instruments planned for the immediate future, produce data in volumes much larger than the typical astronomer is accustomed to deal with. Most astronomers lack both the computer science education and the access to the resources that are required to handle these amounts of data.

Astronomy is therefore facing an avalanche of data that no one can process and exploit in full. Hence, sophisticated and innovative approaches to data exploration, data discovery, data mining analysis and visualization are being developed to extract the full scientific content contained in this data tsunami.

4.1 Virtual Observatory

Although very sophisticated, most of current astronomical archives are just isolated islands of information with unique structure, data formats and access rules (including specific search engine). The returned data has different units, scales, coordinate system or may be expressed in different variables (e.g. energy in keV in X-ray astronomy instead of wavelength or frequency in optical or radio astronomy).

Efficient access to all these scattered datasets is a big problem in astronomy, which prompted astronomers to work on the (astronomical) Virtual Observatory (VO), whose goal is to provide standards describing all astronomical resources worldwide and to enable the standardized discovery and access to these collections as well as powerful tools for scientific analysis and visualisation[11].

Most of the contemporary highly acknowledged astronomical services like Vizier, Simbad, NED or tools as Aladin are practical examples of VO technology

[9] `http://almascience.eso.org` (retrieved on 30.5.2014).
[10] `http://www.lofar.org` (retrieved on 30.5.2014).
[11] `http://www.ivoa.net/about/TheIVOA.pdf` (retrieved on 30.5.2014).

in everyday use. All the complexity of replicated database engines, XML processors, data retrieval protocols as well as distributed grids of supercomputers providing powerful services is hidden under the hood of a simple web-based form or nicely looking visualization clients delivering complex tables, images, previews, graphs or animations „just on the button click".

The key issue for success of interoperability of distinct services is the distributed service-oriented architecture based on strict standardization of data format and metadata. Astronomy has had the advantage of using the same format – FITS[12] – for all astronomical frames for several decades, but this was only a part of current success of VO. The very important part is handling of metadata with formalized semantics.

The VO data provider has to make the final calibrated data VO-compatible. This requires creation of a set of metadata (for curation, provenance and characterization) and preparation of access interface in accordance with appropriate VO standard protocols[13]. These metadata is crucial for the effective extraction of the information contained in the data sets.

The VO development and preparation of standards is coordinated by the International Virtual Observatory Alliance (IVOA)[14] . To the very similar technology seem to converge other astronomy-related science branches like, e.g. the the Virtual Magnetospheric Observatory (VMO)[15], Virtual Solar Terestrial Observatory (VSTO)[16] or Environmental Virtual Observatory (EVO)[17]. The Climatology community has recently established the Earth System Grid Federation[18] infrastructure which even presents its goal as "moving from Petabytes to Exabytes" .

The global interoperability of VO infrastructure is based on several standardized components:

VOTable

Data is not exploitable without metadata. The metadata is describing the same physical variables with the same term despite the original label used in the given table. The same holds for units. Here the most important role plays the controlled semantic vocabulary of Unified Content Descriptors (UCD)[19].

This, together with the standardized access protocols, allows to design clients which can query and retrieve data from all VO-compatible servers at once. Standard data format in VO, the VOTable[20], is an XML standard allowing full

[12] `http://fits.gsfc.nasa.gov/fits_overview.html` (retrieved on 30.5.2014).
[13] `http://www.ivoa.net/Documents/Notes/IVOAArchitecture/index.html` (retrieved on 30.5.2014).
[14] `http://ivoa.net` (retrieved on 30.5.2014).
[15] `http://vmo.nasa.gov` (retrieved on 30.5.2014).
[16] `http://www.vsto.org` (retrieved on 30.5.2014).
[17] `http://www.evo-uk.org` (retrieved on 30.5.2014).
[18] `http://esgf.org` (retrieved on 30.5.2014).
[19] `http://www.ivoa.net/documents/ latest/UCD.html` (retrieved on 30.5.2014).
[20] `http://www.ivoa.net/Documents/VOTable` (retrieved on 30.5.2014).

serialization (metadata is sent first, a stream of numbers follows) and embedded hyperlinks for real data contents (e.g. URL to FITS on remote servers).

All the available astronomical knowledge about acquisition process, observing conditions as well as the whole processing and reduction is included in the self-describing part of VOTable metadata (provenance) together with all proper credits and citations (curation metadata)[21].

All the physical properties of observation are placed in characterization metadata which should describe all the relevant information about spatial, temporal and spectral coverage, resolution, position, exposure length, filters used, etc.

VO Registry

The worldwide knowledge about the particular VO resource requires the global distributed database similar to Internet Domain Name Service (DNS). So all VO resources (catalogues, archives, services) have to be registered in one of the VO Registries[22]

The registry records are encoded in XML. Every VO resource has the unique identifier looking like URL but instead of `http://` having the prefix `ivo://`, which is even planned as one possibility for referring to datasets in astronomical journals.

All the information describing the nature of the data, parameters, characterization or even references and credits put in one registration server is being regularly harvested by all VO registries, so every desktop client may have the fresh list of everything available in VO within hours after putting it on-line.

Data Access Protocols

The transparent access of data from VO servers is accomplished using a number of strictly controlled protocols. The following ones belong among the most commonly used:

- ConeSearch[23]. It returns the catalogue information about objects in given circle (position, radius) on the celestial sphere.
- SIAP[24] (Simple Image Access Protocol) is intended for transfer of images or their part of given size and orientation.
- SSAP[25] (Simple Spectra Access Protocol) is designed to retrieve spectrum of given properties (time, position, spectral range, spectral resolution power etc.).

[21] `http://www.ivoa.net/Documents/latest/RM.html` (retrieved on 30.5.2014).
[22] `http://www.ivoa.net/Documents/RegistryInterface` (retrieved on 30.5.2014).
[23] `http://www.ivoa.net/Documents/latest/ConeSearch.html` (retrieved on 30.5.2014).
[24] `http://www.ivoa.net/Documents/latest/SIA.html` (retrieved on 30.5.2014).
[25] `http://www.ivoa.net/Documents/latest/SSA.html` (retrieved on 30.5.2014).

- SLAP[26] (Simple Line Access Protocol) mostly used in theoretical services returns the atomic or molecular data about given line transitions in selected wavelength or energy range and vice versa.
- TAP[27] (Table Access Protocol) is a complex protocol for querying very large tables (like catalogues, observing logs etc.) from many distributed servers simultaneously. It has asynchronous mode for very long time queries based on Universal Worker Service pattern (UWS)[28]. The queries are written using the specific superset of SQL, called ADQL[29] (Astronomical Data Query Language) with operators allowing selection of objects in subregion of any geometrical shape on the sky or the XMATCH operator allowing to decide the probability of match of two sets of objects in two catalogues with different error box (called cross-matching of catalogues).

VOSpace

As the real astronomical data may have a very big size (even of order of TB), it has to be moved only over high-speed links between data storage and data processing nodes without involvement of storage space on client computer. The user needs as well some virtual storage for large data sets (e.g. the big results of image queries) before being transferred to, e.g., data mining supercomputer facility or visualization node. Only the resulting graph of data mining process or movie of large-scale simulation, which is of relatively small size, need to be downloaded to user's computer. This idea is realized by the concept of virtual network storage or virtual user home directory called VOSpace[30].

VO Applications

The interaction of VO infrastructure with end user (scientist) is provided by a number of VO-compatible applications. Most of them are desktop clients (written in the multi-platform manner – in Java or Python) There are general tools for work with multidimensional data sets – VOPlot[31] or TOPCAT[32], celestial atlases for showing images over-plotted with catalogue data as Aladin[33] as well as applications for specific operations on spectra. Here we have SPLAT-VO[34], VOSpec[35],

[26] http://www.ivoa.net/Documents/latest/SLAP.html (retrieved on 30.5.2014).
[27] http://www.ivoa.net/Documents/TAP (retrieved on 30.5.2014).
[28] http://www.ivoa.net/Documents/UWS (retrieved on 30.5.2014).
[29] http://www.ivoa.net/Documents/latest/ADQL.html (retrieved on 30.5.2014).
[30] http://www.ivoa.net/Documents/VOSpace/ (retrieved on 30.5.2014).
[31] http://vo.iucaa.ernet.in/~voi/voplot.htm (retrieved on 30.5.2014).
[32] http://www.star.bris.ac.uk/~mbt/topcat/ (retrieved on 30.5.2014).
[33] http://aladin.u-strasbg.fr/aladin.gml (retrieved on 30.5.2014).
[34] http://star-www.dur.ac.uk/~pdraper/splat/splat-vo/ (retrieved on 30.5.2014).
[35] http://www.sciops.esa.int/index.php?project=SAT&page=vospec (retrieved on 30.5.2014).

and SpecView[36]. The regularly updated list of all VO applications is maintained at EURO-VO Software page[37].

The so far tedious but very important astronomical technique is the determination of spectral energy distribution (SED), which helps to reveal the physical nature of the astronomical object. The VO technology can help a lot in an aggregation of observed data and theoretical models. Building of SEDs in VO consists of collecting the scattered photometric data, transforming them into common filter systems (using the database of different filter transmission curves) and fitting theoretical model obtained as well from VO databases of model spectra. One recent application for building SEDs is VO Iris[38]. Some more complicated tools are being built as web services or web applications (with query forms etc.). An example of a very useful web-based application is Virtual Observatory SED Analyzer (VOSA)[39]

As every application is written by different developers having in mind specific type of scientific analysis, there does not exist any single complex all-purpose VO tool. Instead of this, in the spirit of UNIX thinking, the isolated applications have common interoperability interface using the Simple Application Messaging Protocol (SAMP)[40]. VO applications supporting SAMP can exchange their data (VO-Tables, spectra, images) with other SAMP-compatible application. This allows (together with command scripting) the building of complex processing and analyzing workflows by chaining the VO applications joining them even with other supercomputer grids or cloud based storages.

The VO infrastructure is the essential component for handling the astronomical Big Data, providing easy access to homogenized pre-filtered and even pre-processed data sources allowing the easy construction of very large and complex knowledge bases for the data mining.

4.2 Astroinformatics

As shown above, accomplishing the analysis in VO infrastructure may benefit from automatic aggregation of distributed archive resources (e.g. the multispectral research), seamless on-the-fly data conversion, and common interoperability of all tools and powerful graphical visualization of measured and derived quantities.

Combining the VO infrastructure power with the high performance computing on grid will allow the advanced analysis of large sky surveys feasible in a reasonable time.

---

[36] `http://www.stsci.edu/resources/software_hardware/specview` (retrieved on 30.5.2014).

[37] `http://www.euro-vo.org/?q=science/software` (retrieved on 30.5.2014).

[38] `http://www.usvao.org/science-tools-services/iris-sed-analysis-tool/` (retrieved on 30.5.2014).

[39] `http://svo2.cab.inta-csic.es/theory/vosa/` (retrieved on 30.5.2014).

[40] `http://www.ivoa.net/documents/SAMP` (retrieved on 30.5.2014).

The crucial role in understanding the results of such an analysis plays the data mining, or more properly the process of the Knowledge Discovery in Databases (KDD), as a methodology allowing the extraction of new physical knowledge from astronomical observations, which is, after all, the final goal of all scientific effort.

E-science is often referred to as the internet-enabled sharing of distributed data, information, computational resources, and team knowledge for the advancement of science. As we said in Section 1, an example of working e-Science technology in astronomy is given in the emerging new kind of astronomical research methodology - the Astroinformatics.

The astroinformatics is placed at the intersection between traditional astronomy, computer science, and information technologies and borrows many concepts from the fields of bioinformatics and geoinformatics (Brescia et al., 2012b). Its main goal is to provide the astronomical community with a new generation of effective and reliable methods and tools needed to analyze and understand massive and complex data sets and data flows which go far beyond the reach of traditionally used methods. This involves distributed database queries and data mining across distributed and integrated virtual sky survey catalogs (Borne et al., 2009).

The astroinformatics is an example of a new science methodology involving machine learning based data mining, where the new discoveries result often from the searching of outliers in common statistical patterns (Ball and Brunner, 2010). Examples of successful application of astroinformatics are the estimation of photometric red shifts (Laurino et al., 2011), quasar candidates identification (D'Abrusco et al., 2009), detection of globular clusters in galaxies (Brescia et al., 2012c), transients (Mahabal et al., 2010), or classification of emission-line galaxies (Cavuoti et al., 2014a).

Data Mining Methods in Astronomy

As said, the astroinformatics is primarily based on machine learning methods involving the KDD and data mining of massive data sets. From a wide variety of data mining methods (beyond the scope of this work) most of the astronomical problems involve partitioning of the objects into classes, thus represent a classification or clustering problem. Here, a lot of applications exploit the common techniques as Artificial Neural Networks, Decision Trees and Random Decision Forests and Support Vector Machines. The petascale of current astronomical data resources, however, presents a big challenge for the current KDD technology.

The serious threat is posed by scalability of existing algorithms and methods. It is well known that most if not all existing data mining methods scale badly with both increasing number of records and/or of features (i.e. input variables). When working on complex and massive data sets this problem is usually circumvented by extracting subsets of data, performing the training and validation of the methods on these decimated data sets and then extrapolating the results to the whole set. This approach obviously introduces biases, which are often difficult to control, but, more important, even the sampled data in case of petabyte archive would still pose serious computational challenges, which would be unmatchable by most users.

Moreover, the data mining praxis requires, for a given problem, a lengthy fine-tuning procedure that implies hundreds of experiments to be performed in order to identify the optimal method or, within the same method, the optimal architecture or combination of parameters.

The astronomical data is extremely heterogeneous and the big data sets need to be accessed and used by a large community of thousands of different users each with different goals, scientific interests and methods. It is therefore unthinkable to move such data volumes across the network from the distributed data repositories to a myriad of different users.

In addition to this the KDD processing of a reasonable sample of real data from big archives is extremely time demanding (of order of days or weeks).

So the size of the data and nature of mathematical processing required in KDD demands the massively parallel and distributed processing (including GPGPUs) as well as the new kind of data mining architecture.

DAME Infrastructure

One of the most promising approaches to solve this problem is a DAME computational infrastructure[41] developed at the University of Naples. The DAME (Data Mining and Exploration) (Brescia et al., 2012a) is an innovative, general purpose, web-based, distributed data mining infrastructure specialized in Massive Data Sets exploration with machine learning methods, and currently consists of a computational grid with more than 2400 nodes. DAME is embedded in standardisation initiatives like the VO and makes use of modern service oriented architectures and web services. It is based on a platform DAMEWARE (Data Mining Web Application Resource)[42], a public data mining service specialized on massive astrophysical data, which allows the scientific community to perform data mining and exploration experiments on massive data sets, by using a simple web browser.

By means of state of the art Web 2.0 technologies (for instance web applications and services), DAME offers several tools which can be seen as working environments where to choose data analysis functionalities such as clustering, classification, regression, feature extraction etc., together with models and algorithms, all derived from the machine learning paradigms.

The user can set-up, configure and execute experiments on his own data on top of a virtualized computing infrastructure, without the need to install any software on his local machine. Furthermore the DAME infrastructure offers the possibility to extend the original library of available tools, by allowing the end users to plug-in and execute their own codes in a simple way, by uploading the programs without any restriction about the native programming language, and automatically installing them through a simple guided interactive procedure.

Moreover, the DAME platform offers a variety of computing facilities, organized as a cloud of versatile architectures, from the single multi-core processor to a grid farm, automatically assigned at runtime to the user task, depending on the specific problem, as well as on the computing and storage requirements.

[41] `http://dame.dsf.unina.it/index.html` (retrieved on 30.5.2014).
[42] `http://dame.dsf.unina.it/dameware.html` (retrieved on 30.5.2014).

The recent DAME research is also focused on GPU processing with promising results (Cavuoti et al., 2014b). The optimistic results achieved here (the speedup factor of more than 200 achieved on genetic algorithms training phase using GPUs) could not be expected, however, in other machine learning problems. It seems to be highly probable that completely new data mining algorithms will need to be developed, tailored to the particular architecture of GPGPUs (e.g., due to the fact of their limited on-the-chip memory or the nature of their stream-based processing) (Gainaru et al., 2011).

Despite its maturity in other science fields, the KDD in astronomy still contains a lot of yet unsolved problems. One of the very specific to the astronomy is the problem of missing data and upper limits stemming from the fact that most data mining algorithms are not robust against missing data and cannot effectively deal with upper limits.

In fields other than astronomy (e.g. market analysis and many if not all bioinformatics applications) this is only a minor problem since the data set to be mined can be cleaned of all those records having incomplete or missing information and the results can be afterwards generalized. For example, if in the citizen records for one citizen the field "age" is missing, it means that the data has not been collected and not that the citizen has no age.

But if an astronomical object is missing a magnitude in a specific photometric band it may either mean that it has not been observed (as in the previous case) or, and this is much more interesting, that it is so faint that it cannot be detected in that band.

In the first case, the missing information can be easily recovered if a proper data model exists and the machine learning literature offers a variety of models, which can reconstruct the missing information.

In the second case, which includes many of the most interesting astronomical applications such as, e.g. the search for obscured or high redshift quasars, the crucial information is in the missing data themselves and this information cannot be reconstructed. The natural way of storing and accessing such a data with variable number of non-null attributes is a column NoSQL database rather than the relational ones.

Such a problem, however, seems not to be currently solved by adapting existing methods. It will probably require the implementation of a new generation of machine learning techniques, since it would need methods based on adaptive metrics capable to work equally well on points, hypersurfaces and hypervolumes of different dimensionality.

Astroinformatics and Social Networks

One of the non-standard methodologies in astroinformatics benefits from the popularity and power of social networks.

It is called Citizen Computing as well as Crowd Sourcing or U-Science. The general idea of this kind of platform is to publish massive datasets to the interested public and use their knowledge to help in classifying datasets and extracting interesting concepts. The verified training sets for classical data mining and machine

learning methods have been acquired by gathering the collective power of thousands or even millions of human brains (thus another name is the Carbon computing).

Since amateurs and laymen are aiding research in this way, this is also called *citizen science*, aiming at harvesting the power of the "wisdom of the crowds" for supporting astronomical research.

Sending the same patterns for human evaluation by multiple participants allows estimating the error bound of logical inferences which is otherwise difficult with classical data mining methods, but can also bring the new discoveries.

Unexpected Discoveries

The receiving of large-scale survey data on a global scale may result in surprising discoveries especially if integrated with human experience as in case of citizen science projects like Galaxy Zoo[43]. An outstanding example is the discovery of strange green compact galaxies called Galactic Peas (Cardamone et al., 2009) and mainly the so far not well understood gaseous blob of galactic size called Hanny's Voorwerp (Hanny's object) after the Dutch primary school teacher Hanny van Arkel. She have noticed the strange object near one galaxy during classification of galaxy shapes in project Galaxy Zoo, predecessor of Zooniverse and asked the professional astronomers about their opinion.

After being investigated by several large telescopes, radio telescopes and space observatories (including HST and Chandra), the nature of object is still unknown, although several theories already appeared (e.g. light echo of faded out quasar) (Lintott et al., 2009).

Zooniverse

The success of Galaxy Zoo project yielding tens of scientific articles resulted in more complex platform Zooniverse[44] , an initiative aiming to harvest the wisdom of the crowds not only in astronomical research, e.g. for discovering phenomena or for manually classifying images but also in Climatology, Ocean Sciences or Biology. Zooniverse currently hosts about twenty projects, exploring the surface of Moon or Mars, solar activity, black holes jets, exoplanets or deep space galaxies as well as Earth climatic changes, tropical cyclones, ocean floor or whales communication. There is even the project focused of study of life of ancient Greeks.

Involvement of crowd sourcing methods can enhance not only the physical knowledge about Universe, but may bring as well the sociological studies about motivation of people contributing to some effort which might help to prepare better setup for further citizen science projects or investigate the psychological aspects of modern IT technology (e.g. how to attract the potential users by simply looking user interface not overcrowded by large amount of menus and buttons) (Raddick et al., 2010).

[43] `http://www.galaxyzoo.org` (retrieved on 30.5.2014).
[44] `https://www.zooniverse.org/` (retrieved on 30.5.2014).

5 Big Data and Evolutionary Algorithms – Perspectives and Possibilities

Evolutionary algorithms are search methods that can be used for solving optimization problems. They mimic working principles from natural evolution by employing a population–based approach, labeling each individual of the population with fitness and including elements of random, albeit the random is directed through a selection process. Evolutionary algorithms, or better evolutionary computational techniques (a part of so called bio-inspired algorithms), are based on principles of evolution which have been observed in nature long time before they were applied to and transformed into algorithms to be executed on computers. When next reviewing some historical facts that led to evolutionary computation, as we know it now, we will mainly focus on the basic ideas, but will also allow glimpsing at the people who did the pioneering work and established the field. Maybe the two most significant persons whose research on evolution and genetics had the biggest impact on modern understanding of evolution and its use for computational purposes are Gregor Johann Mendel and Charles Darwin, for more detail see (Zelinka et al., 2010).

Gregor Johann Mendel (July 20, 1822 - January 6, 1884) was an Augustinian priest and scientist, and is often called the father of genetics for his study of the inheritance of certain traits in pea plants. The most significant contribution of Mendel for science was his discovery of genetic laws, which showed that the inheritance of these traits follows particular laws, published in (Mendel, 1865), which were later named after him.

The other important (and much more well–known and therefore here only briefly introduced) researcher whose discoveries founded the theory of evolution was the British scientist Charles Darwin. Darwin published in his work (Darwin, 1859) the main ideas of the evolutionary theory. In his book On the Origin of Species (1859) he established evolutionary descent with modification as the dominant scientific explanation of diversification in the nature.

The above-mentioned ideas of genetics and evolution have been formulated long before the first computer experiments with evolutionary principles had been done. The beginning of the evolutionary computational techniques is officially dated to the 70s of the 20th century, when famous genetic algorithms were introduced by Holland (Holland, 1975) or to the late 60s with evolutionary strategies, introduced by Schwefel (Schwefel, 1977) and Rechenberg (Rechenberg,1973), and evolutionary programming by Fogel (Fogel et al., 1966). However, when certain historical facts are taken into consideration, then one can see that the main principles and ideas of evolutionary computational techniques as well as its computer simulations had been done earlier than reported above. In fact on the beginning is A.M. Turing, first numerical experiments to the (far less famous) Barricelli and others, see (Barricelli, 1954; Barricelli, 1957) or (Zelinka et al., 2010).

In recent years, a broad class of algorithms has been developed for stochastic optimization, i.e. for optimizing systems where the functional relationship

between the independent input variables and the output (objective function) of a system is not explicitly known. Using stochastic optimization algorithms such as Genetic Algorithms, Simulated Annealing and Differential Evolution, a system is confronted with a random input vector and its response is measured. This response is then used by the algorithm to tune the input vector in such a way that the system produces the desired output or target value in an iterative process. Most engineering problems can be defined as optimization problems, e.g. the finding of an optimal trajectory for a robot arm, the optimal thickness of steel in pressure vessels, the optimal set of parameters for controllers, optimal relations or fuzzy sets in fuzzy models, etc. Solutions to such problems are usually difficult to find as their parameters usually include variables of different types, such as floating point or integer variables. Evolutionary algorithms, such as the Genetic Algorithms, Particle Swarm, Ant Colony Optimization, Scatter Search, Differential Evolution, etc., have been successfully used in the past for these engineering problems, because they can offer solutions to almost any problem in a simplified manner: they are able to handle optimizing tasks with mixed variables, including the appropriate constraints, and they do not rely on the existence of derivatives or auxiliary information about the system, e.g. its transfer function.

The evolutionary computational techniques are numerical algorithms that are based on the basic principles of Darwin's theory of evolution and Mendel's foundation of genetics. The main idea is that every individual of a species can be characterized by its features and abilities that help it to cope with its environment in terms of survival and reproduction. These features and abilities can be termed by its fitness and are inheritable via its genome. In the genome the features/abilities are encoded. The code in the genome can be viewed as a kind of "blue–print" that allows to store, process and transmit the information needed to build the individual. So, the fitness coded in the parent's genome can be handed over to new descendants and support the descendants in performing in the environment. Darwinian participation to this basic idea is the connection between fitness, population dynamics and inheritability while the Mendelian input is the relationship between inheritability, feature/ability and fitness. If the evolutionary principles are used for the purposes of complicated calculations, the following procedure is used, for details see (Zelinka et al., 2010):

1. Specification of the evolutionary algorithms parameters: for each algorithm, parameters must be defined that control the run of the algorithm or terminate it regularly, if the termination criterions defined in advance are fulfilled (for example, the number of cycles - generations). Also the cost function has to be defined. The objective function is usually a mathematical model of the problem, whose minimization or maximization leads to the solution of the problem. This function with possible limiting conditions is some kind of "environmental equivalent" in which the quality of current individuals is used in step 2.
2. Generation of the initial population (generally $N \times M$ matrix, where N is the number of parameters of an individual which is a vector of numbers

having such a number of components as the number of optimized parameters of the objective function. These components are set randomly and each individual thus represents one possible specific solution of the problem. The set of individuals is called *population*.

3. All the individuals are evaluated through a defined objective function and to each of them is assigned a direct value of the return objective function, or so called fitness.
4. Now parents are selected according to their quality (fitness, value of the objective function) or, as the case may be, also according to other criterions.
5. Crossbreeding the parents creates descendants. The process of crossbreeding is different for each algorithm. Parts of parents are changed in classic genetic algorithms, in a differential evolution, crossbreeding is a certain vector operation, etc.
6. Every descendant is mutated by means of a suitable random process.
7. Every new individual is evaluated in the same manner as in step 3.
8. The best individuals (offspring and/or parents) are selected.
9. The selected individuals fill a new population.
10. The old population is forgotten (eliminated, deleted, dies, ...) and is replaced by a new population; go to step 4.

Steps 4 - 10 are repeated until the number of evolution cycles specified before by the user is reached or if the required quality of the solution is not achieved. The principle of the evolutionary algorithm outlined above is general and may more or less differ in specific cases.

Another use of evolutionary approach is in symbolic structures and solutions synthesis, usually done by Genetic Programming (GP) or Grammatical Evolution (GE). Another interesting research was carried out by Artificial Immune Systems (AIS) or/and systems, which do not use tree structures like linear GP and other similar algorithm like Multi Expression Programming (MEP), etc. In this chapter, a different method called Analytic Programming (AP), is presented. AP is a grammar free algorithmic superstructure, which can be used by any programming language and also by any arbitrary Evolutionary Algorithm (EA) or another class of numerical optimization method. AP was used in various tasks, namely in comparative studies with selected well-known case examples from GP as well as applications on synthesis of: controller, systems of deterministic chaos, electronics circuits, etc. For simulation purposes, AP has been co-joined with EA's like Differential Evolution (DE), Self-Organising Migrating Algorithm (SOMA), Genetic Algorithms (GA) and Simulated Annealing (SA).

All case studies has been carefully prepared and repeated in order to get valid statistical data for proper conclusions. The term symbolic regression represents a process during which measured data sets are fitted; thereby a corresponding mathematical formula is obtained in an analytical way. An output of the symbolic expression could be, e.g., $(K\ x_2 + y_3)^2$, and the like. The initial idea of symbolic regression by means of a computer program was proposed in GP (Koza, 1990;

Koza, 1998). The other approach of GE was developed in (Ryan et al., 1998) and AP in (Zelinka et al., 2011). Another interesting investigation using symbolic regression were carried out (Johnson, 2004) on AIS and Probabilistic Incremental Program Evolution (PIPE), which generates functional programs from an adaptive probability distribution over all possible programs. Yet another new technique is the so called Transplant Evolution, see (Weisser and Osmera, 2010a; Weisser and Osmera, 2010b; Weisser et al., 2010), which is closely associated with the conceptual paradigm of AP, and modified for GE. GE was also extended to include DE by (O'Neill and Brabazon, 2006). Generally speaking, it is a process which combines, evaluates and creates more complex structures based on some elementary and non-complex objects, in an evolutionary way. Such elementary objects are usually simple mathematical operators (+, −, ×, ...), simple functions (sin, cos, And, Not, ...), user-defined functions (simple commands for robots – MoveLeft, TurnRight, ...), etc. An output of symbolic regression is a more complex "object" (formula, function, command,...), solving a given problem like data fitting of the so-called Sextic and Quintic problem described in (Koza et al., 2003; Zelinka et al., 2014), randomly synthesized function (Zelinka et al., 2005), Boolean problems of parity and symmetry solution (basically logical circuits synthesis) (Koza et al., 2003; Zelinka et al., 2014), or synthesis of quite complex robot control command (Koza et al., 1998; Oplatkova and Zelinka, 2006).

Genetic Programming

GP was the first tool for symbolic regression carried out by means of computers instead of humans. The main idea comes from GA, which was used in GP (Koza, 1990; Koza et al., 1998). Its ability to solve very difficult problems is well proven; e.g., GP performs so well that it can be applied to synthesize highly sophisticated electronic circuits (Koza et al., 2003).

The main principle of GP is based on GA, which is working with populations of individuals represented in the LISP programming language. Individuals in a canonical form of GP are not binary strings, different from GA, but consist of LISP symbolic objects (commands, functions, ...), etc. These objects come from LISP, or they are simply user-defined functions. Symbolic objects are usually divided into two classes: functions and terminals. Functions were previously explained and terminals represent a set of independent variables like x, y, and constants like π, 3.56, etc.

The main principle of GP is usually demonstrated by means of the so-called trees (basically graphs with nodes and edges). Individuals in the shape of a tree, or formula like $0.234Z + X - 0.789$, are called programs. Because GP is based on GA, evolutionary steps (mutation, crossover, ...) in GP are in principle the same as GA, see (Koza, 1990; Koza et al., 1998; Zelinka, 2011).

Grammatical Evolution

GE is another program developed in (O'Neill and Ryan, 2003), which performs a similar task to that of GP. GE has one advantage over GP, and this is the ability to use any arbitrary programming language, not only LISP as in the case of the canonical version of GP. In contrast to other EA's, GE was used only with a few

search strategies, and with a binary representation of the populations (O'Neill and Ryan, 2003). The last successful experiment with DE applied on GE was reported in (O'Neill and Brabazon, 2006). GE in its canonical form is based on GA, thanks to a few important changes it has in comparison with GP. The main difference is in the individual coding.

While GP manipulates in LISP symbolic expressions, GE uses individuals based on binary strings segmented into so-called *codons*. These are transformed into integer sequences and then mapped into a final program in the Backus-Naur form, the rule used for individuals transforming into a program is based on operation modulo, see (O'Neill and Ryan, 2003). Codons are then transformed into an integer domain (in the range of values in 0-255) and by means of defined grammar it is transformed into appropriate structure, see (O'Neill and Ryan, 2003; Zelinka et al., 2011).

Analytic Programming

The last method described here is called Analytic programming (AP), see (Zelinka et al., 2011), which has been compared to GP with very good results (see, e.g., (Zelinka and Oplatkova, 2003; Zelinka et al., 2005; Oplatkova and Zelinka, 2006; Zelinka et al., 2008)).

The basic principles of AP were developed in 2001 and first published in (Zelinka, 2001) and (Zelinka, 2002). AP is also based on the set of functions, operators and terminals, which are usually constants or independent variables, in the same way like in GA or GE.

All these objects create a set, from which AP tries to synthesize an appropriate solution. Because of the variability of the content of this set, it is called a *general functional set* (GFS). The structure of GFS is nested, i.e., it is created by subsets of functions according to the number of their arguments. The content of GFS is dependent only on the user. Various functions and terminals can be mixed together. For example, GFS_{all} is a set of all functions, operators and terminals, GFS_{3arg} is a subset containing functions with maximally three arguments, GFS_{0arg} represents only terminals, etc.

AP, as further described later, is a mapping from a set of individuals into a set of possible programs. Individuals in population and used by AP consist of non-numerical expressions (operators, functions, ...), as described above, which are in the evolutionary process represented by their integer position indexes (Zelinka et al., 2011). This index then serves as a pointer into the set of expressions and AP uses it to synthesize the resulting function (program) for cost function evaluation, see (Zelinka et al., 2011).

GFS need not consist only of pure mathematical functions as demonstrated above, but may also be constructed from other user-defined functions, e.g., logical functions, functions which represent elements of electrical circuits or robot movement commands, linguistic terms, etc.

Above mentioned methods of symbolic regression based on evolutionary algorithms can be used for synthesis of various programs that in symbolic regression terminology means arbitrary complex structures like mathematical formulas, e.g.

((A ∧ (((((B ∧ A) ∨ (C ∧ A) ∨ (¬C ∧ B) ∨ (¬C ∧ ¬A)) ∧ ((B ∧ A) ∨ (C ∧ A) ∨ (¬ C ∧ B) ∨ (¬ C ∧ ¬ A))) ∨ (B ∨¯ A)) ∧ ((A ∨ (B ∧ A) ∨ (C ∧ A) ∨(¬C ∧ B) ∨ (¬C ∧ ¬A)) ∧ B ∧ ((B ∧ A) ∨ (C ∧ A) ∨ (¬C ∧ B) ∨ (¬C ∧ ¬A))))) ∧C) ∧ (C ∨ (C ∨¯ (A ∧ (C ∧ ((B ∧ A) ∨ (C ∧ A) ∨ (¬C ∧ ¬A))))))

or

$$x(x^2(x(K_7 + x) - K_2) + x(K_4 - K_5) + xK_6 + K_1 - K_3 - 1)$$

Visualization of behavior of such programs can be very interesting as depicted in Figure 1, see also (Zelinka et al., 2011).

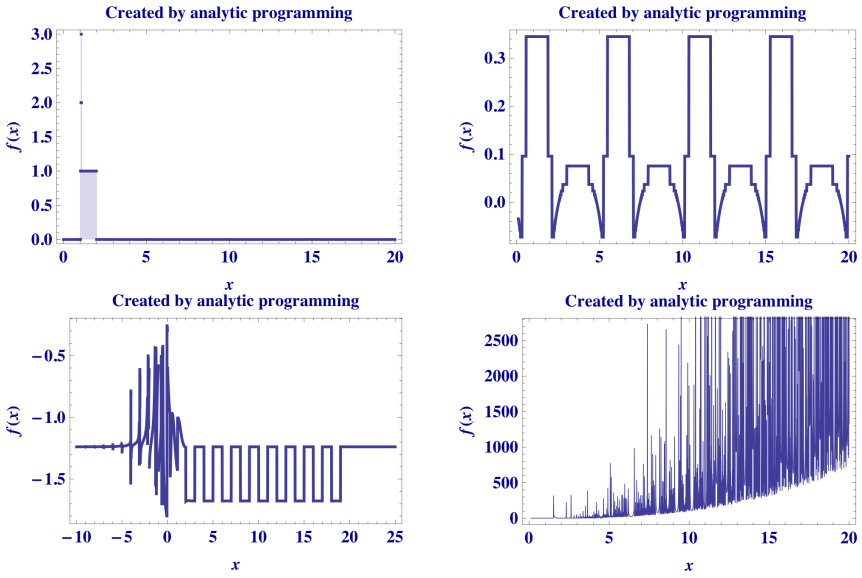

Fig. 1 Visualization of various programs behavior in time

Also synthesis of user programs that can e.g. control dynamics of robot. like

IfFoodAhead[Move, Prog3[IfFoodAhead[Move, Right], Prog2[Right, Prog2[Left, Right]], Prog2[IfFoodAhead[Move, Left], Move]]],

can be visualized as a tree of commands, see Figure 2.

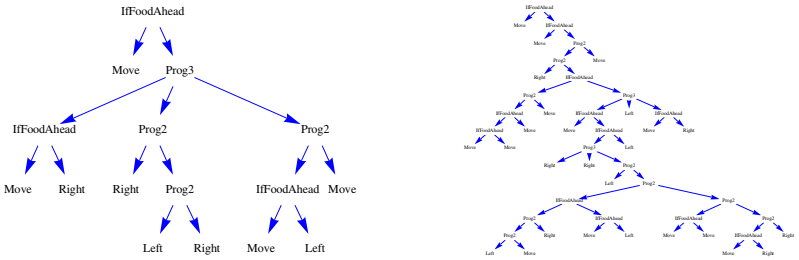

Fig. 2 Two different programs for artificial robot control in the so-called "tree" format

The power of symbolic regression and its capability from Big Data point of view is in fact that symbolic regression can synthesize programs in general, not only math formulas, electronic circuits etc. It is also possible to synthesize other algorithms, as was demonstrated in (Oplatkova, 2009; Oplatkova et al., 2010a; Oplatkova et al., 2010b; Oplatkova et al., 2010b). Then possibility to synthesize algorithms for Big Data pre-processing by means of evolutionary techniques is quite clear. This idea is already mentioned in research papers like (Yadav et al., 2013; Tan et al., 2009; Flockhart and Radcliffe, 1996; Khabzaoui et al., 2008) amongst the others. Symbolic regression can be used to estimate parameters or complete synthesize such methods like outlier detection, classification trees, model synthesis, synthesis and/or learning of artificial neural networks (ANN), Pattern Clustering, as well as in wavelets methods in data mining and more, see (Maimon and Rokach, 2010) . Also another tasks can be solved by symbolic regression like parameter estimation and algorithm synthesis in data cleaning (data preparation for the next process, i.e. noise and irrelevant data rejection), data integration (removing of redundant and inconsistent data), attribute selection (selection of the relevant data), data mining, models and predictive models synthesis, classification clustering and regression tasks amongst the others.

For example, evolutionary synthesis of ANNs has been done by wide spectra of researchers like (Dorogov, 2000; Babkin and Karpunina, 2005) and resulted structures, solving problems on defined datasets were fully functional despite its unusual structure, see Figure 3.

Another example of use of symbolic regression is a model of synthesis of emission spectral line profile in Be stars, as reported in (Zelinka et al., 2013), which may be used on Big Data sets from astrophysical databases. Be stars are characterized by prominent emission lines in their spectrum. In the past research has attention been given to creation a feature extraction method for classification of Be stars with focusing on the automated classification of Be stars based on typical shapes of their emission lines. The aim was to design a reduced, specific set of features characterizing and discriminating the shapes of Be lines. Possibility to create in an evolutionary way the model of spectra lines of Be stars is discussed there. We focus on the evolutionary synthesis of the mathematical models of Be

stars based on typical shapes of their emission lines. Analytical programming powered by classical random as well as chaotic random-like number generator is used there. Experimental data is used from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. As reported in article (Zelinka et al., 2014) various models with the same level of quality (fitting of measured data) has been synthesized as is demonstrated in Figure 4.

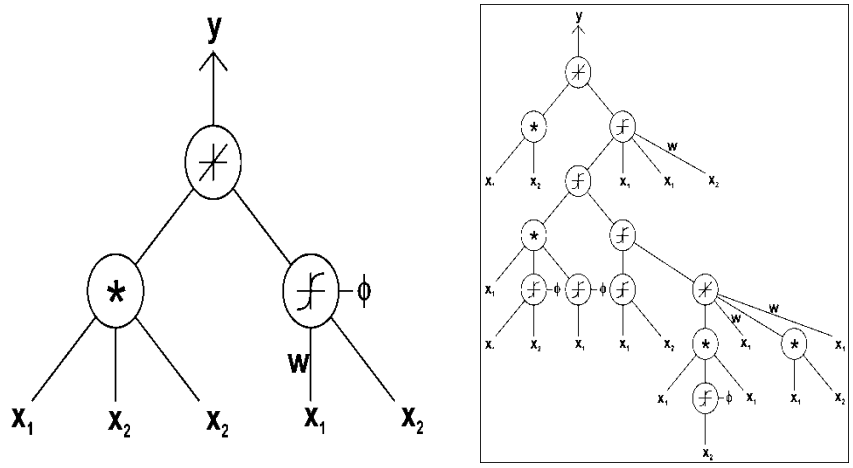

Fig. 3 An example of evolutionary synthesized ANNs

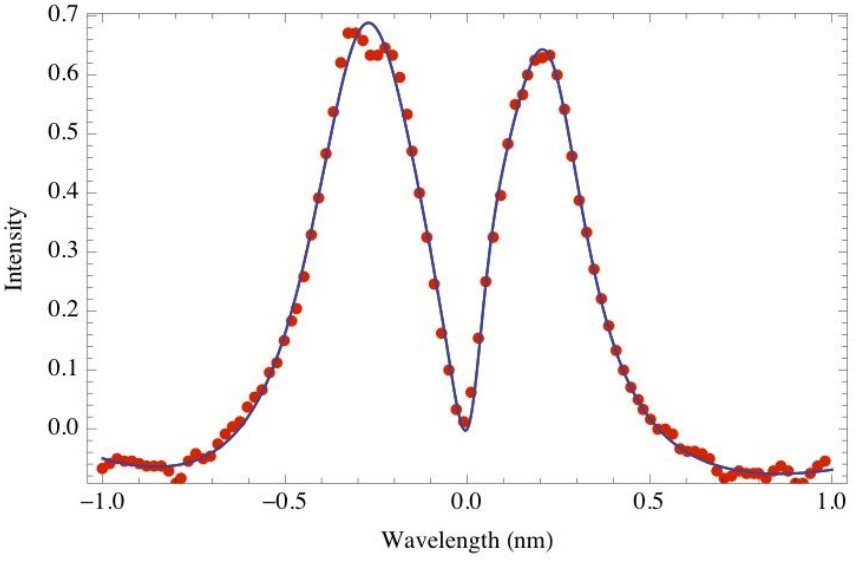

Fig. 4 Astrophysical example: Be stars spectral line profile synthesis, dots are data, solid line fitting by synthesized formula.

Limits of Computation

Unfortunately, no matter what kind of sufficiently powerful computer and elegant algorithm do we have, still some limits are there. There is still class of problems that cannot be solved algorithmically due to their nature. More exactly, there is not enough of time for their solution (Zelinka et al., 2010).

Part of these restrictions come from theoretical part of computer science (Amdahl, 1967), and part of them from physics in the form of so called physical limits that follow from the thermodynamics, quantum physics, etc. It restricts the output of every computer and algorithm by its space-time and quantum-mechanical properties. These limits, of course, are based on the contemporary state of our knowledge in physical sciences, which means that they might be re-evaluated in the case of new experimentally confirmed theories. Basic restriction in this direction is the so-called Bremermann's limit (Bremermann, 1962), according to which it is not possible to process more than 10^{51} bits per second in every kilogram of matter. In the original work of this author (Bremermann, 1962), the value of $2x10^{47}$ bites per second in one gram of matter is indicated. At first sight, this limit does not look bad, till we take "elementary" real examples for (Zelinka et al., 2010).

Another researchers (Lloyd et al., 2004) apply those limits to the transfer of information through real physical information channel (computation can also be considered as a transfer of information through a special channel) and come with very exotic results. Beside other things, it was found that if a certain mass (or energy) of the transfer medium is reached, further information could not be transferred through the channel, because the channel collapses theoretically into an astrophysical object called black hole.

It is clear that for processing large databases it is needed a sophisticated combination of computer hardware and algorithms. The promising approach is parallelization of bio-inspired methods, as many results show today.

6 Conclusion

This chapter focused on solving several data analysis problems in the domain of scientific data in accordance with more general trends in Big Data (e.g., a forecast for 2014[45]). The main research tasks of the day are attempting to improve the quality and scalability of data mining methods according to these trends. The processes of query composition - especially in the absence of a schema - and the interpretation of the obtained answers may be non-trivial since the query results may be too large to be easily interpreted by a human. Much of the present data is not stored natively in a structured format; hence, transforming the content into a suitable representation for later analysis is also a challenge. Data mining techniques, which are already widely applied to extract frequent correlations of values

[45] http://www.csc.com/big_data/publications/91710/105057-trends_in_big_data_a_forecast_for_2014 (retrieved on 14.3.2014).

from both structured and semi-structured datasets in BI can be applied for Big Analytics as well, if they are properly extended and accommodated.

So far the mining process is guided by the analyst, whose knowledge of the application scenario determines the portion of the data from which the useful patterns can be extracted. More advanced approaches are an automated mining process and an approximate extraction of synthetic information on both the structure and the contents of large datasets.

As mentioned in the previous sections, the contemporary astronomy is flooded with enormous amounts of data which rise exponentially. In order to survive this data avalanche and to extract useful knowledge, a new scientific discipline had to be introduced - the astroinformatics. The massive computing power of distributed massively parallel data mining infrastructure together with large astronomical archives accessible world-wide by VO technology augmented by the power of human brains of more than half million of volunteers participating in Zooniverse represent a great potential for a number of new outstanding discoveries.

It is also important to remember that not only sophisticated classical algorithms and its parallelization is used for Big Data processing, but also hybridization with unconventional methods like bio-inspired methods is possible and was demonstrated by another researchers like – algorithm synthesis by means of symbolic regression, or use in big data processing and analysis, amongst the others.

Acknowledgement. The following two grants are acknowledged for the financial support provided for this research: Grant Agency of the Czech Republic - GACR P103/13/08195S and P103-14-14292P, by the Development of human resources in research and development of latest soft computing methods and their application in practice project, reg. no. CZ.1.07/2.3.00/20.0072 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic. The Astronomical Institute of the Academy of Sciences of the Czech Republic is also supported by project RVO 67985815. This research has used data obtained with Perek 2m telescope at Ondřejov observatory. We are indebted to the DAME team, namely prof. Longo, dr. Brescia and dr. Cavuoti, and their collaborators, dr. Laurino and dr. D'Abrusco for inspiring ideas and introduction into the problems of astronomical machine learning.

References

Ahn, C.P., Alexandroff, R., Allende Prieto, C., et al.: The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment (2013), arXiv:1307.7735

Amdahl, G.M.: Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities. In: AFIPS Conference Proceedings, vol. (30), pp. 483–485 (1967), doi:10.1145/1465482.1465560.

Babkin, E., Karpunina, M.: Towards application of neural networks for optimal structural synthesis of distributed database systems. In: Proceedings of 12th IEEE Int. Conf. on Electronics, Circuits and Systems, Satellite Workshop Modeling, Computation and Services, Gammarth, Tunisia, pp. 486–490 (2005)

Ball, N.M., Brunner, R.M.: Data mining and machine learning in astronomy. International Journal of Modern Physics D 19(07), 1049–1107 (2010)

Barricelli, N.A.: Esempi Numerici di processi di evoluzione. Methodos, 45–68 (1954)

Barricelli, N.A.: Symbiogenetic evolution processes realized by artificial methods. Methodos 9(35-36), 143–182 (1957)

Bednárek, D., Dokulil, J., Yaghob, J., Zavoral, F.: Data-Flow Awareness in Parallel Data Processing. In: Fortino, G., Badica, C., Malgeri, M., Unland, R. (eds.) IDC 2012. SCI, vol. 446, pp. 149–154. Springer, Heidelberg (2012)

Borkar, V., Carey, M.J., Li, C.: Inside "Big Data management": ogres, onions, or parfaits? In: Proceedings of EDBT Conference, Berlin, Germany, pp. 3–14 (2012)

Borne, K., Accomazzi, A., Bloom, J.: The Astronomy and Astrophysics Decadal Survey. Astro 2010, Position Papers, No. 6. arXiv:0909.3892 (2009)

Bremermann, H.: Optimization through evolution and recombination. In: Yovits, M., Jacobi, G., Goldstine, G. (eds.) Self-Organizing Systems, pp. 93–106. Spartan Books, Washington, DC (1962)

Brescia, M., Longo, G., Castellani, M., et al.: DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases. Memorie della Societa Astronomica Italiana Supplementi 19, 324–329 (2012)

Brescia, M., Cavuoti, S., Djorgovski, G.S., et al.: Extracting Knowledge from Massive Astronomical Data Sets. In: Astrostatistics and Data Mining. Springer Series in Astrostatistics, vol. 2, pp. 31–45. Springer (2012), arXiv:1109.2840

Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Puzia, T.: The detection of globular clusters in galaxies as a data mining problem. Monthly Notices of the Royal Astronomical Society 421(2), 1155–1165 (2012)

Brewer, E.A.: CAP twelve years later: how the 'rules' have changed. Computer 45(2), 23–29 (2012)

Cardamone, C., Schawinski, K., Sarzi, M., et al.: Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. Monthly Notices of the Royal Astronomical Society 399(3), 1191–1205 (2009), doi:10.1111/j.1365-2966.2009.15383.x

Cattell, R.: Scalable SQL and NoSQL Data Stores. SIGMOD Record 39(4), 12–27 (2010)

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., Paolillo, M.: Photometric classification of emission line galaxies with Machine Learning methods. Monthly Notices of the Royal Astronomical Society 437(1), 968–975 (2014)

Cavuoti, S., Garofalo, M., Brescia, M., et al.: Astrophysical data mining with GPU. A case study: genetic classification of globular clusters. New Astronomy 26, 12–22 (2014)

D'Abrusco, R., Longo, G., Walton, N.A.: Quasar candidates selection in the Virtual Observatory era. Monthly Notices of the Royal Astronomical Society 396(1), 223–262 (2009)

Darwin, C.: On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 1st edn. John Murray, London (1859)

Dean, D., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM 51(1), 107–113 (2008)

Djorgovski, S.G., Baltay, C., Mahabal, A.A., et al.: The Palomar-Quest digital synoptic sky survey. Astron. Nachr. 329(3), 263–265 (2008)

Dorogov, A.Y.: Structural synthesis of fast two-layer neural networks. Cybernetics and Systems Analysis 36(4), 512–519 (2000)

Drake, A.J., Djorgovski, S.G., Mahabal, A., et al.: First Results from the Catalina Real-time Transient Survey. Astrophys. Journal 696, 870–884 (2009)

Flockhart, I.W., Radcliffe, N.J.: A Genetic Algorithm-Based Approach to Data Mining. In: Proceedings of 2nd Int. Conf. AAAI: Knowledge Discovery and Data Mining, Portland, Oregon, pp. 299–302 (1996)

Fogel, L., Owens, J., Walsh, J.: Artificial Intelligence through Simulated Evolution. John Wiley, Chichester (1966)

Gainaru, A., Slusanschi, E., Trausan-Matu, S.: Mapping data mining algorithms on a GPU architecture: A study. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS, vol. 6804, pp. 102–112. Springer, Heidelberg (2011)

Gamble, M., Goble, C.: Quality, Trust and Utility of Scientific Data on the Web: Towards a Joint model. In: Proceedings of ACM WebSci 2011 Conference, Koblenz, Germany, 8 p. (2011)

Gartner, Inc., Pattern-Based Strategy: Getting Value from Big Data. Gartner Group (2011), http://www.gartner.com/it/page.jsp?id=1731916 (accessed May 30, 2014)

Ghemawat, S., Gobioff, H., Leung, S.-L.: The Google File System. ACM SIGOPS Operating Systems Review 37(5), 29–43 (2003)

Härder, T., Reuter, A.: Concepts for Implementing and Centralized Database Management System. In: Proceedings of Int. Computing Symposium on Application Systems Development, Nürnberg, Germany, B.G., pp. 28–104 (1983)

Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2010)

Holland, J.: Adaptation in natural and artificial systems. Univ. of Michigan Press, Ann Arbor (1975)

Hwu, W., Keutzer, K., Mattson, T.G.: The Concurrency Challenge. IEEE Des. Test of Computers 25(4), 312–320 (2008)

Johnson, C.: Artificial immune systems programming for symbolic regression. In: Ryan, C., Soule, T., Keijzer, M., Tsang, E., Poli, R., Costa, E. (eds.) EuroGP 2003. LNCS, vol. 2610, pp. 345–353. Springer, Heidelberg (2003)

Kaiser, N.: The Pan-STARRS Survey Telescope Project. In: Advanced Maui Optical and Space Surveillance Technologies Conference (2007)

Kaiser, N., Burgett, W., Chambers, K., et al.: The pan-STARRS wide-fieldoptical/NIR imaging survey. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 7733, p. 12 (2010)

Keutzer, K., Mattson, T.G.: A Design Pattern Language for Engineering (Parallel) Software. Addressing the Challenges of Tera-scale Computing. Intel Technology Journal 13(04), 6–19 (2008)

Khabzaoui, M., Dhaenens, C., Talbi, E.G.: Combining Evolutionary Algorithms and Exact Approaches for Multi-Objective Knowledge Discovery. Rairo-Oper. Res. 42, 69–83 (2008), doi:10.1051/ro:2008004

Khan, M.F., Paul, R., Ahmed, I., Ghafoor, A.: Intensive data management in parallel systems: A survey. Distributed and Parallel Databases 7(4), 383–414 (1999)

Koza, J.: Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford University, Computer Science Department, Technical Report STAN-CS-90-1314 (1990)

Koza, J.: Genetic programming. MIT Press (1998)

Koza, J.R., Bennett, F.H., Andre, D., Keane, M.A.: Genetic Programming III; Darwinian Invention and problem Solving. Morgan Kaufmann Publisher (1999)

Koza, J., Keane, M., Streeter, M.: Evolving inventions. Scientific American 288(2), 52–59 (2003)

Laurino, O., D'Abrusco, R., Longo, G., Riccio, G.: Monthly Notices of the Royal Astronomical Society 418, 2165–2195 (2011)

Lintott, C.J., Lintott, C., Schawinski, K., Keel, W., et al.: Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo? Monthly Notices of Royal Astronomical Society 399(1), 129–140 (2009)

Lloyd, S., Giovannetti, V., Maccone, L.: Physical limits to communication. Phys. Rev. Lett. 93, 100501 (2004)

Mahabal, A., Djorgovski, S.G., Donalek, C., Drake, A., Graham, M., Williams, R., Moghaddam, B., Turmon, M.: Classification of Optical Transients: Experiences from PQ and CRTS Surveys. In: Turon, C., Arenou, F., Meynadier, F. (eds.) Gaia: At the Frontiers of Astrometry. EAS Publ. Ser., vol. 45, EDP Sciences, Paris (2010)

Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer (2010)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Inst. (2011)

Mellier, Y., Laureijs, R., Amiaux, J., et al.: EUCLID definition study report (Euclid Red Book). European Space Agency (2011), http://sci.esa.int/euclid/48983-euclid-definition-study-report-esa-sre-2011-12 (accessed May 30, 2014)

Mendel, J.: Versuche uber Pflanzenhybriden Verhandlungen des naturforschenden Vereines in Brunn. Bd. IV fur das Jahr. Abhandlungen, 3–47 (1865); For the English translation, see: Druery, C.T., Bateson, W.: Experiments in plant hybridization. Journal of the Royal Horticultural Society 26, 1–32 (1901), http://www.esp.org/foundations/genetics/classical/gm-65.pdf (accessed May 30, 2014)

Morgan, T.P.: IDC: Big data biz worth $16.9 BILLION by 2015. The Register (2012)

Mueller, R., Teubner, J., Alonso, G.: Data processing on FPGAs. Proc. VLDB Endow. 2(1), 910–921 (2009)

O'Neill, M., Brabazon, A.: Grammatical differential evolution. In: Proceedings of International Conference on Artificial Intelligence, pp. 231–236. CSEA Press (2006)

O'Neill, M., Ryan, C.: Grammatical Evolution, Evolutionary Automatic Programming in an Arbitrary Language. Springer, New York (2003)

Oplatkova, Z.: Optimal trajectory of robots using symbolic regression. In: Proceedings of 56th International Astronautics Congress, Fukuoka, Japan (2005)

Oplatkova, Z.: Metaevolution: Synthesis of Optimization Algorithms by means of Symbolic Regression and Evolutionary Algorithms. Lambert Academic Publishing, New York (2009)

Oplatkova, Z., Zelinka, I.: Investigation on artificial ant using analytic programming. In: Proceedings of Genetic and Evolutionary Computation Conference, Seattle, WA, pp. 949–950 (2006)

Oplatkova, Z., Senkerik, R., Belaskova, S., Zelinka, I.: Synthesis of control rule for synthesized chaotic system by means of evolutionary techniques. In: Proceedings of 16th International Conference on Soft Computing Mendel 2010, Technical university of Brno, Brno, Czech Republic, pp. 91–98 (2010)

Oplatkova, Z., Senkerik, R., Zelinka, I., Holoska, J.: Synthesis of control law for chaotic Henon system - preliminary study. In: Proceedings of 24th European Conference on Modelling and Simulation, ECMS 2010, Kuala Lumpur, Malaysia, pp. 277–282 (2010)

Oplatkova, Z., Senkerik, R., Zelinka, I., Holoska, J.: Synthesis of control law for chaotic logistic equation - preliminary study. In: IEEE Proceedings of AMS 2010, ASM, Kota Kinabalu, Borneo, Malaysia, pp. 65–70 (2010)

Perryman, M.A.C.: Overview of the Gaia Mission. In: Proceedings of the Three-Dimensional Universe with Gaia, ESA SP-576, p. 15 (2005)

Pokorny, J.: NoSQL Databases: a step to databases scalability in Web environment. International Journal of Web Information Systems 9(1), 69–82 (2013)

Quinn, P., Lawrence, A., Hanisch, R.: The Management, Storage and Utilization of Astronomical Data in the 21st Century, IVOA Note (2004), http://www.ivoa.net/documents/latest/OECDWhitePaper.html (accessed May 30, 2014)

Raddick, J.M., Bracey, G., Gay, P.L., Lintott, C.J., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J.: Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. Astronomy Education Review 9(1), 010103 (2010)

Rajaraman, A., Leskovec, J., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (2013)

Rechenberg, I.: Evolutionsstrategie - Optimierung technischer Systeme nach Prin-zipien der biologischen Evolution. PhD thesis, Printed in Fromman-Holzboog (1973)

Ryan, C., Collins, J.J., O'Neill, M.: Grammatical evolution: Evolving programs for an arbitrary language. In: Banzhaf, W., Poli, R., Schoenauer, M., Fogarty, T.C. (eds.) EuroGP 1998. LNCS, vol. 1391, pp. 83–95. Springer, Heidelberg (1998)

Schwefel, H.: Numerische Optimierung von Computer-Modellen, PhD thesis (1974), reprinted by Birkhauser (1977)

Strauch, C.: NoSQL Databases. Lecture Selected Topics on Software-Technology Ultra-Large Scale Sites, Stuttgart Media University, manuscript (2011), http://www.christof-strauch.de/nosqldbs.pdf (accessed May 30, 2014)

Szalay, A., Gray, J.: The World Wide Telescope. Science 293, 2037–2040 (2001)

Szalay, A.S., Gray, J., van den Berg, J.: Petabyte scale data mining: Dream or reality? In: SPIE Conference Proceedings, vol. 4836, p. 333 (2002), doi:10.1117/12.461427

Tan, K.C., Teoh, E.J., Yu, Q., Goh, K.C.: A hybrid evolutionary algorithm for at-tribute selection in data mining. Expert Systems with Applications 36, 8616–8630 (2009)

van Haarlem, M.P., Wise, M.W., Gunst, A.W., et al.: LOFAR: The LOw-Frequency Array. Astronomy and Astrophysics 556(A2), 53 (2013)

Vinayak, R., Borkar, V., Carey, M.-J., Chen Li, C.: Big data platforms: what's next? ACM Cross Road 19(1), 44–49 (2012)

Weisser, R., Osmera, P.: Two-level transplant evolution. In: Proceedings of 17th Zittau Fuzzy Colloquium, Zittau, Germany, pp. 63–70 (2010)

Weisser, R., Osmera, P.: Two-level transplant evolution for optimization of general controllers. In: New Trends in Technologies, Devices, Computer, Communication and Industrial Systems, pp. 55–68. Sciyo (2010)

Weisser, R., Osmera, P., Matousek, R.: Transplant evolution with modified schema of differential evolution: Optimization structure of controllers. In: Proceedings of 16th International Conference on Soft Computing MENDEL, Brno, Czech Republic, pp. 113–120 (2010)

Yadav, C., Wang, S., Kumar, M.: Algorithm and approaches to handle large Data - A Survey. IJCSN International Journal of Computer Science and Network 2(3), 37–41 (2013)

Zelinka, I., Guanrong, C., Celikovsky, S.: Chaos synthesis by means of evolutionary algorithms. International Journal of Bifurcation and Chaos 18(4), 911–942 (2008)

Zelinka, I.: Analytic programming by means of new evolutionary algorithms. In: Proceedings of 1st International Conference on New Trends in Physics 2001, Brno, Czech Republic, pp. 210–214 (2001)

Zelinka, I.: Analytic programming by means of soma algorithm. In: Proceedings of First International Conference on Intelligent Computing and Information Systems, Cairo, Egypt, pp. 148–154 (2002)

Zelinka, I., Oplatkova, Z.: Analytic programming – comparative study. In: Proceedings of Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems, Singapore (2003)

Zelinka, I., Oplatkova, Z., Nolle, L.: Analytic programming – symbolic regression by means of arbitrary evolutionary algorithms. Int. J. of Simulation, Systems, Science and Technology 6(9), 44–56 (2005)

Zelinka, I., Skanderova, L., Saloun, P., Senkerik, R., Pluhacek, M.: Chaos Powered Symbolic Regression in Be Stars Spectra Modeling. In: Proceedings of the ISCS 2013, Praha, pp. 131–139. Springer (2014)

Zelinka, I., Celikovsky, S., Richter, H., Chen, G. (eds.): Evolutionary Algorithms and Chaotic Systems. SCI, vol. 267. Springer, Heidelberg (2010)

Zelinka, I., Davendra, D., Senkerik, R., Jasek, R., Oplatkova, Z.: Analytical Program-ming - a Novel Approach for Evolutionary Synthesis of Symbolic Structures. In: Kita, E. (ed.) Evolutionary Algorithms, pp. 149–176. InTech (2011), doi:10.5772/16166

Zhang, Y., Zheng, H., Zhao, Y.: Knowledge discovery in astronomical data. In: SPIE Conference Proceedings, vol. 701938, p. 108 (2008), doi:10.1117/12.788417

Zhao, Y., Raicu, I., Foster, I.: Scientific workflow systems for 21st century, new bottle or new wine? In: Proceedings of IEEE Congress on Services - Part I, pp. 467–471 (2008)

Zhao, G., Zhao, Y., Chu, Y., Jing, Y., Deng, L.: LAMOST Spectral Survey. Research in Astron. Astrophys. 12(7), 723–734 (2012)

Towards Robust Performance Guarantees for Models Learned from High-Dimensional Data

Rui Henriques and Sara C. Madeira

Abstract. Models learned from high-dimensional spaces, where the high number of features can exceed the number of observations, are susceptible to overfit since the selection of subspaces of interest for the learning task is prone to occur by chance. In these spaces, the performance of models is commonly highly variable and dependent on the target error estimators, data regularities and model properties. High-variable performance is a common problem in the analysis of omics data, healthcare data, collaborative filtering data, and datasets composed by features extracted from unstructured data or mapped from multi-dimensional databases. In these contexts, assessing the statistical significance of the performance guarantees of models learned from these high-dimensional spaces is critical to validate and weight the increasingly available scientific statements derived from the behavior of these models. Therefore, this chapter surveys the challenges and opportunities of evaluating models learned from big data settings from the less-studied angle of big dimensionality. In particular, we propose a methodology to bound and compare the performance of multiple models. First, a set of prominent challenges is synthesized. Second, a set of principles is proposed to answer the identified challenges. These principles provide a roadmap with decisions to: *i)* select adequate statistical tests, loss functions and sampling schema, *ii)* infer performance guarantees from multiple settings, including varying data regularities and learning parameterizations, and *iii)* guarantee its applicability for different types of models, including classification and descriptive models. To our knowledge, this work is the first attempt to provide a robust and flexible assessment of distinct types of models sensitive to both the dimensionality and size of data. Empirical evidence supports the relevance of these principles as they offer a coherent setting to bound and compare the performance of models learned in high-dimensional spaces, and to study and refine the behavior of these models.

Keywords: high-dimensional data, performance guarantees, statistical significance of learning models, error estimators, classification, biclustering.

Rui Henriques · Sara C. Madeira
KDBIO, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
e-mail: {rmch,sara.madeira}@tecnico.ulisboa.pt

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_3

1 Introduction

High-dimensional data has been increasingly used to derive implications from the analysis of biomedical data, social networks or multi-dimensional databases. In high-dimensional spaces, it is critical to guarantee that the learned relations are statistically significant, that is, they are not learned by chance. This is particularly important when these relations are learned from subspaces of the original space and when the number of observations is not substantially larger than the number of features. Examples of data where the number of observations/instances is either lower or not significantly higher than the number of features include: collaborative filtering data, omics data (such as gene expression data, structural genomic variations and biological networks), clinical data (such as data integrated from health records, functional magnetic resonances and physiological signals), and random fields (Amaratunga, Cabrera, and Shkedy 2014). In order to bound or compare the performance of models composed by multiple relations, the impact of learning in these high-dimensional spaces on the statistical assessment of these models needs to be properly considered.

Despite the large number of efforts to study the effects of dimensionality and data size (number of instances) on the performance of learning models (Kanal and Chandrasekaran 1971; Jain and Chandrasekaran 1982; Raudys and Jain 1991; Adcock 1997; Vapnik 1998; Mukherjee et al. 2003; Hua et al. 2005; Dobbin and Simon 2007; Way et al. 2010; Guo et al. 2010), an integrative view of their potentialities and limitations is still lacking. In this chapter, we identify a set of major requirements to assess the performance guarantees of models learned from high-dimensional spaces and survey critical principles for their adequate satisfaction. These principles can also be used to affect the learning methods and to estimate the minimum sample size that guarantees the inference of statistically significant relations.

Some of the most prominent challenges for this task are the following. First, assessing the performance of models based on simulated surfaces and on fitted learning curves often fail to provide robust statistical guarantees. Typically under these settings, the significance of the estimations is tested against loose models learned from permuted data and the performance guarantees are not affected by the variability of the observed errors (Mukherjee et al. 2003; Way et al. 2010). Second, many of the existing assessments assume independence among features (Dobbin and Simon 2005; Hua et al. 2005). This assumption does not hold for datasets in high-dimensional spaces where the values of few features can discriminate classes by chance. This is the reason why learning methods that rely on subsets of the original features, such as rule-based classifiers, have higher variance on the observed errors, degrading the target performance bounds. Third, error estimators are often inadequate since the common loss functions for modeling the error are inappropriate and the impact of test sample size is poorly studied leading to the collection error estimates without statistical significance (Beleites et al. 2013). Fourth, assessment methods from synthetic data commonly rely on simplistic data distributions, such as multivariate Gaussian class-conditional distributions (Dobbin and Simon 2007). However, features in real-world data (biomedical features

such as proteins, metabolites, genes, physiological features, etc.) exhibit highly skewed mixtures of distributions (Guo et al. 2010). Finally, existing methods are hardly extensible towards more flexible settings, such as the performance evaluation of descriptive models (focus on a single class) and of classification models in the presence of multiple and unbalanced classes.

In this context, it is critical to define principles that are able to address these drawbacks. In this chapter, we rely on existing contributions and on additional empirical evidence to derive these structural principles. Additionally, their integration through a new methodology is discussed. Understandably, even in the presence of datasets with identical sample size and dimensionality, the performance is highly dependent on data regularities and learning setting, as they affect the underlying significance and composition of the learned relations. Thus, the proposed methodology is intended to be able to establish both data-independent and data-dependent assessments. Additionally, it is suitable for distinct learning tasks in datasets with either single or multiple classes. Illustrative tasks include classification of tumor samples, prediction of healthcare needs, biclustering of genes, proteomic mass spectral classification, chemosensitivity prediction, or survival analysis.

The proposed assessment methodology offers three new critical contributions to the big data community:

- Integration of statistical principles to provide a solid foundation for the definition of robust estimators of the true performance of models learned in high-dimensional spaces, including adequate loss functions, sampling schema (or parametric estimators), statistical tests and strategies to adjust performance guarantees in the presence of high variance and bias of performance;
- Inference of general performance guarantees for models tested over multiple high-dimensional datasets with varying regularities;
- Applicability for different types of models, including classification models with class-imbalance, regression models, local or (bi)clustering models and global descriptive models.

This chapter is organized as follows. In what follows, we provide the background required to define and understand the target task – assessing models learned from high-dimensional spaces. *Section 2* surveys research streams with important contributions for this task, covering their major challenges. *Section 3* introduces a set of key principles derived from existing contributions to address the identified challenges. These are then coherently integrated within a simplistic assessment methodology. *Section 4* discusses the relevance of these principles based on experimental results and existing literature. Finally, concluding remarks and future research directions are synthesized.

1.1 Problem Definition

Consider a dataset described by n pairs (x_i, y_i) from (X, Y), where $x_i \in \mathbb{R}^m$ and Y is either described by a set of labels $y_i \in \Sigma$ or numeric values $y_i \in \mathbb{R}$. A space described by $n \in \mathbb{N}$ observations and $m \in \mathbb{N}$ features is here referred as a (n, m)-space, $X^{n,m} \subseteq X$.

Assuming data is characterized by a set of underlying stochastic regularities, $P_{X|Y}$, a *learning task* aims to infer a model M from a (n,m)-space such that the error over $P_{X|Y}$ is minimized. The M model is a composition of relations (or abstractions) from the underlying stochastic regularities.

Under this setting, two major types of models can be considered. First, *supervised models*, including classification models ($M : X \rightarrow Y$, where Y=Σ is a set of categoric values) and regression models ($M : X \rightarrow Y$, with Y=$\mathbb{R}$), focus on the discriminative aspects of the conditional regularities $P_{X|Y}$ and their error is assessed recurring to loss functions (Toussaint 1974). Loss functions are typically based on accuracy, area under *roc*-curve or sensitivity metrics for classification models, and on the normalized or root mean squared errors for regression models. In supervised settings, there are two major types of learning paradigms with impact on the assessment of performance: *i)* learning a relation from all features, including multivariate learners based on discriminant functions (Ness and Simpson 1976), and *ii)* learning a composition of relations inferred from specific subspaces $X^{q,p} \subseteq X^{n,m}$ of interest (e.g. rule-based learners such as decision trees and Bayesian networks). For the latter case, capturing the statistical impact of feature selection is critical since small subspaces are highly prone to be discriminative by chance (Iswandy and Koenig 2006). To further clarify the impact of dimensionality when assessing the performance of these models, consider a subset of the original features, $X^{n,p} \subseteq X^{n,m}$, and a specific class or real interval, $y \in Y$. Assuming that these discriminative models can be decomposed in mapping functions of the type $M : X^{n,p} \rightarrow y$, comparing or bounding the performance of these models needs to consider the fact that the (n,p)-space is not selected aleatory. Instead, this subspace is selected as a consequence of an improved discriminatory power. In high-dimensional spaces, it is highly probable that a small subset of the original features is able to discriminate a class by chance. When the statistical assessment is based on error estimates, there is a resulting high-variability of values across estimates that needs to be considered. When the statistical assessment is derived from the properties of the model, the effect of mapping the original (n,m)-space into a (n,p)-space needs to be consider.

Second, *descriptive models* ($|Y|$=1) either globally or locally approximate P_X regularities. Mixtures of multivariate distributions are often used as global descriptors, while (bi)clustering models define local descritpors. The error is here measured either recurring to merit functions or match scores when there is knowledge regarding the underlying regularities. In particular, a local descriptive model is a composition of learned relations from subspaces of features J=$X^{n,p} \subseteq X^{n,m}$, samples I=$X^{q,m} \subseteq X^{n,m}$, or both (I,J). Thus, local models define a set of k (bi)clusters such that each (bi)cluster (I_k,J_k) satisfies specific criteria of homogeneity. Similarly to supervised models, it is important to guarantee a robust collection and assessment of error estimates or, alternatively, that the selection of the (q_k,p_k)-space of each (bi)cluster (where q_k=$|I_k|$ and p_k=$|J_k|$) is statistical significant, that is, the observed homogeneity levels for these subspaces do not occur by chance.

Consider that the asymptotic probability of misclassification of a particular model M is given by ε_{true}, and a non-biased estimator of the observed error in a (n,m)-space is given by $\theta(\varepsilon_{true})$. The problem of computing the *performance guarantees*

for a specific model M in a (n,m)-space can either be given by its performance bounds or by its ability to perform better than other models. The task of computing the $(\varepsilon_{min}, \varepsilon_{max})$ *performance bounds* for a M model in a (n,m)-space can be defined as:

$$[\varepsilon_{min}, \varepsilon_{max}] : P(\varepsilon_{min} < \theta(\varepsilon_{true}) < \varepsilon_{max} \mid n, m, M, P_{X|Y}) = 1 - \delta, \quad (1)$$

where the performance bounds are intervals of confidence tested with 1-δ statistical power.

In this context, the task of *comparing a set of models* $\{M_1, .., M_l\}$ in a (n,m)-space can be defined as the discovery of significant differences in performance between groups of models while controlling the family-wise error, the probability of making one or more false comparisons among all the $l \times l$ comparisons. Defining an adequate estimator of the true error $\theta(\varepsilon_{true})$ for a target $(n, m, M, P_{X|Y})$ setting is, thus, the central role of these assessments.

In literature, similar attempts have been made for testing the minimum number of observations, by comparing the estimated error for n observations with the true error, $min_n : P(\theta_n(\varepsilon_{true}) < \varepsilon_{true} \mid m, M, P_{X|Y}) > 1\text{-}\delta$ rejected at α, or by allowing relaxation factors $\theta_n(\varepsilon_{true}) < (1+\gamma)\varepsilon_{true}$ when the observed error does not rapidly converge to ε_{true}, $\lim_{n\to\infty} \theta_n(\varepsilon_{true}) \neq \varepsilon_{true}$. In this context, the ε_{true} can be theoretically derived from assumptions regarding the regularity $P_{X|Y}$ or experimentally approximated using the asymptotic behavior of learning curves estimated from data.

To illustrate the relevance of target performance bounding and comparison tasks, let us consider the following model: a linear hyperplane $M(x)$ in $\mathbb{R}^m$ defined by a vector w and point b to either separate two classes, $sign(w \cdot x + b)$, predict a real-value, $w \cdot x + b$, or globally describe the observations, $X \sim w \cdot x + b$. In contexts where the number of features exceeds the number of observations ($m > n$), these models are not able to generalize (perfect overfit towards data). As illustrated in Fig.1, a linear hyperplane in $\mathbb{R}^m$ can perfectly model up to $m+1$ observations, either as classifier $X \to \{\pm 1\}$, as regression $X \to \mathbb{R}$ or as descriptor of X. Thus, a simple assessment of the errors of these models using the same training data would lead to $\theta(\varepsilon_{true})$=0 without variance across estimates ε_i and, consequently, to ε_{min}=ε_{max}=0, which may not be true in the presence of an additional number of observations.

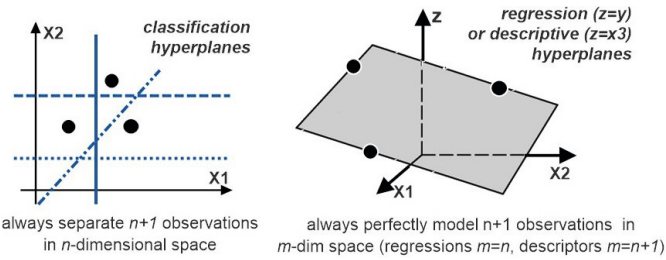

Fig. 1 Linear hyperplanes cannot generalize when dimensionality is larger than the number of observations (data size), $m \geq n+1$

Moreover, the performance of these models using new testing observations tends to be high-variable. These observations should be considered when selecting the assessment procedure, including the true error estimator $\theta(\varepsilon_{true})$, the statistical tests and the assumptions underlying data and the learning method.

2 Related Work

Classic statistical methods to *bound the performance* of models as a function of the data size include power calculations based on frequentist and Bayesian methods (Adcock 1997), deviation bounds (Guyon et al. 1998), asymptotic estimates of the true error ε_{true} (Raudys and Jain 1991; Niyogi and Girosi 1996), among others (Jain and Chandrasekaran 1982). Here, the impact of the data size in the observed errors is essentially dependent on the entropy associated with the target (n,m)-space. When the goal is the *comparison of multiple models*, Wilcoxon signed ranks test (two models) and the Friedman test with the corresponding post-hoc tests (more than two models) are still state-of-the-art methods to derive comparisons either from error estimates or from the performance distributions given by classic statistical methods (Demšar 2006; García and Herrera 2009).

To generalize the assessment of performance guarantees for an unknown sample size n, learning curves (Mukherjee et al. 2003; Figueroa et al. 2012), theoretical analysis (Vapnik 1998; Apolloni and Gentile 1998) and simulation studies (Hua et al. 2005; Way et al. 2010) have been proposed. A critical problem with these latter approaches is that they either ignore the role of dimensionality in the statistical assessment or the impact of learning from subsets of overall features.

We grouped these existing efforts according to six major streams of research: *1)* classic statistics, *2)* risk minimization theory, *3)* learning curves, *4)* simulation studies, *5)* mutivariate model's analysis, and *6)* data-driven analysis. Existing approaches have their roots on, at least, one of these research streams, which assess the performance significance of a single learning model as a function of the available data size, a key factor when learning from high-dimensional spaces. Understandably, comparing multiple models is a matter of defining robust statistical tests from the assessed performance per model.

First, *classic statistics* covers a wide-range of methods. They are either centered on power calculations (Adcock 1997) or on the asymptotic estimates of ε_{true} by using approximation theory, information theory and statistical mechanics (Raudys and Jain 1991; Opper et al. 1990; Niyogi and Girosi 1996). Power calculations provide a critical view on the model errors (performance) by controlling both sample size n and statistical power 1-γ, $P(\theta_n(\varepsilon_{true}) < \varepsilon_{true})$=1-$\gamma$, where $\theta_n(\varepsilon_{true})$ can either rely on a frequentist view, from counts to estimate the discriminative/descriptive ability of subsets of features, or on a Bayesian view, more prone to deal with smaller and noisy data (Adcock 1997).

Second, *theoretical analysis of empirical risk minimization* (Vapnik 1982; Apolloni and Gentile 1998) studies the performance of a model by analyzing the trade-off between the model capacity and the observed error. To further understand the

concept of risk minimization, consider the two following models: one simplistic model that achieves a good generalization (high model capacity) but has a high observed error, and a model able to minimize the observed error but overfitted to the available data. The trade-off analysis for the minimization of these errors is illustrated in Fig.2. Core contributions from this research stream comes from Vapnik-Chervonenkis (VC) theory (Vapnik 1998), where the sample size and the dimensionality is related through the VC-dimension (h), a measure of the model capacity that defines the minimum number of observations required to generalize the learning in a m-dimensional space. As illustrated in Fig.1, linear hyperplanes have $h = m+1$. The VC-dimension can be theoretically or experimentally estimated for different models and used to compare the performance of models and approximate lower-bounds. Although the target overfitting problem is addressed under this stream, the resulting assessment tends to be conservative.

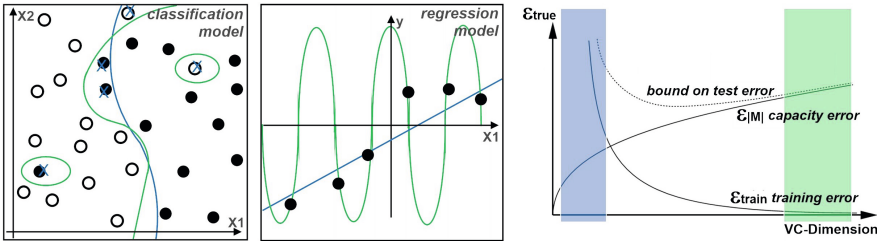

Fig. 2 Capacity and training error impact on true error estimation for classification and regression models

Third, *learning curves* use the observed performance of a model over a given dataset to fit inverse power-law functions that can extrapolate performance bounds as a function of the sample size or dimensionality (Mukherjee et al. 2003; Boonyanunta and Zeephongsekul 2004). An extension that weights estimations according to their confidence has been applied for medical data (Figueroa et al. 2012). However, the estimation of learning curves in high-dimensional spaces requires large data ($n>m$), which are not always available, and does not consider the variability across error estimates.

Fourth, *simulation studies* infer performance guarantees by studying the impact of multiple parameters on the learning performance (Hua et al. 2005; Way et al. 2010; Guo et al. 2010). This is commonly accomplished through the use of a large number of synthetic datasets with varying properties. Statistical assessment and inference over the collected results can be absent. This is a typical case when the simulation study simply aims to assess major variations of performance across settings.

Fifth, true performance estimators can be derived from a *direct analysis of the learning models* (Ness and Simpson 1976; El-Sheikh and Wacker 1980; Raudys and Jain 1991), mainly driven by the assessment of multivariate models that preserve the dimensionality of space (whether described by the original m features or by a subset

of the original features after feature selection) when specific regularities underlying data are assumed. Illustrative models include classifiers based on discriminant functions, such as Euclidean, Fisher, Quadratic or Multinomial. Unlike learning models based on tests over subsets of features selected from the original high-dimensional space, multivariate learners consider the values of all features. Despite the large attention given by the multivariate analysis community, these models only represent a small subset of the overall learning models (Cai and Shen 2010).

Finally, model-independent size decisions derived from data regularities are reviewed and extended by Dobbin and Simon (2005; 2007). *Data-driven formulas* are defined from a set of approximations and assumptions based on dimensionality, class prevalence, standardized fold change, and on the modeling of non-trivial sources of errors. Although dimensionality is used to affect both the testing significance levels and the minimum number of features (i.e., the impact of selecting subspaces is considered), the formulas are independent from the selected models, forbidding their extension for comparisons or the computation of performance bounds.

These six research streams are closely related and can be mapped through concepts of information theory. In fact, an initial attempt to bridge contributions from statistical physics, approximation theory, multivariate analysis and VC theory within a Bayesian framework was proposed by Haussler, Kearns, and Schapire (1991).

2.1 Challenges and Contributions

Although each of the introduced research streams offers unique perspectives to solve the target task, they suffer from drawbacks as they were originally developed with a different goal – either minimum data size estimation or performance assessments in spaces where $n \gg m$. These drawbacks are either related with the underlying approximations, with the assessment of the impact of selecting subspaces (often related with a non-adequate analysis of the variance of the observed errors) or with the poor extensibility of existing approaches towards distinct types of models or flexible data settings. Table 1 details these drawbacks according to three major categories that define the ability to: *A)* rely on robust statistical assessments, *B)* deliver performance guarantees from multiple flexible data settings, and *C)* extend the target assessment towards descriptive models, unbalanced data, and multi-parameter settings. The latter two categories trigger the additional challenge of inferring performance guarantees from multiple settings where data regularities and model parameters are varied.

Since each one of the introduced streams of research were developed with a specific goal and under a large set of assumptions, it is natural that their limitations fall into several of the identified categories in Table 1. In Table 2, we identify the major problems of these streams of research when answering the target tasks.

Table 1 Common challenges when defining performance guarantees of models learned from high-dimensional data

Category	Problem Description
A. Statistical Robustness	1. Non-robust estimators of the true performance of models. First, the probability of selecting informative features by chance is higher in high-dimensional spaces, leading to an heightened variability of error estimates and, in some cases, turning inviable the inference of performance guarantees. Second, when the number of features exceeds the number of observations, errors are prone to systemic biases. The simple use of mean and deviation metrics from error estimates to compare and bound the performance is insufficient in these spaces; 2. Inappropriate sampling scheme for the collection of error estimates in high-dimensional spaces (Beleites et al. 2013). Assessing the variance of estimations within and across folds, and the impact of the number of folds and test sample size is critical to tune the level of conservatism of performance guarantees; 3. Inadequate loss functions to characterize the observed error. Examples of loss functions of interest that are commonly ignored include sensitivity for unbalanced classification settings (often preferred against accuracy) or functions that provide a decomposition of errors; 4. Inadequate underlying density functions to test the significance of error estimates. Significance is typically assessed against very loose null settings (Mukherjee et al. 2003), and rarely assessed over more meaningful settings. Additionally, many of the proposed estimators are biased (Hua et al. 2005); 5. Others: approximated and asymptotic error estimators derived from multivariate model analysis (Raudys and Jain 1991) are only applicable for a specific subset of learning models; model-independent methods, such as formulae-based methods for minimum size estimation (Dobbin and Simon 2005), are non-extensible to compare models or bound performance; performance guarantees provided by a theoretical analysis of the learning properties, such as in VC-theory (Vapnik 1982), tend to be very conservative; dependency of large datasets to collect feasible estimates (Mukherjee et al. 2003).
B. Data Flexibility	1. Performance guarantees are commonly only assessed in the context of a specific dataset (e.g. classic statistics, learning curves), and, therefore, the implied performance observations cannot be generalized; 2. Performance comparisons and bounds are computed without assessing the regularities underlying the inputted data (Guo et al. 2010). These regularities provide a context to understand the learning challenges of the task and, thus, providing a frame to assess the significance of the scientific implications; 3. Contrasting with data size, dimensionality is rarely considered a variable to compare and bound models' performance (Jain and Chandrasekaran 1982). Note that dimensionality m and performance $\theta(\varepsilon_{true})$ are co-dependent variables as it is well-demonstrated by the VC theory (Vapnik 1998); 4. Independence among features is assumed in some statistical assessments. However, most of biomedical features (such as molecular units) and extracted features from collaborative data are functionally correlated; 5. Non-realistic synthetic data settings. Generated datasets should follow properties of real datasets, which are characterized by mixtures of distributions with local dependencies, skewed features and varying levels of noise; 6. The impact of modeling additional sources of variability, such as pooling, dye-swap samples and technical replicates for biomedical settings, is commonly disregarded (Dobbin and Simon 2005).
C. Extensibility	1. Inadequate statistical assessment of models learned from datasets with heightened unbalance among classes and non-trivial conditional distributions $P_{X\|y_i}$; 2. Weaker guidance for computing bounds for multi-class models ($\|\Sigma\| > 2$); 3. Existing methods are not extensible to assess the performance bounds of descriptive models, including (single-class) global and local descriptive models; 4. Lack of criteria to establish performance guarantees from settings where the impact of numerous parameters is studied (Hua et al. 2005; Way et al. 2010).

Table 2 Limitations of existing approaches according to the introduced challenges

Approach	Major Problems (non-exhaustive observations)
Bayesian & Frequentist Estimations	Originally proposed for the estimation of minimum data size and, thus, not prepared to deliver performance guarantees; applied in the context of a single dataset; impact of feature selection is not assessed; no support as-is for descriptive tasks and hard data settings;
Theoretical Methods	Delivery of worst-case performance guarantees; learning aspects need to be carefully modeled (complexity); guarantees are typically independent from data regularities (only the size and dimensionality of the space are considered); no support as-is for descriptive tasks and hard data settings.
Learning Curves	Unfeasible for small datasets or high-dimensional spaces where $m>n$; dimensionality and the variability of errors does not explicitly affect the curves; guarantees suitable for a single input dataset; no support as-is for descriptive tasks and hard data settings.
Simulation Studies	Driven by error minimization and not by the statistical significance of performance; data often rely on simplistic conditional regularities (optimistic data settings); poor guidance to derive decisions from results.
Multivariate Analysis	Limited to multivariate models from discriminant functions; different models require different parametric analyzes; data often rely on simplistic conditional regularities; no support as-is for descriptive tasks and hard data settings; approximations can lead to loose bounds;
Data-driven Formula	Not able to deliver performance guarantees (model-independent); estimations only robust for specific data settings; Independence among features is assumed; suitable for a single inputted dataset; unfeasible for small samples.

Although the surveyed challenges are lengthy in number, many of them can be answered recurring to contributions provided in literature. In Table 3, we describe illustrative sets of contributions that can be used to satisfy the requirements derived from the identified limitations. This analysis triggers the need to isolate sets of principles to address each challenge, and to study whether it is possible to integrate these dispersed contributions for the development of more robust comparisons and bound estimation procedures.

Table 3 Contributions with potential to satisfy the target set of requirements

Requirements	Contributions
Guarantees from High-Variable Performance (A.1)	Statistical tests to bound and compare performance sensitive to error distributions and loss functions (Martin and Hirschberg 1996; Qin and Hotilovac 2008; Demšar 2006); VC theory and discriminant-analysis (Vapnik 1982; Raudys and Jain 1991); unbiasedness principles from feature selection (Singhi and Liu 2006; Iswandy and Koenig 2006).
Bias Effect (A.1)	Bias-Variance decomposition of the error (Domingos 2000).
Adequate Sampling Schema (A.2)	Criteria for sampling decisions (Dougherty et al. 2010; Toussaint 1974); test-train splitting impact (Beleites et al. 2013; Raudys and Jain 1991).
Expressive Loss Functions (A.3)	Error views in machine learning (Glick 1978; Lissack and Fu 1976; Patrikainen and Meila 2006).
Feasibility (A.4)	Significance of estimates against baseline settings (Adcock 1997; Mukherjee et al. 2003).
Flexible Data Settings (B.1/4/5)	Simulations with hard data assumptions: mixtures of distributions, local dependencies and noise (Way et al. 2010; Hua et al. 2005; Guo et al. 2010; Madeira and Oliveira 2004).
Retrieval of Data Regularities (B.2)	Data regularities to contextualize assessment (Dobbin and Simon 2007; Raudys and Jain 1991).
Dimensionality Effect (B.3)	Extrapolate guarantees by sub-sampling features (Mukherjee et al. 2003; Guo et al. 2010).
Advanced Data Properties (B.6)	Modeling of additional sources of variability (Dobbin and Simon 2005).
Unbalanced/Difficult Data (C.1)	Guarantees from unbalanced data and adequate loss functions (Guo et al. 2010; Beleites et al. 2013).
Multi-class Tasks (C.2)	Integration of class-centric performance bounds (Beleites et al. 2013).
Descriptive Models (C.3)	Adequate loss functions and collection of error estimates for global and (bi)clustering models (Madeira and Oliveira 2004; Hand 1986).
Guidance Criteria (C.4)	Weighted optimization methods for robust and compact multi-parameter analysis (Deng 2007).

3 Principles to Bound and Compare the Performance of Models

The solution space is proposed according to the target tasks of bounding or comparing the performance of a model M learned from high-dimensional spaces. The adequate definition of estimators of the true error is the central point of focus. An illustrative simplistic estimator of the performance of classification model can be described by a collection of observed errors obtained under a k-fold cross-validation, with its expected value being their average:

$E[\theta(\varepsilon_{true})] \approx \frac{1}{k}\Sigma_{i=1}^{k}(\varepsilon_i \mid M, n, m, P_{X|Y})$,

where ε_i is the observed error for the i^{th} fold. When the number of observations is not significantly large, the errors can be collected under a leave-one-out scheme, where k=n and the ε_i is, thus, simply given by a loss function L applied over a single testing instance (x_i, y_i): $L(M(x_i)$=$\hat{y}_i, y_i)$.

In the presence of an estimator for the true error, finding performance bounds can rely on non-biased estimators from the collected error estimates, such as the mean and q-percentiles to provide a bar-envelope around the mean estimator (e.g. $q \in \{20\%, 80\%\}$). However, such strategy does not robustly consider the variability of the observed errors. A simple and more robust alternative is to derive the confidence intervals for the expected true performance based on the distribution underlying the observed error estimates.

Although this estimator considers the variability across estimates, it still may not reflect the true performance bounds of the model due to poor sampling and loss function choices. Additionally, when the number of features exceeds the number of observations, the collected errors can be prone to systemic biases and even statistically inviable for inferring performance guarantees. These observations need to be carefully considered to shape the statistical assessment.

The definition of good estimators is also critical for comparing models, as these comparisons can rely on their underlying error distributions. For this goal, either the traditional t-Student, McNemar and Wilcoxon tests can be adopted to compare pairs of classifiers, and Friedman tests with the corresponding post-hoc tests (Demšar 2006) or less conservative tests[1] (García and Herrera 2009) can be adopted for either comparing distinct models, models learned from multiple datasets or models with different parameterizations.

Motivated by the surveyed contributions to tackle the limitations of existing approaches, this section derives a set of principles for a robust assessment of the performance guarantees of models learned from high-dimensional spaces. First, these principles are incrementally provided according to the introduced major sets of challenges. Second, we show that these principles can be consistently and coherently combined within a simplistic assessment methodology.

3.1 Robust Statistical Assessment

Variability of Performance Estimates. Increasing the dimensionality m for a fixed number of observations n introduces variability in the performance of the learned model that must be incorporated in the estimation of performance bounds for a specific sample size. A simplistic principle is to compute the confidence intervals from error estimates $\{\varepsilon_1, .., \varepsilon_k\}$ obtained from k train-test partitions by fitting an underlying distribution (e.g. Gaussian) that is able to model their variance.

However, this strategy has two major problems. First, it assumes that the variability is well-measured for each error estimate. This is commonly not true as each

[1] Friedman tests rely on pairwise Nemenyi tests that are conservative and, therefore, may not reveal a significant number of differences among models (García and Herrera 2009)

error estimate results from averaging a loss function across testing instances within a partitioning fold, which smooths and hides the true variability. Second, when the variance across estimates is substantially high, the resulting bounds and comparisons between models are not meaningful. Thus, four additional strategies derived from existing research are proposed for: *1)* a robust assessment of models that preserve the original dimensionality, *2)* correcting performance guarantees of models that rely on subspaces of the original space, *3)* reducing the variability of errors in $m \gg n$ settings, and *4)* obtaining more conservative guarantees.

First, the discriminant properties of multivariate models learned over the original space can be used to approximate the observed error for a particular setting $\theta_n(\varepsilon_{true} \mid m, M, P_{X|Y})$ and the asymptotic estimate of the true error $\lim_{n\to\infty} \theta_n(\varepsilon_{true} \mid m, M, P_{X|Y})$ (Ness and Simpson 1976). An analysis on the deviations of the observed error from the true error as a function of data size n, dimensionality m and discriminant functions M was initially provided by Raudys and Jain (1991) and extended by more recent approaches (Bühlmann and Geer 2011; Cai and Shen 2010).

Second, the unbiasedness principle from feature selection methods can be adopted to affect the significance of performance guarantees. Learning models M that rely on decisions over subsets of features either implicitly or explicitly use a form of feature selection driven by core metrics, such as Mahalanobis, Bhattacharyya, Patrick-Fisher, Matusita, divergence, mutual Shannon information, and entropy (Raudys and Jain 1991). In this context, statistical tests can be made to guarantee that the value of a given metric per feature is sufficiently better than a random distribution of values when considering the original dimensionality (Singhi and Liu 2006; Iswandy and Koenig 2006). These tests return a p-value that can be used to weight the probability of the selected set of features being selected by chance over the (n,m)-space and, consequently, to affect the performance bounds and the confidence of comparisons of the target models. Singhi and Liu (2006) formalize selection bias, analyze its statistical properties and how they impact performance bounds.

Third, when error estimates are collected, different methods have been proposed for controlling the observed variability across estimates (Raeder, Hoens, and Chawla 2010; Jain et al. 2003), ranging from general principles related with sampling schema and density functions to more specific statistical tests for a correct assessment of the true variability in specific biomedical settings where, for instance, replicates are considered. These options are revised in detail in the next subsections.

Fourth, conservative bounds for a given dimensionality can be retrieved from the VC-dimension (capacity) of a target model (Vapnik 1982; Blumer et al. 1989). These bounds can be used to guide model comparison. The VC-dimension can be obtained either theoretically or experimentally (Vayatis and Azencott 1999). A common experimental estimation option for the VC-dimension is to study the maximum deviation of error rates among independently labeled datasets. An illustrative lower-bound for the estimator of the true performance of a M model composed by h mapping functions (number of decisions from the values of m features) is: $\theta_n(\varepsilon_{true}) \geq \frac{1}{n}(log\frac{1}{\delta} + logh)$ (Apolloni and Gentile 1998), where δ is the

statistical power[2]. In high-dimensional spaces, h tends to be larger, degrading the performance bounds if the number of instances is small. For more complex models, such as Bayesian learners or decision trees, the VC-dimension can be adopted using assumptions that lead to less conservative bounds[3] (Apolloni and Gentile 1998). Still, bounds tend to be loose as they are obtained using a data-independent analysis and rely on a substantial number of approximations.

Bias Associated with High-Dimensional Spaces. In (n,m)-spaces where $n < m$, the observed error associated with a particular model can be further decomposed in bias and variance components to understand the major cause of the variability across error estimates. While variance is determined by the ability to generalize a model from the available observations (see Fig.2), the bias is mainly driven by the complexity of the learning task from the available observations. High levels of bias are often found when the collection of instances is selected from a specific stratum, common in high-dimensional data derived from social networks, or affected by specific experimental or pre-processing techniques, common in biomedical data. For this reason, the bias-variance decomposition of error provides useful frame to study the error performance of a classification or regression model, as it is well demonstrated by its effectiveness across multiple applications (Domingos 2000). To this end, multiple metrics and sampling schemes have been developed for estimating bias and variance from data, including the widely used holdout approach of Kohavi and Wolpert (Kohavi and Wolpert 1996).

Sampling Schema. When the true performance estimator is not derived from the analysis of the parameters of the learned model, it needs to rely on samples from the original dataset to collect estimates. Sampling schema are defined by two major variables: sampling criteria and train-test size decisions. Error estimations in high-dimensional data strongly depend on the adopted resampling method (Way et al. 2010). Many principles for the selection of sampling methods have been proposed (Molinaro, Simon, and Pfeiffer 2005; Dougherty et al. 2010; Toussaint 1974). Cross-validation methods and alternative bootstrap methods (e.g. randomized bootstrap, 0.632 estimator, mc-estimator, complex bootstrap) have been compared and assessed for a large number of contexts. Unlike cross-validation, bootstrap was shown to be pessimistically biased with respect to the number of training samples. Still, studies show that bootstrap becomes more accurate than its peers for spaces with very large observed errors as it is often observed in high-dimensional spaces where $m > n$ (Dougherty et al. 2010). Resubstitution methods are optimistically biased and should be avoided. We consider both the use of k-folds cross-validation and bootstrap to be acceptable. In particular, the number of folds, k, can be adjusted based on the minimum number of estimates for a statistical robust assessment of confidence

[2] Inferred from the probability $P(\varepsilon_{true} \mid M,m,n)$ to be consistent across the n observations.

[3] The number and length of subsets of features can be used to affect the performance guarantees. For instance, a lower-bound on the performance of decision lists relying on tests with at most p features chosen from a m-dimensional space and d-depth is $\theta(\varepsilon_{true}) \geq \frac{1}{n}(log\frac{1}{\delta} + \Theta(p^d log_2 p^d))$.

intervals. This implies a preference for a large number of folds in high-dimensional spaces with either high-variable performance or $n \ll m$.

An additional problem when assessing performance guarantees in (n,m)-spaces where $n < m$, is to guarantee that the number of test instances per fold offers a reliable error estimate since the observed errors within a specific fold are also subjected to systematic (bias) and random (variance) uncertainty. Two options can be adopted to minimize this problem. First option is to find the best train-test split. Raudys and Jain (1991) proposed a loss function to find a reasonable size of the test sample based on the train sample size and on the estimate of the asymptotic error, which essentially depends on the dimensionality of the dataset and on the properties of the learned model M. A second option is to model the testing sample size independently from the number of training instances. This guarantees a robust performance assessment of the model, but the required number of testing instances can jeopardize the sample size and, thus, compromise the learning task. Error assessments are usually described as a Bernoulli process: n_{test} instances are tested, t successes (or failures) are observed and the true performance for a specific fold can be estimated, $\hat{p}=t/n_{test}$, as well as its variance $p(1\text{-}p)/n_{test}$. The estimation of n_{test} can rely on confidence intervals for the true probability p under a pre-specified precision[4] (Beleites et al. 2013) or from the expected levels of type I and II errors using the statistical tests described by Fleiss (1981).

Loss Functions. Different loss functions capture different performance views, which can result in radically different observed errors, $\{\varepsilon_1, \ldots \varepsilon_k\}$. Three major views can be distinguished to compute each of these errors for a particular fold from these loss functions. First, error counting, the commonly adopted view, is the relative number of incorrectly classified/predicted/described testing instances. Second, smooth modification of error counting (Glick 1978) uses distance intervals, and it is applicable for classification models with probabilistic outputs (correctly classified instances can contribute to the error) and for regression models. Finally, posterior probability estimate (Lissack and Fu 1976) is often adequate in the presence of the class-conditional distributions. These two latter metrics provide a critical complementary view for models that deliver probabilistic outputs. Additionally, their variance is more realistic than the simple error counting. The problem with smooth modification is its dependence on the error distance function, while posterior probabilities tend to be biased for small datasets.

Although error counting (and the two additional views) are commonly parameterized with an accuracy-based loss function (incorrectly classified instances), other metrics can be adopted to turn the analysis more expressive or to be extensible towards regression models and descriptive models. For settings where the use of confusion matrices is of importance due to the difficulty of the task for some

[4] For some biomedical experiments (Beleites et al. 2013), 75-100 test samples are commonly necessary to achieve reasonable validation and 140 test samples (confidence interval widths 0.1) are necessary for an expected sensitivity of 90%. When this number is considerably higher than the number of available observations, there is the need to post-calibrate the test-train sizes according to the strategies depicted for the first option.

classes/ranges of values, the observed errors can be further decomposed according to type-I and type-II errors.

A synthesis of the most common performance metrics per type of model is provided in Table 4. A detailed analysis of these metrics is provided in Section 3.3 related with extensibility principles. In particular, in this section we explain how to derive error estimates from descriptive settings.

The use of complementary loss functions for the original task (1) is easily supported by computing performance guarantees multiple times, each time using a different loss function to obtain the error estimates.

Table 4 Performance views to estimate the true error of discriminative and descriptive models

Model	Performance views
Classification model	accuracy (percentage of samples correctly classified); area under receiver operating characteristics curve (AUC); critical complementary performance views can be derived from (multi-class) confusion matrices, including sensitivity, specificity and the F-Measure.
Regression model	simple, average normalized or relative root mean squared error; to draw comparisons with literature results, we suggest the use of the normalized root mean squared error (NRMSE) and the symmetric mean absolute percentage of error (SMAPE).
Descriptive Local model (presence of hidden bics.)	entropy, F-measure and match score clustering metrics (Assent et al. 2007; Sequeira and Zaki 2005); F-measure can be further decomposed in terms of recall (coverage of found samples by a hidden cluster) and precision (absence of samples present in other hidden clusters); match scores (Prelić et al. 2006) assess the similarity of solutions based on the Jaccard index; Hochreiter et al. (2010) introduced a consensus score by computing similarities between all pairs of biclusters; biclustering metrics can be delivered by the application of a clustering metric on both dimensions or by the relative non-intersecting area (RNAI) (Bozdağ, Kumar, and Catalyurek 2010; Patrikainen and Meila 2006).
Descriptive Local model (absence of hidden bics.)	merit functions can be adopted as long as they are not biased towards the merit criteria used within the approaches under comparison (mean squared residue introduced by Cheng and Church (2000) or the Pearson's correlation coefficent; domain-specific evaluations can be adopted by computing statistical enrichment p-values (Madeira and Oliveira 2004).
Descriptive Global model	merit functions to test the fit in the absence of knowledge regarding the regularities; equality tests between multivariate distributions; similarity functions between the observed and approximated distributions.

Feasibility of Estimates. As previously prompted, different estimators of the true error can be defined to find confidence intervals or significant differences associated with the performance of a specific model M. For this goal, we covered how to derive estimators from the parametric analysis of the learned models or from error estimates gathered under a specific sampling scheme and loss function. Nevertheless, the performance guarantees defined by these estimators are only valid if they

are able to perform better than a null (random) model under a reasonable statistical significance level. An analysis of the significance of these estimators indicates whether we can estimate the performance guarantees of a model or, otherwise, we would need a larger number of observations for the target dimensionality.

A simplistic validation option is to show the significant superiority of M against permutations made on the original dataset (Mukherjee et al. 2003). A possible permutation procedure is to construct for each of the k folds, t samples where the classes (discriminative models) or domain values (descriptive models) are randomly permuted. From the errors computed for each permutation, different density functions can be developed, such as:

$$P_{n,m}(x) = \frac{1}{kt}\Sigma_{i=1}^{k}\Sigma_{j=1}^{t}\theta(x-\varepsilon_{i,j,n,m}), \tag{2}$$

where $\theta(z) = 1$ if $z \geq 0$ and 0 otherwise. The significance of the model is $P_{n,m}(x)$, the percentage of random permutations with observed error smaller than x, where x can be fixed using a estimator of the true error for the target model M. The average estimator, $\varepsilon_{n,m} = \frac{1}{k}\Sigma_{i=1}^{k}(\varepsilon_i \mid n,m)$, or the θ^{th} percentile of the sequence $\{e_1,...,e_k\}$ can be used as an estimate of the true error. Both the average and θ^{th} percentile of error estimates are unbiased estimators. Different percentiles can be used to define error bar envelopes for the true error.

However, there are two major problems with this approach. First, the variability of the observed errors does not affect the significance levels. To account for the variability of the error estimates across the $k \times t$ permutations, more robust statistical tests can be used, such as one-tailed t-test with $(k \times t)$-1 degrees of freedom to test the unilateral superiority of the target model. Second, the significance of the learned relations of a model M is assessed against permuted data, which is a very loose setting. Instead, the same model should be assessed against data generated with similar global regularities in order to guarantee that the observed superiority does not simply result from an overfitting towards the available observations. Similarly, stastical t-tests are suitable options for this scenario.

When this analysis reveals that error estimates cannot be collected with statistical significance due to data size constraints, two additional strategies can be applied. A first strategy it to adopt complementary datasets by either: *1)* relying on identical real data with more samples (note, however, that distinct datasets can lead to quite different performance guarantees (Mukherjee et al. 2003)), or by *2)* approximating the regularities of the original dataset and to generated larger synthetic data using the retrieved distributions. A second strategy is to relax the significance levels for the inference of less conservative performance guarantees. In this case, results should be provided as indicative and exploratory.

3.2 Data Flexibility

Deriving performance guarantees from a single dataset is of limited interest. Even in the context of a specific domain, the assessment of models from multiple datasets with varying regularities of interest provides a more complete and general frame to

validate their performance. However, in the absence of other principles, the adoption of multiple datasets leads to multiple, and potentially contradicting, performance guarantees. Principles for the generalization of performance bounds and comparisons[5] retrieved from distinct datasets are proposed in *Section 3.4*.

When real datasets are used, their regularities should be retrieved for a more informative context of the outputted performance guarantees. For this goal, distribution tests (with parameters estimated from the observed data) to discover global regularities, biclustering approaches to identify (and smooth) meaningful local correlations, and model reduction transformations to detect (and remove) redundancies (Hocking 2005) can be adopted. When the target real datasets are sufficiently large, size and dimensionality can be varied to approximate learning curves or to simply deliver performance bounds and comparisons for multiple (n,m)-spaces. Since performance bounds and comparisons for the same (n,m)-space can vary with the type of data[6], it is advisable to only combine estimates from datasets that share similar conditional regularities $P_{X|Y}$.

In simulation studies, synthetic datasets should be generated using realistic regularities. Common distribution assumptions include either single or multiple multivariate Gaussian distributions (Way et al. 2010; Guo et al. 2010; Hua et al. 2005; El-Sheikh and Wacker 1980), respectively, for descriptive ($M(X)$) or discriminative models ($M : X \rightarrow Y$). In classification settings, it is common to assume unequal means and equal covariance matrices ($X_i \mid y_1 \sim Gaussian(\mu_1, \sigma^2)$, $X_j \mid y_2 \sim Gaussian(\mu_2, \sigma^2)$, where $\mu_1 \neq \mu_2$). The covariance-matrix can be experimentally varied or estimated from real biomedical datasets. In (Way et al. 2010), unequal covariance matrices that differ by a scaling factor are considered. While a few datasets after proper normalization have a reasonable fit, the majority of biomedical datasets cannot be described by such simplistic assumption. In these cases, the use of mixtures, such as the mixture of the target distribution with Boolean feature spaces (Kohavi and John 1997), is also critical to assess non-linear capabilities of the target models. Hua et al. (2005) proposes a hard bimodal model, where the conditional distribution for class y_1 is a Gaussian centered at μ_0=(0,...,0) and the conditional distribution for class y_2 is a mixture of equiprobable Gaussians centered at $\mu_{1,0}$=(1,...,1) and $\mu_{1,1}$=(-1,...,-1). In Guo et al. (2010) study, the complexity of Gaussian conditional distributions was tested by fixing μ_0=0 and by varying μ_1 from 0.5 to 0 in steps of 0.05 for $\sigma_0^2 = \sigma_1^2 = 0.2$. Additionally, one experimental setting generated data according to a mixture of Uniform $U(\mu + 3\sigma, \mu + 6.7\sigma)$ and Gaussian $N(\mu, \sigma^2)$ distributions.

Despite these flexible data assumptions, some datasets have features exhibiting highly skewed distributions. This is a common case with molecular data (particularly from human tissues). The study by Guo et al. (2010) introduces varying levels of signal-to-noise in the dataset, which resulted in a critical decrease of the observed statistical power for the computed bounds. Additionally, only a subset of overall fea-

[5] The comparison of performance of models can be directly learned from multiple datasets using the introduced Friedman framework based on Nemenyi tests (Demšar 2006).

[6] Distinct datasets with identical (n,m)-spaces can have significantly different learning complexities (Mukherjee et al. 2003).

tures was generated according to class-conditional distributions in order to simulate the commonly observed compact set of discriminative biomarker features.

The majority of real-world data settings is also characterized by functionally correlated features and, therefore, planting different forms of dependencies among the m target features is of critical importance to infer performance guarantees. Hua et al. (2005) propose the use of different covariance-matrices by dividing the overall features into correlated subsets with varying number of features ($p \in \{1,5,10,30\}$), and by considering different correlation coefficients ($\rho \in \{0.125,0.25,0.5\}$). The increase in correlation among features, either by decreasing g or increasing ρ, increases the Bayes error for a fixed dimensionality. Guo et al. (2010) incorporate a correlation factor just for a small portion of the original features. Other studies offer additional conditional distributions tested using unequal covariance matrices (Way et al. 2010). Finally, biclusters can be planted in data to capture flexible functional relations among subsets of features and observations. Such local dependencies are commonly observed in biomedical data (Madeira and Oliveira 2004).

Additional sources of variability can be present, including technical biases from the collected sample of instances or replicates, pooling and dye-swaps in biological data. This knowledge can be used to shape the estimators of the true error or to further generate new synthetic data settings. Dobbin and Simon (2005; 2007) explored how such additional sources of variability impact the observed errors. The variability added by these factors is estimated from the available data. These factors are modeled for both discriminative (multi-class) and descriptive (single-class) settings where the number of independent observations is often small. Formulas are defined for each setting by minimizing the difference between the asymptotic and observed error, $(\lim_{n\to\infty} \varepsilon_{true|n}) - \varepsilon_{true|n}$, where $\varepsilon_{true|n}$ depends on these sources of variability. Although this work provides hints on how to address advanced data aspects with impact on the estimation of the true error, the proposed formulas provide loose bounds and have been only deduced in the scope of biological data under the independence assumption among features. The variation of statistical power using ANOVA methods has been also proposed to assess these effects on the performance of models (Surendiran and Vadivel 2011).

Synthesizing, flexible data assumptions allow the definition of more general, complete and robust performance guarantees. Beyond varying the size n and dimensionality m, we can isolate six major principles. First, assessing models learned from real and synthetic datasets with disclosed regularities provide complementary views for robust and framed performance guarantees. Second, when adopting multivariate Gaussian distributions to generate data, one should adopt varying distances between their means, use covariance-matrices characterized by varying number of features and correlation factors, and rely on mixtures to test non-linear learning properties. Non-Gaussian distributions can be complementary considered. Third, varying degrees of noise should be planted by, for instance, selecting a percentage of features with skewed values. Fourth, impact of selecting a subset of overall features with more discriminative potential (e.g. lower variances) should be assessed. Fifth, other properties can be explored, such as the planting of local regularities with different properties to assess the performance guarantees of descriptive models and

the creation of imbalance between classes to assess classification models. Finally, additional sources variability related with the specificities of the domains of interest can be simulated for context-dependent estimations of performance guarantees.

3.3 Extensibility

Performance Guarantees from Imbalanced Data Settings. Imbalance in the representativity of classes (classification models), range of values (regression models) and among feature distributions affect the performance of models and, consequently, the resulting performance guarantees. In many high-dimensional contexts, such as biomedical labeled data, case and control classes tend to be significantly unbalanced (access to rare conditions or diseases is scarce). In these contexts, it is important to compute performance guarantees in (n,m)-spaces from unbalanced real data or from synthetic data with varying degrees of imbalance. Under such analysis, we can frame the performance guarantees of a specific model M with more rigor. Similarly, for multi-class tasks, performance guarantees can be derived from real datasets and/or synthetic datasets (generated with a varying number and imbalance among the classes) to frame the true performance of a target model M.

Additionally, an adequate selection of loss functions to compute the observed errors is required for these settings. Assuming the presence of c classes, one strategy is to estimate performance bounds c times, where each time the bounds are driven by a loss function based on the sensitivity of that particular class. The overall upper and lower bounds across the c estimations can be outputted. Such illustrative method is critical to guarantee the robustness assessment of the performance of classification models for each class.

Performance Guarantees of Descriptive Models. The introduced principles in *Sections 3.1* and *3.2* to derive performance guarantees of discriminative models can be extended for descriptive models under a small set of assumptions. Local and global descriptive models can be easily used when considering one of the loss functions proposed in Table 4. The evaluation of local descriptive models can either be performed in the presence or absence of hidden (bi)clusters, H. Similarly, global descriptive models that return a mixture of distributions that approximate the population from which the sample was retrieved, $X \sim \pi$, can be evaluated in the presence and absence of the underlying true regularities.

However, both descriptive and global models cannot rely on traditional sampling schema to collect error estimates. Therefore, in order to have multiple error estimates for a particular (n,m)-space, which is required for a robust statistical assessment, these estimates should be computed from:

- alternative subsamples of a particular dataset (testing instances are discarded);
- multiple synthetic datasets with fixed number of observations n and features m generated under similar regularities.

3.4 Inferring Performance Guarantees from Multiple Settings

In the previous sections, we proposed alternative estimators of the true performance, and the use of datasets with varying regularities. Additionally, the performance of learning methods can significantly vary depending on their parameterizations. Some of the variables that can be subject to variation include: data size, data dimensionality, loss function, sampling scheme, model parameters, distributions underlying data, discriminative and skewed subsets of features, local correlations, degree of noise, among others. Understandably, the multiplicity of views related with different estimators, parameters and datasets results in a large number of performance bounds and comparison-relations that can hamper the assessment of a target model. Thus, inferring more general performance guarantees is critical and valid for studies that either derive specific performance guarantees from collections of error estimates or from the direct analysis of the learned models.

Guiding criteria need to be considered to frame the performance guarantees of a particular model M based on the combinatorial explosion of hyper-surfaces that assess performance guarantees from these parameters. When *comparing models*, simple statistics and hierarchical presentation of the inferred relations can be available. An illustrative example is the delivery of the most significant pairs of values that capture the percentage of settings where a particular model had a superior and inferior performance against another model.

When *bounding performance*, a simple strategy is to use the minimum and maximum values over similar settings to define conservative lower and upper bounds. More robustly, error estimates can be gathered for the definition of more general confidence intervals. Other criteria based on weighted functions can be used to frame the bounds from estimates gathered from multiple estimations (Deng 2007). In order to avoid very distinct levels of difficulty across settings that penalized the inferred performance bounds, either a default parameterization can be considered for all the variables and only one variable be tested at a time or distinct settings can be clustered leading to a compact set of performance bounds.

3.5 Integrating the Proposed Principles

The retrieved principles can be consistently and coherently combined according to a simple methodology to enhance the assessment of the performance guarantees of models learned from high-dimensional spaces. First, the decisions related with the definition of the estimators, including the selection of adequate loss functions and sampling scheme and the tests of the feasibility of error estimates, provide a structural basis to bound and compare the performance of models.

Second, to avoid biased performance guarantees towards a single dataset, we propose the estimation of these bounds against synthetic datasets with varying properties. In this context, we can easily evaluate the impact of assuming varying regularities $X|Y$, planting feature dependencies, dealing with different sources of variability, and of creating imbalance for discriminative models. Since varying a large number of parameters can result in a large number of estimations, the identified strategies to deal with the inference of performance guarantees from multiple settings should be adopted in order to collapse these estimations into a compact frame of performance guarantees.

Third, in the presence of a model that is able to preserve the original space (e.g. support vector machines, global descriptors, discriminant multivariate models), the impact of dimensionality in the performance guarantees is present by default, and it can be further understood by varying the number of features. For models that rely on subsets of overall features, as the variability of the error estimates may not reflect the true performance, performance guarantees should be adjusted through the unbiasedness principle of feature selection or conservative estimations should be considered recurring to VC-theory.

Finally, for both discriminative and descriptive models, the estimator of the true performance should be further decomposed to account for both the bias and variance underlying error estimates. When performance is highly-variable (loose performance guarantees), this decomposition offers an informative context to understand how the model is able to deal with the risk of overfitting associated with high-dimensional spaces.

4 Results and Discussion

In this section we experimentally assess the relevance of the proposed methodology. First, we compare alternative estimators and provide initial evidence for the need to consider the proposed principles when assessing performance over high-dimensional datasets when $n<m$. Second, we bound and compare the performance of classification models learned over datasets with varying properties. Finally, we show the importance of adopting alternative loss functions for imbalanced multi-class and single-class (descriptive) models.

For these experiments, we rely on both real and synthetic data. Two distinct groups of real-world datasets were used: high-dimensional datasets with small number of instances ($n<m$) and high-dimensional datasets with a large number of instances. For the first group we used expression data for tumor classification collected from BIGS repository[7]: *colon* cancer data (m=2000, n=62, 2 labels), *lymphoma* data (m=4026, n=96, 9 labels), and *leukemia* data (m=7129, n=72, 2 labels). For the second group we selected a random population from the healthcare heritage prize database[8] (m=478, n=20000) which integrates claims across hospitals,

[7] `http://www.upo.es/eps/bigs/datasets.html`

[8] `http://www.heritagehealthprize.com/c/hhp/data` (under a granted permission).

pharmacies and laboratories. The original relational scheme was denormalized by mapping each patient as an instance with features extracted from the collected claims (400 attributes), the monthly laboratory tests and taken drugs (72 attributes), and the patient profile (6 attributes). We selected the tasks of classifying the need for upcoming interventions (2 labels) and the level of drug prescription ({*low,moderate, high*} labels), considered to be critical tasks for care prevention and drug management.

Two groups of synthetic datasets were generated: multi-label datasets for discriminative models and unlabeled datasets for descriptive models. The labeled datasets were obtained by varying the following parameters: the ratio and the size of the number of observations and features, the number of classes and their imbalance, the conditional distributions (mixture of Gaussians and Poissons per class), the amount of planted noise, the percentage of skewed features, and the area of planted local dependencies. The adopted parameterizations are illustrated in Table 5. To study the properties of local descriptive models, synthetic datasets with varying number and shape of planted biclusters were generated. These settings, described in Table 6, were carefully chosen in order to follow the properties of molecular data (Serin and Vingron 2011; Okada, Fujibuchi, and Horton 2007). In particular, we varied the size of these matrices up to m=4000 and n=400, maintaining the proportion between rows and columns commonly observed in gene expression data.

Table 5 Parameters for the generation of the labeled synthetic datasets

Features	$m \in \{500, 1000, 2000, 5000\}$
Observations/Instances	$n \in \{50, 100, 200, 500, 1000, 10000\}$
Number of Classes	$c \in \{2, 3, 5\}$
Distributions (illustrative)	(c=3) $\{N(1,\sigma), N(0,\sigma), N(-1,\sigma)\}$ with $\sigma \in \{3,5\}$ (easy setting) (c=3) $\{N(u_1,\sigma), N(0,\sigma), N(u_3,\sigma)\}$ with $u_1 \in \{-1,2\}$, $u_2 \in \{-2,1\}\}$ (c=3) mixtures of $N(u_i,\sigma)$ and $P(\lambda_i)$ where λ_1=4, λ_2=5, λ_3=6
Noise (% of values' range)	{0%,5%,10%,20%,40%}
Skewed Features	{0%,30%,60%,90%}
Degree of Imbalance (%)	{0%,40%,60%,80%}

Table 6 Properties of the generated set of unlabeled synthetic datasets

Features×Observations (♯m×♯n)	100×30	500×60	1000×100	2000×200	4000×400
Nr. of hidden biclusters	3	5	10	15	20
Nr. columns in biclusters	[5,7]	[6,8]	[6,10]	[6,14]	[6,20]
Nr. rows in biclusters	[10,20]	[15,30]	[20,40]	[40,70]	[60,100]
Area of biclusters	9.0%	2.6%	2.4%	2.1%	1.3%

The software implementing the methodology that combines the introduced principles was coded in Java (JVM version 1.6.0-24). The selected supervised learners

were run from WEKA. The following experiments were computed using an Intel Core i3 1.80GHz with 6GB of RAM.

Challenges. An initial assessment of the performance of two simplistic classification models learned from real high-dimensional datasets is given in Fig.3. The performance bounds[9] from real datasets where $m>n$ confirm the high-variability of performance associated with the learning in these spaces. In particular, the difference between the upper and lower bounds is over 30% for cross-validation options with 10 folds and n folds (leave-one-out). In general, leave-one-out sampling scheme has higher variability than 10-fold cross-validation. Although leave-one-out is able to learn from more observations (decreasing the variability of performance), the true variability of 10-fold cross-validation is masked by averaging errors per fold. The smooth effect of cross-validation sampling supports the need to increase the levels of significance to derive more realistic performance bounds. Additionally, the use of bootstrap schema with resampling methods to increase the number of instances seems to optimistically bias the true performance of the models. Contrasting with these datasets, models learned from the heritage data setting, where $n \gg m$, have a more stable performance across folds. Consistently with these observations, the use of Friedman tests reveals a higher number of superiority relations among classification models learned from the heritage data.

Bounding performance using VC inference or specific percentiles of error estimates introduces undesirable bias. In fact, under similar experimental settings, the VC bounds were very pessimistic (>10 percentage points of difference), while the use of the 0.15 and 0.85 percentiles (to respectively define lower and upper bounds) led to more optimistic bounds against the bounds provided in Fig.3. Although changing percentiles easily allows to tune the target level of conservatism, they do not capture the variability of the error estimates.

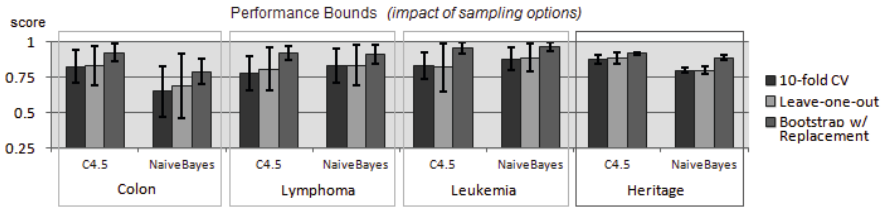

Fig. 3 Performance guarantees from real datasets with varying $\frac{n}{m}$ degree for two classifiers tested under different sampling options

A set of views on the significance of the learned relations from real and synthetic high-dimensional datasets is respectively provided in Table 7 and Fig.4. Different

[9] Confidence intervals of a mean estimator from the sample of error estimates assumed to be normally distributed with the same expectation mean, a standard error $\frac{\sigma}{\sqrt{n}}$ and significance α=0.05.

methods were adopted to compute the significance (p-value) associated with a collection of error estimates. These methods basically compute a p-value by comparing the collected error estimates against estimates provided by loose settings where: *1)* the target model is learned from permuted data, *2)* a *null* classifier[10] is learned from the original data, and *3)* the target model is learned from *null* data (preservation of global conditional regularities). We also considered the setting proposed by Mukherjee et al. (2003) (2). Comparisons are given by one-tailed t-tests. For this analysis, we compared the significance of the learned C4.5, Naive Bayes and support vector machines (SVM) models for real datasets and averaged their values for synthetic datasets. A major observation can be retrieved: p-values are not highly significant ($\ll 1\%$) when $n<m$, meaning that the performance of the learned models is not significantly better than very loose learners. Again, this observation pinpoints the importance of carefully framing assessments of models learned from high-dimensional spaces. Additionally, different significance views can result in quite different p-values, which stresses the need to choose an appropriate robust basis to validate the collected estimates. Comparison against null data is the most conservative, while the counts performed under (2) (permutations density function) are not sensitive to distances among error mismatches and easily lead to biased results.

Table 7 Significance of the collected error estimates of models learned from real datasets using improvement p-values. p-values are computed by comparing the target models vs. baseline classification models, and error estimates collected from the original dataset vs. a permuted dataset or null dataset (where basic regularities are preserved)

	Colon			Leukemia			Heritage		
	C4.5	NBayes	SVM	C4.5	NBayes	SVM	C4.5	NBayes	SVM
Comparison Against Permuted Data	1.5%	41.3%	1.2%	0.6%	0.1%	0.2%	∼0%	∼0%	∼0%
Comparison Against Null Model	1.1%	32.2%	1.2%	0.1%	0.1%	0.1%	∼0%	∼0%	∼0%
Comparison Against Null Dataset	15.2%	60.3%	9.3%	9.7%	12.0%	7.2%	1.3%	3.8%	1.7%
Permutations Density Function (2)	14.0%	36.0%	8.4%	8.4%	1.2%	0.8%	0.0%	0.4%	0.0%

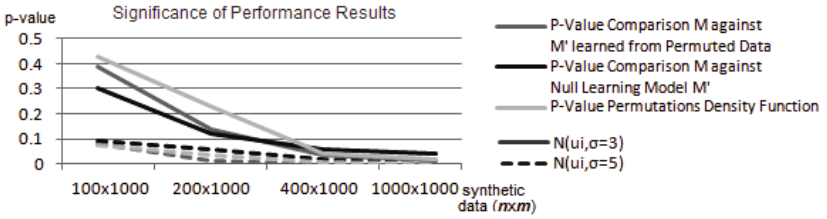

Fig. 4 Significance views on the error estimates collected by classification models from $m>n$ synthetic datasets under easy N(u_i,σ=3) and moderate N(u_i,σ=5) settings against loose baseline settings

[10] A classifier that defines the average conditional values per feature during the training phase and the mode of feature-based classes during the testing phase was considered.

To further understand the root of the variability associated with the performance of models learned from high-dimensional datasets, Fig.5 provides its decomposition in two components: bias and variance. Bias provides a view on how the expected error deviates across folds for the target dataset. Variance provides a view on how the model behavior differs across distinct training folds. We can observe that the bias component is higher than the variance component, which is partially explained by the natural biased incurred from samples in $n<m$ high-dimensional spaces. The disclosed variance is associated with the natural overfitting of the models in these spaces. Interestingly, we observe that the higher the $\frac{m}{n}$ ratio is, the higher the bias/variance ratio. The sum of these components decreases for an increased number of observations, n, and it also depends on the nature of the conditional distributions of the dataset, as it is shown by the adoption of synthetic datasets with conditional Gaussian distributions with small-to-large overlapping areas under the density curve. The focus on each one of these components for the inference of novel performance guarantees is critical to study the impact of the capacity error and training error associated with the learned model (see Fig.2).

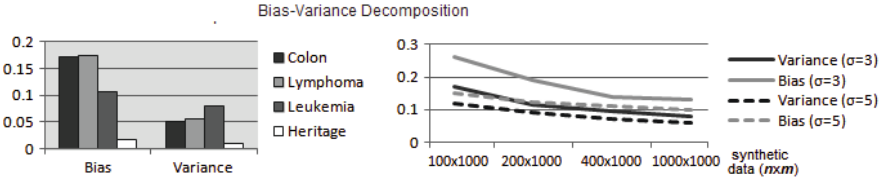

Fig. 5 Decomposition of the performance variability from real and synthetic data (see Table 5) using C4.5: understanding the model capacity (*variance* component) and the model error (*bias* component)

Imbalanced Multi-class Data. The importance of selecting adequate performance views to retrieve realistic guarantees is shown in Fig.6. This is still an underestimated problem that needs to be addressed for both: *1)* balanced datasets where the class-conditional distributions differ in complexity (see the sensitivity associated with the classes from Colon and Leukemia datasets in Fig.6a), and *2)* for imbalanced datasets, where the representativity of each class can hamper the learning task even if the complexity of the class-conditional distributions is similar (see the sensitivity of the classes from datasets with different degrees of imbalance Fig.6b). In this analysis, we adopted sensitivity as a simplistic motivational metric, however many other loss functions hold intrinsic properties of interest to derive particular implications from the performance of the target models. Table 4 synthesizes some of the most common performance views. The chosen view not only impacts the expected true error, but the variability of the error as it is well-demonstrated in Fig.6a. This impacts both the inferred bounds and the number of significant comparisons.

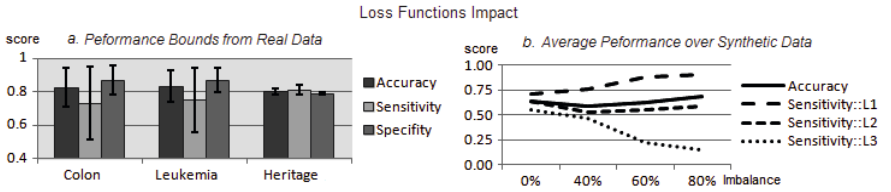

Fig. 6 Impact of adopting alternative loss functions on the: *a*) performance variability of real datasets, and *b*) true performance of synthetic datasets (n=200 and m=500) with varying degrees of imbalance among classes

Performance Guarantees from Flexible Data Settings. To understand how performance guarantees varies across different data settings for a specific model, we computed C4.5 performance bounds from synthetic datasets with varying degree of planted noise and skewed features. Inferring performance guarantees across settings is important to derive more general implications on the performance of models. This analysis is provided in Fig.7. Generalizing performance bounds from datasets with different learning complexity may result in very loose bounds and, therefore, should be avoided. In fact, planting noise and skewing features not only increases the expected error but also its variance. Still, some generalizations are possible when the differences between collections of error estimates is not high. In these cases, collections of error estimates can be joint for the computation of new confidence intervals (as the ones provided in Fig.7). When the goal is to compare sets of models, superiority relations can be tested for each setting under relaxed significance levels, and outputted if the same relation appears across all settings. In our experimental study we were only able to retrieve a small set of superiority relations between C4.5 and Naive Bayes using the Friedman-test under loose levels of significance (10%).

Fig.8 assesses the impact of using different conditional distributions for the inference of general performance guarantees for C4.5. Understandably, the expected error increases when the overlapping area between conditional distributions is higher or when a particular class is described by a mixture of distributions. Combining

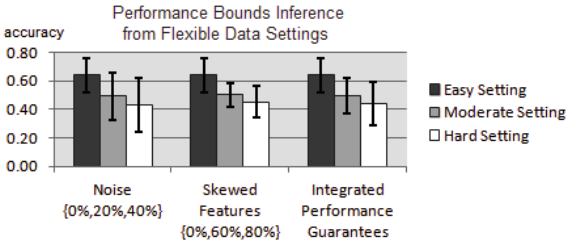

Fig. 7 Inference of performance guarantees in a (n=200,m=500)-space with varying degree of planted noise (as a percentage of domain values) and skewed features (as a percentage of total features)

such hard settings with more easy settings gives rise to loose performance bounds and to a residual number of significant superiority relations between models. Still, this assessment is required to validate and weight the increasing number of data-independent implications of performance from the recent studies.

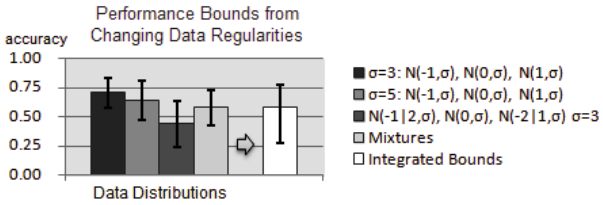

Fig. 8 Inference of performance guarantees from (n=200,m=500)-spaces with different regularities described in Table 5

Descriptive Models. The previous principles are extensible towards descriptive models under an adequate loss function and sampling method to collect estimates. This means that the introduced significance views, decomposition of the error and inference of guarantees from flexible data settings become applicable to different types of models, such as (bi)clustering models and global descriptive models. Fig.9 illustrates the performance bounds of BicPAM biclustering model[11] using three distinct loss functions computed from estimates collected from datasets generated with identical size, dimensionality and underlying regularities (according to Table 6). The target loss functions are the traditional match scores (Prelić et al. 2006), which assess the similarity of the discovered biclusters B and planted biclusters H based on the Jaccard index[12], and the Fabia consensus[13] (Hochreiter et al. 2010). The observed differences on the mean and variability of performance per loss function are enough to deliver distinct Friedman-test results when comparing multiple descriptive models. Therefore, the retrieved implications should be clearly contextualized as pertaining to a specific loss function, sampling scheme, data setting and significance threshold.

[11] `http://web.ist.utl.pt/~rmch/software/bicpam/`

[12] $MS(B,H)$ defines the extent to what found biclusters match with hidden biclusters, while $MS(H,B)$ reflects how well hidden biclusters are recovered:
$\mathbf{MS}(B,H) = \frac{1}{|B|}\Sigma_{(I_1,J_1)\in B} max_{(I_2,J_2)\in H} \frac{|I_1 \cap I_2|}{I_1 \cup I_2|}$

[13] Let S_1 and S_2 be, respectively, the larger and smaller set of biclusters from $\{B,H\}$, and MP be the pairs $B \leftrightarrow H$ assigned using the Munkres method based on overlapping areas (Munkres 1957):
$\mathbf{FC}(B,H) = \frac{1}{|S_1|}\Sigma_{((I_1,J_1)\in S_1,(I_2,J_2)\in S_2)\in MP} \frac{|I_1 \cap I_2| \times |J_1 \cap J_2|}{|I_1| \times |J_1| + |I_2| \times |J_2| - |I_1 \cap I_2| \times |J_1 \cap J_2|}$

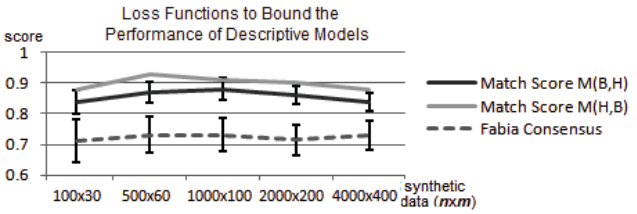

Fig. 9 Performance assessment over biclustering models (BicPAM) using distinct loss functions – Fabia consensus, and match scores $M(B,H)$ and $M(H,B)$ – and a collection of error estimates from 20 data instances per data setting

Final Discussion. In this chapter, we synthesized critical principles to bound and compare the performance of models learned from high-dimensional datasets. First, we surveyed and provided empirical evidence for the challenges related with this task for (n,m)-spaces where $n<m$. This task is critical as implications are derived from studies where the differences in performance of classification models learned over these spaces against permuted and *null* spaces is not significant. Also, the width between the estimated confidence intervals of performance is considerably high, leading to the absence of significant results from Friedman comparisons.

Second, motivated by these challenges, we showed the importance of adopting robust statistical principles to test the feasibility of the collected estimates. Different tests for computing significance levels have been proposed, each one providing different levels of conservatism, which can be used to validate and weight the increasing number of implications derived from the performance of models in high-dimensional spaces.

Third, understanding the source of variability of the performance of the models learned in these spaces is critical as the variability can be either related with the overfitting aspect of the models or with the learning complexity associated with the dataset. The variability of performance can, thus, be further decomposed in variance and bias. While the variance captures the differences on the behavior of the model across samples from the target population, which is indicative of the model capacity (see Fig.2), the bias captures the learning error associated within the available samples. These components disclose the why behind the inferred performance guarantees and, thus, are critical to understand and refine the model behavior.

Fourth, we compared alternative ways of bounding and comparing performance, including different sampling schema, loss functions and statistical tests. In particular, we used initial empirical evidence to show how different estimators can bias the true error or smooth its variability.

An alternative to the inference of performance guarantees from estimates is to approximate the true performance from the properties of the learned models. For this latter line of research two strategies can be followed. A first strategy is to retrieve guarantees from the learned parameters from multivariate models that preserve the original dimensionality (Ness and Simpson 1976; El-Sheikh and Wacker 1980). A

second strategy is to understand the discriminative significance of the selected local subspaces from the original space when a form of feature-set selection is adopted during the learning process (Singhi and Liu 2006; Iswandy and Koenig 2006).

Fifth, the impact of varying data regularities on the performance guarantees was also assessed, including spaces with varying degrees of the $\frac{n}{m}$ ratio, (conditional) distributions, noise, imbalance among classes (when considering classification models), and uninformative features. In particular, we observed that inferring general bounds and comparisons from flexible data settings is possible, but tends to originate very loose guarantees when mixing data settings with very distinct learning complexities. In those cases, a feasible trade-off would be to simply group data settings according to the distributions of each collection of error estimates.

Finally, we showed the applicability of these principles for additional types of models, such as descriptive models.

5 Conclusions and Future Work

Motivated by the challenges of learning from high-dimensional data, this chapter established a solid foundation on how to assess the performance guarantees given by different types of learners in high-dimensional spaces. The definition of adequate estimators of the true performance in these spaces, where the learning is associated with high variance and bias from error estimates, is critical. We surveyed a set of approaches that provide distinct principles on how to bound and compare the performance of models as a function of the data size. A taxonomy to understand their major challenges was proposed. These challenges mainly result from their underlying assumptions and task goals. Existing approaches often fail to provide a robust performance guarantees, are not easily extensible to support unbalanced data settings or assess non-discriminative models (such as local and global descriptive models), and are not able to infer guarantees from multiple data settings with varying properties, such as locally correlated features, noise, and underlying complex distributions.

In this chapter, a set of principles is proposed to answer the identified challenges. They offer a solid foundation to select adequate estimators (either from data sampling or direct model analysis), loss functions, and statistical tests sensitive to the pecularities of the performance of models in high-dimensional spaces. Additionally, these principles provide critical strategies for the generalization of performance guarantees from flexible data settings where the underlying global and local regularities can vary. Finally, we briefly show that these principles can be integrated within a single methodology. This methodology offers a robust, flexible and complete frame to bound and compare the performance of models learned over high-dimensional datasets. In fact, it provides critical guidelines to assess the performance of upcoming learners proposed for high-dimensional settings or, complementary, to determine the appropriate data size and dimensionality required to support decisions related with experimental, collection or annotation costs.

Experimental results support the relevance of these principles. We provided empirical evidence for the importance of computing adequate significance views to

adjust the statistical power when bounding and comparing the performance of models, of selecting adequate error estimators, of inferring guarantees from flexible data settings, and of decomposing the error to gain further insights on the source of its variability. Additionally, we have experimentally shown the extensibility of these decisions for descriptive models under adequate performance views.

This work opens a new door for understanding, bounding and comparing the performance of models in high-dimensional spaces. First, we expect the application of the proposed methodology to study the performance guarantees of new learners, parameterizations and feature selection methods. Additionally, these guarantees can be used to weight and validate the increasing number of implications derived from the application of these models over high-dimensional data. Finally, we expect the extension of this assessment towards models learned from structured spaces, such as high-dimensional time sequences.

Acknowledgments. This work was partially supported by FCT under the projects PTDC/EIA-EIA/ 111239/2009 (Neuroclinomics) and PEst-OE/ EEI/LA0021/2013, and under the PhD grant SFRH/BD/75924/2011.

Software Availability

The generated synthetic datasets and the software implementing the proposed statistical tests are available in http://web.ist.utl.pt/rmch/software/bsize/.

References

1. Adcock, C.J.: Sample size determination: a review. J. of the Royal Statistical Society: Series D (The Statistician) 46(2), 261–283 (1997)
2. Amaratunga, D., Cabrera, J., Shkedy, Z.: Exploration and Analysis of DNA Microarray and Other High-Dimensional Data. Wiley Series in Probability and Statistics. Wiley (2014)
3. Apolloni, B., Gentile, C.: Sample size lower bounds in PAC learning by algorithmic complexity theory. Theoretical Computer Science 209(1-2), 141–162 (1998)
4. Assent, I., et al.: DUSC: Dimensionality Unbiased Subspace Clustering. In: ICDM, pp. 409–414 (2007)
5. Beleites, C., et al.: Sample size planning for classification models. Analytica Chimica Acta 760, 25–33 (2013)
6. Blumer, A., et al.: Learnability and the Vapnik-Chervonenkis dimension. J. ACM 36(4), 929–965 (1989)
7. Boonyanunta, N., Zeephongsekul, P.: Predicting the Relationship Between the Size of Training Sample and the Predictive Power of Classifiers. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 529–535. Springer, Heidelberg (2004)
8. Bozdağ, D., Kumar, A.S., Catalyurek, U.V.: Comparative analysis of biclustering algorithms. In: BCB, Niagara Falls, pp. 265–274. ACM, New York (2010)
9. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics. Springer (2011)
10. Cai, T., Shen, X.: High-Dimensional Data Analysis (Frontiers of Statistics). World Scientific (2010)

11. Cheng, Y., Church, G.M.: Biclustering of Expression Data. In: Intelligent Systems for Molecular Biology, pp. 93–103. AAAI Press (2000)
12. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. J. Machine Learning Res. 7, 1–30 (2006)
13. Deng, G.: Simulation-based optimization. University of Wisconsin–Madison (2007)
14. Dobbin, K., Simon, R.: Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6(1), 27+ (2005)
15. Dobbin, K.K., Simon, R.M.: Sample size planning for developing classifiers using high-dimensional DNA microarray data. Biostatistics 8(1), 101–117 (2007)
16. Domingos, P.: A Unified Bias-Variance Decomposition and its Applications. In: IC on Machine Learning, pp. 231–238. Morgan Kaufmann (2000)
17. Dougherty, E.R., et al.: Performance of Error Estimators for Classification. Current Bioinformatics 5(1), 53–67 (2010)
18. El-Sheikh, T.S., Wacker, A.G.: Effect of dimensionality and estimation on the performance of gaussian classifiers. Pattern Recognition 12(3), 115–126 (1980)
19. Figueroa, R.L., et al.: Predicting sample size required for classification performance. BMC Med. Inf. & Decision Making 12, 8 (2012)
20. Fleiss, J.L.: Statistical Methods for Rates and Proportions. Wiley P. In: Applied Statistics. Wiley (1981)
21. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694 (2009)
22. Glick, N.: Additive estimators for probabilities of correct classification. Pattern Recognition 10(3), 211–222 (1978)
23. Guo, Y., et al.: Sample size and statistical power considerations in highdimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 11(1), 1–19 (2010)
24. Guyon, I., et al.: What Size Test Set Gives Good Error Rate Estimates? IEEE Trans. Pattern Anal. Mach. Intell. 20(1), 52–64 (1998)
25. Hand, D.J.: Recent advances in error rate estimation. Pattern Recogn. Lett. 4(5), 335–346 (1986)
26. Haussler, D., Kearns, M., Schapire, R.: Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In: IW on Computational Learning Theory, pp. 61–74. Morgan Kaufmann Publishers Inc., Santa Cruz (1991)
27. Hochreiter, S., et al.: FABIA: factor analysis for bicluster acquisition. Bioinformatics 26(12), 1520–1527 (2010)
28. Hocking, R.: Methods and Applications of Linear Models: Regression and the Analysis of Variance. Wiley Series in Probability and Statistics, p. 81. Wiley (2005)
29. Hua, J., et al.: Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21(8), 1509–1515 (2005)
30. Iswandy, K., Koenig, A.: Towards Effective Unbiased Automated Feature Selection. In: Hybrid Intelligent Systems, pp. 29–29 (2006)
31. Jain, A., Chandrasekaran, B.: Dimensionality and Sample Size Considerations. In: Krishnaiah, P., Kanal, L. (eds.) Pattern Recognition in Practice, pp. 835–855 (1982)
32. Jain, N., et al.: Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics 19(15), 1945–1951 (2003)
33. Kanal, L., Chandrasekaran, B.: On dimensionality and sample size in statistical pattern classification. Pattern Recognition 3(3), 225–234 (1971)
34. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97(1-2), 273–324 (1997)

35. Kohavi, R., Wolpert, D.H.: Bias Plus Variance Decomposition for Zero-One Loss Functions. In: Machine Learning, pp. 275–283. Morgan Kaufmann Publishers (1996)
36. Lissack, T., Fu, K.-S.: Error estimation in pattern recognition via Ldistance between posterior density functions. IEEE Transactions on Information Theory 22(1), 34–45 (1976)
37. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. Comput. Biol. Bioinformatics 1(1), 24–45 (2004)
38. Martin, J.K., Hirschberg, D.S.: Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests. Tech. rep. DICS (1996)
39. Molinaro, A.M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
40. Mukherjee, S., et al.: Estimating dataset size requirements for classifying DNA Microarray data. Journal of Computational Biology 10, 119–142 (2003)
41. Munkres, J.: Algorithms for the Assignment and Transportation Problems. Society for Ind. and Applied Math. 5(1), 32–38 (1957)
42. van Ness, J.W., Simpson, C.: On the Effects of Dimension in Discriminant Analysis. Technometrics 18(2), 175–187 (1976)
43. Niyogi, P., Girosi, F.: On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. Neural Comput. 8(4), 819–842 (1996)
44. Okada, Y., Fujibuchi, W., Horton, P.: A biclustering method for gene expression module discovery using closed itemset enumeration algorithm. IPSJ Transactions on Bioinformatics 48(SIG5), 39–48 (2007)
45. Opper, M., et al.: On the ability of the optimal perceptron to generalise. Journal of Physics A: Mathematical and General 23(11), L581 (1990)
46. Patrikainen, A., Meila, M.: Comparing Subspace Clusterings. IEEE TKDE 18(7), 902–916 (2006)
47. Prelić, A., et al.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinf. 22(9), 1122–1129 (2006)
48. Qin, G., Hotilovac, L.: Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat. Methods Med. Res. 17(2), 207–221 (2008)
49. Raeder, T., Hoens, T.R., Chawla, N.V.: Consequences of Variability in Classifier Performance Estimates. In: ICDM, pp. 421–430 (2010)
50. Raudys, S.J., Jain, A.K.: Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13(3), 252–264 (1991)
51. Sequeira, K., Zaki, M.: SCHISM: a new approach to interesting subspace mining. Int. J. Bus. Intell. Data Min. 1(2), 137–160 (2005)
52. Serin, A., Vingron, M.: DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach. Algorithms for Molecular Biology 6(1), 1–12 (2011) (English)
53. Singhi, S.K., Liu, H.: Feature subset selection bias for classification learning. In: IC on Machine Learning, pp. 849–856. ACM, Pittsburgh (2006)
54. Surendiran, B., Vadivel, A.: Feature Selection using Stepwise ANOVA Discriminant Analysis for Mammogram Mass Classification. IJ on Signal Image Proc. 2(1), 4 (2011)
55. Toussaint, G.: Bibliography on estimation of misclassification. IEEE Transactions on Information Theory 20(4), 472–479 (1974)

56. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer Series in Statistics. Springer-Verlag New York, Inc., Secaucus (1982)
57. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
58. Vayatis, N., Azencott, R.: Distribution-Dependent Vapnik-Chervonenkis Bounds. In: Fischer, P., Simon, H.U. (eds.) EuroCOLT 1999. LNCS (LNAI), vol. 1572, pp. 230–240. Springer, Heidelberg (1999)
59. Way, T., et al.: Effect of finite sample size on feature selection and classification: A simulation study. Medical Physics 37(2), 907–920 (2010)

Stream Clustering Algorithms: A Primer

Sharanjit Kaur, Vasudha Bhatnagar, and Sharma Chakravarthy

Abstract. Stream data has become ubiquitous due to advances in acquisition technology and pervades numerous applications. These massive data gathered as continuous flow, are often accompanied by dire need for real-time processing. One aspect of data streams deals with storage management and processing of continuous queries for aggregation. Another significant aspect pertains to discovery and understanding of hidden patterns to derive actionable knowledge using mining approaches. This chapter focuses on stream clustering and presents a primer of clustering algorithms in data stream environment.

Clustering of data streams has gained importance because of its ability to capture natural structures from unlabeled, non-stationary data. Single scan of data, bounded memory usage, and capturing data evolution are the key challenges during clustering of streaming data. We elaborate and compare the algorithms on the basis of these constraints. We also propose a taxonomy of algorithms based on the fundamental approaches used for clustering. For each approach, a systematic description of contemporary, well-known algorithms is presented. We place special emphasis on *synopsis* data structure used for consolidating characteristics of streaming data and feature it as an important

Sharanjit Kaur
Department of Computer Science, Acharya Narendra Dev College,
University of Delhi, Delhi, India
e-mail: sharanjitkaur@andc.du.ac.in

Vasudha Bhatnagar
Department of Computer Science, University of Delhi, Delhi, India
e-mail: vbhatnagar@cs.du.ac.in

Sharma Chakravarthy
Computer Science and Engineering Department,
University of Texas at Arlington, TX, USA
e-mail: sharma@cse.uta.edu

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_4

issue in design of a stream clustering algorithms. We argue that a number of functional and operational characteristics (e.g. quality of clustering, handling of outliers, number of parameters etc.) of a clustering algorithm are influenced by the choice of synopsis. A summary of clustering features that are supported by different algorithms is given. Finally, research directions for improvement in the usability of stream clustering algorithms are suggested.

1 Introduction

Data mining has seen enormous success in building and deploying models that have been used in diverse commercial and scientific domains. The technology, conceptualized in the late eighties of the previous century, was motivated by high growth rate of data repositories owing to rapid advances in data acquisition, storage, and processing technologies. Late nineties saw emergence of data sources that would continuously generate data resulting in potentially unending or unbounded *data streams*. Presently, such sources of *streaming* data are commonly found in network traffic, web click stream, power consumption measurement, sensor networks, stock market, and ATM transactions, to name a few. Size, varying input rates, and streaming nature of data have posed challenges to both database management (Chakravarthy and Jiang, 2009) and data mining communities (Hirsh, 2008).

Research efforts by the database community for dealing with streaming data have resulted in several approaches for Data Stream Management Systems or DSMS (Abadi et al., 2003; Arasu et al., 2004; Chakravarthy and Jiang, 2009). These systems are designed to process continuous queries over incoming data streams to satisfy user-defined Quality of Service (QoS) requirements (e.g., memory usage, response time, and throughput).

In contrast to processing stream data for monitoring applications, the goal of stream data mining is to identify significant, valid, and current patterns in a stream to gain insights into the behavior of the underlying system. Mining of data streams has been recognized as an important research area because of its wide applications in both scientific and commercial domains (Aggarwal, 2007; Babcock et al., 2002; Barbára, 2002; Domingos and Hulten, 2000; Gaber et al., 2005; Gama, 2010; Guha et al., 2000, 2002). The goals of data stream mining tasks are the same as those of mining static data, albeit in the context of possibly changing data distributions and input rates. Classification task in a data stream aims to classify unseen data in real-time using a predictive model trained from an evolving stream. The goal is to create a system in which the training model can adapt quickly to the changes in the underlying data stream (Aggarwal, 2007; Aggarwal et al., 2006). Frequent-pattern mining in streams aims to study data evolution by tracking changing status of item-sets (Cormode and Muthukrishnan, 2003; Fan et al., 2004; Gama, 2010; Giannella et al., 2003). Since an *infrequent* item-set may be-

come *frequent* over time and vice-versa, the storage structure needs to be dynamically adjusted to keep track of data evolution.

Clustering of data streams is performed to study time-changing groupings of data. Need to identify embedded structures from continuously growing streaming data has been the driving force behind the popularity of this data mining technique. Practical utility of stream clustering spans over a wide range of scientific and commercial applications. Scientific applications include weather monitoring, observation and analysis of astronomical or seismic data, patient monitoring for clinical observations, tracking spread of epidemics. Commercial applications cover e-commerce intelligence, call monitoring in telecommunications, stock-market analysis, and web logs monitoring.

The problem of clustering data streams was formally defined by Guha et al. in 2000 and *STREAM* was the first published algorithm to explicitly address this problem (Guha et al., 2002). The algorithm processes data stream in numeric data space using an approximation algorithm to obtain pre-specified number of clusters. In little over a decade, a large number of algorithms have been published[1]. Single scan of data, bounded memory usage, and data evolution are the key challenges for streaming clustering algorithms. Daniel Barbára (Barbára, 2002) has explicitly stated three basic requirements of clustering data streams, viz. compact representation of clusters, fast processing of incoming data points, and on-line outlier detection. In the previous decade, many algorithms have been designed which incorporate these requirements to efficiently mine clusters from data streams (Aggarwal et al., 2003; Cao et al., 2006; Charikar et al., 2003; Gao et al., 2005; Guha et al., 2002; Motoyoshi et al., 2004; Park and Lee, 2004; Tasoulis et al., 2006). However, mining of bursty streams with unpredictable input rates has not received as much attention as processing of data streams with uniform rate.

The goal of this chapter is to present a comprehensive account of the approaches used for clustering streams along with taxonomy, and categorize published algorithms systematically. A comparative analysis of approaches along with their characterization is presented. Within each category, selected algorithms are summarized along with their strengths and weaknesses with respect to core issues that arise because of no-second-look requirement of data streams. The choice of synopsis, which significantly influences functional and operational characteristics of stream clustering algorithms is elaborated. Two short surveys on the subject are available with limited coverage (Amini et al., 2011; Mahdiraji, 2009).

Two recent surveys in the area of stream clustering are particularly informative and demand mention. Survey by de Andrade Silva et al. (2013) provides a thorough discussion of the important aspects of stream clustering algorithms. It is an informative source of references to stream clustering applications, software packages and data repositories that may be useful for researchers and practitioners. The taxonomy presented in the article allows

[1] At the time of preparing the manuscript, a Google scholar search for *stream clustering algorithm* yielded 151,000 matches.

the reader to identify surveyed work with respect to important aspects of stream clustering. It also dwells upon the experimental methodologies used for assessing effectiveness of an algorithm. Amini et al. present a detailed account of density based stream clustering algorithms (Amini et al., 2014). Nearly twenty state-of-the art algorithms have been reviewed and chronology is presented. These algorithms are categorized in two broad groups called *density micro-clustering algorithms* and *density grid-based clustering algorithms*. Though the survey delineates algorithms under these two categories, it falls short of in-depth analysis of synopsis structure and its influence on the output clustering scheme.

The remainder of the chapter is organized as follows: Section 2 describes generic architecture of stream clustering algorithms. Section 3 discusses core issues in clustering of data streams. Common approaches for clustering streams are given in Section 4. Synopses used by contemporary algorithms are explained and compared in Section 5. Finally, Section 6 presents suggestions for strengthening stream clustering research to improve the applicability of these algorithms to stream data mining applications, and Section 7 consolidates the ideas and presents concluding remarks.

2 Architecture of Stream Clustering Algorithms

A data stream is defined as a set of d-dimensional data points $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_m}, \cdots$ arriving at time stamps $t_1, t_2, \cdots, t_m, \cdots$ where a data point $\mathbf{X_i} =< x_i^1, \cdots, x_i^d >$ is a d-dimensional vector. The time stamps may be either logical or physical, implicit or explicit. A data stream is potentially never-ending and the order of arrival of data points is unpredictable. Continuous arrival of data in an unbounded stream and a single scan constraint translate to the requirement of compact representation of data characteristics and fast processing of incoming data points.

A stream clustering algorithm is required to perform two basic tasks: i) process incoming data points to incorporate them into synopsis, and ii) generate clusters from the synopsis. This is typically handled by designing algorithms with two distinct components: on-line and off-line. The on-line component processes incoming data points and updates the synopsis, which is a compact representation of the data seen so far. Synopsis summarizes the data stream by including sufficient temporal and spatial information to discover clusters. The off-line component delivers clusters either periodically or on demand (by the user). Figure 1 illustrates generic architecture of a clustering algorithm with two components (described in Section 4). There exist some algorithms that are are capable of generating clusters in a single phase (Gao et al., 2005; Luhr and Lazarescu, 2009). However, such algorithms lack ability to handle fast streams.

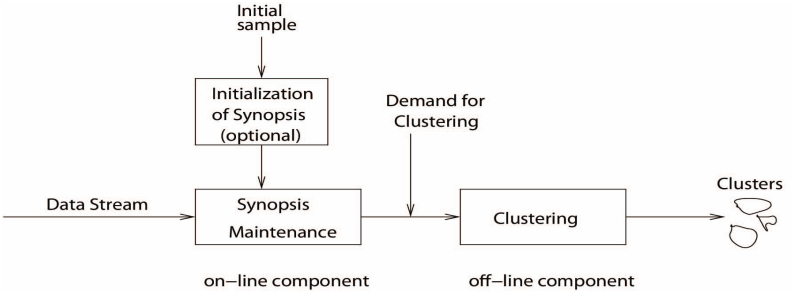

Fig. 1 Generic architecture for stream clustering algorithm

3 Issues in Clustering of Data Streams

This section discusses issues that arise as a consequence of constraint that arise because of unbounded, continuous flow of data. Some issues are inherited from algorithms used for clustering of static data, but get exacerbated in the context of stream processing. For example, handling of mixed attributes and deliverability of the expected relationship (hard or fuzzy) between clusters get accentuated in streaming environment due to the evolving nature of stream data characteristics.

3.1 Synopsis Representation

Due to the continuous nature of streaming data, it is not practical to look for point membership among the clusters discovered so far (Barbára, 2002; Garofalakis et al., 2002). This necessitates employing a synopsis, which summarizes the characteristics of incoming data points. Each incoming data point is processed and amalgamated into synopsis. which is used for building clustering scheme on user demand.

Design of synopsis is one of the critical issues in clustering data streams and must be done carefully, lest it would hamper meeting the goals of the clustering algorithm. Time to insert a data point into synopsis must be constant and strictly independent of its order of arrival. Further, insertion time should be predictable. This ensures that each data point is inserted into synopsis successfully and the resulting clustering scheme is representative. Format of the synopsis forms the backbone of an algorithm and influences functional characteristics of the algorithm, such as handling of outliers and quality of clustering; and operational characteristics, such as initialization, bounded memory usage and constant per-point processing time. Detailed discussion of these points is presented in Section 5.

3.2 Efficient Incremental Processing of Incoming Data Points

Incoming data points are processed by on-line component for insertion in synopsis. In order to avoid data loss, each point from the stream must be processed within a constant time period that is smaller than the inter-arrival time of points (Barbára, 2002; Gaber et al., 2005; Garofalakis et al., 2002). Bounding per-point processing time despite increasing size of the synopsis is a major challenge in design of the on-line component. This issue is auxiliary to the issue of synopsis design. Further, in order to prevent data loss, the on-line component must buffer incoming data points while the synopsis is being read by the off-line component for clustering.

Constant per-point processing time for a uniform data stream leads to accurate estimation of error due to predictable data loss. In case of bursty streams, an application dependent measure needs to be devised to minimize this error. Application of load shedding techniques used in stream data management (e.g., random and semantic load shedding) can be explored for clustering of bursty streams.

3.3 Handling of Mixed Attributes

Stream clustering applications, such as monitoring spread of illness, retrieving documents using text matching, network monitoring etc., require processing of both numerical and categorical data. Hence, the function computing similarities among points should be capable of efficient handling of mixed type attributes.

In case of mixed-type data streams, each categorical attribute is suitably transformed for similarity computation (Han and Kamber, 2005; Tan et al., 2006). This increases per-point processing time thereby causing potential data loss. Recently, several algorithms have been designed for clustering categorical and transactional data streams (Aggarwal and Yu, 2006; He et al., 2004; Li and Gopalan, 2006). Mixed data types have been handled by building histograms for categorical attributes and applying binning technique on numeric attributes in (He et al., 2004). However, systematic evaluation of the performance of these extensions is not yet available.

3.4 Capturing Recency and Data Evolution

Data distribution underlying a stream may change over time (Barbára and Chen, 2001; Gaber et al., 2005). In order to ensure recency of discovered clustering scheme, a mechanism is required to discount historical data in a continuous manner with minimum dragging effect on per-point processing time. Further, the mechanism must be sufficiently robust to handle changing arrival rates of data streams while ensuring that

no structure is lost before being included in the clustering model at least once. Such losses are likely in case of rapid change in data distribution, when newly formed clusters are discounted too early.

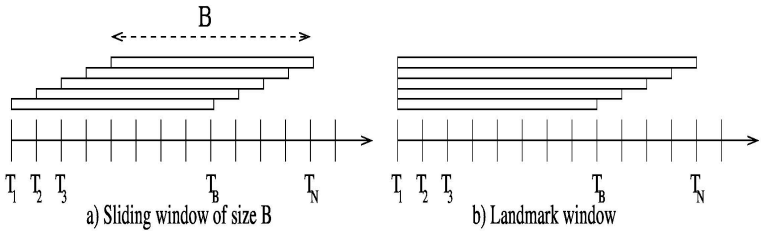

Fig. 2 *Window models* a) Sliding window of size B retains recent B data points and remains fixed in size b) Landmark window needs a land mark (T_1 in the figure) and grows in size with time

Sliding window, landmark window, and damped window models are commonly used to reduce the influence of obsolete data on the discovery of current structures (Aggarwal, 2007). In the *sliding window* model, each data point is time stamped and a data point expires after exactly B time stamps, where B is the size of window (Dang et al., 2009; Dong et al., 2003). The set of last B elements is considered recent (Figure 2(a)) and is used for model building. The window keeps on sliding by one time period to incorporate new data points and discard older points. Guaranteed constant per-point processing time makes this model attractive for on-line component.

In *landmark window* model, size of the window increases monotonically with time (Figure 2(b)) and all data points received within the window, since a predefined landmark are used in building the model (Aggarwal et al., 2003; Guha et al., 2000). This approach is less sensitive to data evolution as dominant older trends may overshadow the recent ones in the stream. As the window size grows, per-point processing time gradually increases making the model computationally more expensive. This model is not favoured in contemporary stream clustering algorithms.

The *damped window* model uses complete stream history to discover clusters, but assigns weight to each data point which is inversely proportional to its age. This is accomplished by using a fading function that periodically modifies the weights of data points (Aggarwal et al., 2004; Chen and Tu, 2007; Lu et al., 2005). Eventually, points with weight less than a threshold are ignored while the rest are used for building the model. Though fading factor is tuned on the basis of anticipated speed of the stream, it does not change dynamically. Consequently, some short-lived structures may go undiscovered. Computationally, damped window model fits in between the earlier two models. Per-point processing time is data dependent and therefore unpredictable. This is an adverse feature of this model.

Interestingly the choice of window model is influenced by the requirements of *completeness* and *recency* of clustering scheme. To ensure *completeness* of result, a stream clustering algorithm must consider every point received since last clustering for building the model. In contrast, an algorithm can capture *recency* only if it has an effective mechanism to *forget* older data. The sliding window model effectively captures *recency* of patterns, though *completeness* is not assured. On the other hand, the landmark window model guarantees *completeness* at the cost of capturing obsolete patterns. The damped window model balances between these two requirements, provided a fading factor is suitably chosen.

3.5 Hard vs. Fuzzy Clustering

Hard or exclusive clustering is an approach in which each data point is a member of exactly one cluster, whereas in fuzzy clustering, a data point may be a member of multiple clusters and clusters may overlap. In applications such as market basket analysis, network analysis, trend analysis, research paper acceptance analysis, a record belongs to exactly one of the several non-overlapping classes. Hence, algorithms used in these applications must place a record exactly in one cluster. In contrast, applications such as deciphering sloppy handwriting, image analysis are relatively tolerant towards imprecision, uncertainty, and approximate reasoning. This tolerance allows a point to be placed in multiple clusters leading to fuzzy or overlapping clusters (Baraldi and Blonda, 1999; Coppi et al., 2006; Kim and Mitra, 1993; Solo, 2008). Appropriateness of the expected relationship between clusters (hard or fuzzy) is application dependent.

In the context of data streams, generation of exclusive (hard) clusters requires the global picture of data captured by updated synopsis. A stream clustering algorithm can deliver hard clustering iff the synopsis is able to capture topological information about incoming points. Otherwise, in absence of original data the algorithm delivers overlapping clusters, which are approximations discovered from synopsis.

3.6 Detection of Outliers

Outliers in streams are those points which do not fit well into the clusters identified so far (Barbára, 2002). Often the role of outliers and clusters are interchangeable in streaming environment (Cao et al., 2006). As data streams evolve, new clusters emerge and old clusters fade out with time in unpredictable ways. Hence, a stream clustering algorithm must incorporate a mechanism to discriminate emerging clusters from outliers in an efficient manner without impacting the quality of the clustering scheme.

4 Approaches for Stream Clustering

After elaborating issues that must be resolved before designing a stream clustering algorithm, we describe various approaches that have been proposed for clustering streams. We focus on algorithms that use all points received in stream for clustering, and exclude sampling-based approaches, which merit exclusive consideration. Typically sampling-based approaches for stream clustering compute a small weighted sample of data stream, called *coresets*, on which approximation-based methods are applied for yielding net clustering scheme (Ackermann et al., 2010; Braverman et al., 2011).

Algorithms for clustering of data streams have been categorized into four groups based on the underlying approach for clustering (Figure 3). The figure also shows the algorithms associated with each approach using the color representation. Density-based approach is shown overlapping with both distance- and grid-based approaches in the figure. The rationale for this depiction is that computing density around a point (in density-based approach) requires the use of a distance function to determine neighborhood of a point. Grid-based approach uses the notion of density as the mechanism to identify structures in the data space without explicitly using distance. This explains the overlap of density and grid-based approaches. Statistics-based approach does not have anything common with the other three approaches and hence is shown separately.

It is natural that this categorization is similar to that of static clustering algorithms because differences in algorithms (for clustering static and streaming data) arise primarily due to: i) incremental processing of data, ii) maintenance of the synopsis, and iii) mechanisms to handle recency, all of which are specific to the characteristics of data streams.

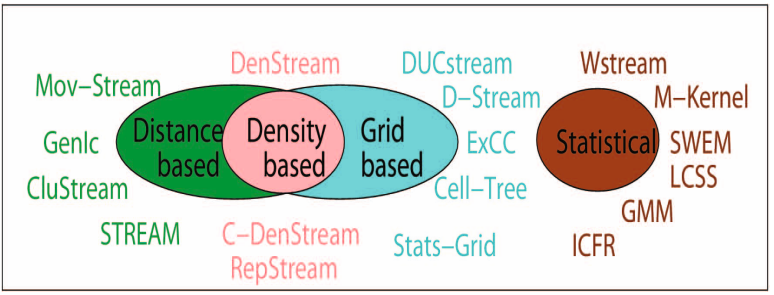

Fig. 3 Approaches for Clustering Data Streams

4.1 Distance- and Density-Based Approaches

Distance- and density-based approaches have been among the most popular approaches in static clustering. These approaches are equally popular for clustering of streaming data as well.

Distance-based approaches use a distance metric to place incoming data points into appropriate clusters (discovered so far). These approaches are known to generate convex clusters and are non-robust in the presence of outliers. Use of distance metric for similarity computation makes them suitable for numeric data. On the other hand, density-based approaches seek distance and density thresholds as user parameters and deliver arbitrary shaped clusters. Distance threshold is used for assessing the span of neighborhood and density threshold keeps check on the number of data points located in the neighborhood. Since density-based approach indirectly uses distance metric to assess the density in the neighborhood of a point, we place these two approaches in the same subsection.

These approaches are adapted for stream clustering by maintaining a synopsis suitable for incremental processing of incoming data. Initial sample of the stream is clustered as static data to generate a representative set of clusters, to which the incoming data points in the stream are assigned. Typically, the centroids/core-points of the representative set of clusters are used to generate net clustering scheme on user demand. Some well-known stream clustering algorithms using these approaches are described below.

4.1.1 STREAM Algorithm

STREAM algorithm is one of the earliest stream clustering algorithms (Guha et al., 2002). It performs clustering in single pass using distance-based approach. It employs a divide-and-conquer strategy for processing data points in stream and generates k optimal clusters using the landmark window model. Stream is treated as a sequence of *chunks* (batches) such that each *chunk* fits in the main memory. For every *chunk* of size n, distinct data points are found along with their respective frequencies. This computation leads to what is termed as *weighted chunk*.

An approximation algorithm (*localsearch*), which is a *Lagrangian* relaxation of *k-Median* problem, is applied on each weighted *chunk* to retain k weighted cluster centres. This constitutes synopsis for a batch. Weight of each cluster center is computed using the frequencies of the distinct points (say, m) as weights. Subsequently, the same algorithm is applied on the retained weighted cluster centres for each *chunk* to get optimal number of clusters. The algorithm is memory efficient and has $O(nm + nklogk)$ running time.

STREAM algorithm is particularly suitable for applications which require complete clustering. It finds exceptionally good solutions compared to *k-Means* and *BIRCH* (Zhang et al., 1996) algorithms with fixed synopsis size.

Divide-and-conquer strategy followed by the algorithm achieves a constant-factor approximation result with small memory requirement (Cao et al., 2006).

The main weakness of the algorithm is its assumption of stationary data stream and its inability to explicitly handle data evolution. The algorithm maintains a fixed number of cluster centres which can change or merge throughout its execution resulting in fuzzy clustering. Use of the landmark window does not allow clustering over varying time horizons.

4.1.2 CluStream Algorithm

CluStream algorithm (Aggarwal et al., 2003) uses distance-based approach and summarizes information about incoming points into micro-clusters (μCs). The μCs are temporal extensions of the *cluster feature vector* defined in *BIRCH* (Zhang et al., 1996), which is one of the earliest approaches for incremental clustering. Each μC summarizes information about member data points, and the set of μCs forms the synopsis representing data locality in stream at any point in time. These micro-clusters are used as pseudo-points to generate clusters. Clusters are generated over a user-specified time horizon using a pyramidal time frame. A micro-cluster is defined as follows.

Definition 1. A micro-cluster (μC) for a set of d-dimensional points $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_n}, \cdots$ with respective time-stamps $t_1, t_2, \cdots, t_n$ is defined as the quintuple $(\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n)$. The entries of the quintuple are defined as follows:

- For each dimension, $\overline{CF2^x}$ maintains the sum of squares of the data values in a vector of size d. The p-th entry of $\overline{CF2^x}$ is equal to $\sum_{j=1}^{n} (x_j^p)^2$.
- For each dimension, $\overline{CF1^x}$ maintains the sum of the data values in a vector of size d. The p-th entry of $\overline{CF1^x}$ is equal to $\sum_{j=1}^{n} x_j^p$.
- $CF2^t$ maintains the sum of the squares of time stamps $t_1, t_2, \cdots, t_n$.
- $CF1^t$ maintains the sum of the time stamps $t_1, t_2, \cdots, t_n$.
- n is the number of data points in the micro-cluster.

The space required to store a micro-cluster is $O(2 \cdot d + 3)$. During initialization phase of the algorithm, *k-Means* algorithm is applied to a sample taken from the early stream to generate a set of q μCs. This set forms the synopsis. The *on-line* component absorbs incoming data points into one of the q micro-clusters in the synopsis, based on a distance threshold. If the new incoming point cannot be absorbed into any of the existing μCs, then a new μC is created. Since it is imperative to ensure constant size of the synopsis, one of the following two actions are taken: i) either a μC with few points or least recent time stamp is deleted or ii) two μCs that are close to each other are merged. If a μC is deleted, then it is reported to the user as an outlier.

Pyramidal time frame is used to store snapshot of the synopsis at different time instances so that clusters can be discovered in a user-specified time horizon h. The *off-line* component of the algorithm discovers clusters by

applying *k-Means* on all μCs reported in horizon h. Subtractive property of μCs is exploited to generate higher-level clusters from the stored synopsis at different snapshots.

CluStream is the most popular stream clustering algorithm so far. Practically constant size synopsis makes the algorithm memory efficient and its main strength. Another positive feature of the synopsis is the strict upper bound ($O(q)$) on the processing time of an incoming point. Use of pyramidal time frame provides the users with a flexibility to explore evolution of clusters over different time periods. This functionality is useful in financial domain for applications such as stock monitoring, mutual funds comparison.

Sensitivity to the input parameters - number of μCs (q), number of final clusters (k) and distance threshold, is some of the limitations of the algorithm. Initialization phase of the algorithm induces a bias towards initial clustering scheme, which is another weakness. Use of a distance function for computing similarity precludes discovery of arbitrary shaped clusters. Further, appearance of an outlier in the stream may lead to creation of a new μC at the cost of a genuine but old cluster. Outlier handling is rather weak in the *CluStream* because sometimes a outlying point may displace a genuine cluster. Since the micro-clusters dynamically change with the data evolution in stream, the algorithm does not guarantee complete clustering.

4.1.3 DenStream Algorithm

DenStream proposed by Cao et al. (Cao et al., 2006) is a density-based stream clustering algorithm that handles evolving data streams using a damped window model. It extends the concept of μC to Potential-microcluster (P-μC) and Outlier-microcluster (O-μC), which are used in conjunction as synopsis. This extension is designed to capture the dynamics of evolving data stream where P-μCs capture the stable structures and O-μCs capture recent patterns and outliers.

The algorithm begins with an initialization phase during which *DBSCAN* algorithm (Ester et al., 1996) is applied to initial n points to generate P-μCs. Later, incoming data points are added to the nearest P-μC, iff their addition does not cause increase in the radius of the P-μC beyond a predefined threshold. If a point cannot be added to a P-μC, either a new O-μC is created or the incoming point is added to the nearest existing O-μC and its weight is computed. If its weight is greater than a threshold, then it is converted into P-μC. The algorithm uses a fading mechanism to reduce impact of older data on current trends. Weight of each P-μC is updated and checked periodically to ensure its validity. All invalid P-μCs are deleted as obsolete structures. When clustering is demanded, the off-line component applies a variant of the *DBSCAN* algorithm on the set of P-μCs.

This algorithm captures outliers effectively by periodically distinguishing between potential clusters and anomalies. Use of density-based approach for clustering allows discovery of arbitrary shaped clusters. Furthermore, as the

number of O-μCs may increase with time, a pruning strategy is used to delete real outliers after reporting them to the user, thereby ensuring complete clustering.

However, the algorithm requires several parameters to be supplied by the user such as maximum permissible radius of a μC, fading factor for pruning, threshold for distinguishing P-μCs from O-μCs. Since these parameters are sensitive to data distribution, capturing valid clustering scheme requires their tuning based on significant domain knowledge and experience, coupled with experimentation and exploration. Incorrect setting of these parameters may capture distorted patterns in streams lowering the quality decision making.

4.1.4 RepStream Algorithm

RepStream algorithm (Luhr and Lazarescu, 2009) is a single phase, graph-based clustering algorithm capable of discovering arbitrary shaped clusters. Graph-based clustering is particularly suitable for modeling of spatio-temporal relationships in data space since it preserves spatial relationship among data points. Instead of placing this algorithm under a distinct graph partitioning approach, we place it under distance- and density-based category for two reasons: i) graph partitioning directly or indirectly uses a distance metric for similarity and ii) literature available on stream clustering using graph partitioning is very limited.

The algorithm maintains two sparse graphs of connected *k-nearest neighbors* to identify clusters. The first graph is used to capture the connectivity relationships among the recently processed data points. Vertices in graph which meet density threshold are termed as *representative* points. The second graph is used to keep track of connectivity among the selected representative points. Clusters are formed around representative points. The representative points are further classified as *exemplars* and *predictors*. Exemplars represent persistent and consistent clusters whose information is stored in a knowledge repository, whereas predictors capture potential clusters. This differentiation between representative points facilitates handling of evolving nature of the data stream.

This algorithm uses complex data structures like *AVL-tree* and *KD-tree* for speedy updates and retrievals from the two graphs. The weight of each vertex in the graph is decayed with time and graph is pruned periodically in order to constrain memory usage and maintain recency. Pruning strategy takes into account both count of points and recency of the vertex. The oldest representative point with large count is considered more useful as compared to recent representative points with small count.

High computational expense is the main drawback of this algorithm. It arises because of the use of two tree-based data structures for the maintenance of graph. Pruning of graph may result into deletion of evolving patterns. Thus the mechanism used to capture current structures manifests as algorithm's deficiency in capturing drift in stream data. Tuning of three parameters –

minimum number of neighbors (k), density scale (α) and decaying factor (λ) – influences the quality of the resultant clustering scheme.

4.1.5 Other Algorithms

We briefly summarize other stream clustering algorithms which are either extensions or incremental modifications of the algorithms described earlier.

Generalized Incremental algorithm (*GenIc*) (Gupta and Grossman, 2004) is a single-phase algorithm that uses *k-Means* for clustering as in *CluStream*. The stream is divided into windows of fixed size. Incremental *k-Means* clustering algorithm is used for each window to assign points to k cluster centers, which are subsequently used for generating final clusters. The algorithm uses evolutionary techniques to improve its search for globally optimal solution to find k cluster centres. Evolutionary techniques (Eiben and Smith, 2007) generate solutions to optimization problems using processes of natural evolution, such as inheritance, mutation, selection, and crossover. These solutions are randomly updated and evaluated with a fitness function until no improvement is attained.

Mov-Stream algorithm (Tang et al., 2008) also uses *k-Means* for clustering and captures data evolution using the sliding window model. Each window is examined to identify different types of cluster movements like decline, drift, expand, and shrink. The algorithm focuses on evolution of individual cluster and falls short of giving an aggregated view of evolution in stream in user-defined time horizon.

C-denStream algorithm (Ruiz et al., 2009) is an extension of *DenStream*. It uses domain knowledge in the form of constraints and performs semi-supervised clustering. Domain knowledge is exploited for validating and evaluating clustering model for establishing its usefulness. However, prerequisite of labeled data to impose constraints makes this algorithm unsuitable for applications that lack labelled data. *HUE-Stream* (Meesuksabai et al., 2011) maintains synopsis as done in *CluStream* but with additional information using histogram for each feature within the synopsis to capture clustering structure evolution for heterogeneous data stream.

ClusTree algorithm (Kranen et al., 2010) relies on an index structure for storing and maintaining a compact view of the recent clustering scheme. It maintains cluster features similar to *CluStream* to store compact representation of all incoming data points. A hierarchy of micro-clusters is built using R-tree for retrieving information at different levels of granularity. The algorithm caters to very fast stream by maintaining aggregates of incoming points and inserting them in tree to avoid frequent insertions and data loss.

4.2 Grid-Based Approach

Grid-based approach for stream clustering dissects multi-dimensional bounded data space into a set of non-overlapping data regions (cells) to maintain detailed data distribution of the incoming points. It covers the data space with a grid to construct a spatial summary of the data and organizes the space *enclosing* the patterns (Akodjenou et al., 2007; Gama, 2010; Schikuta, 1996).

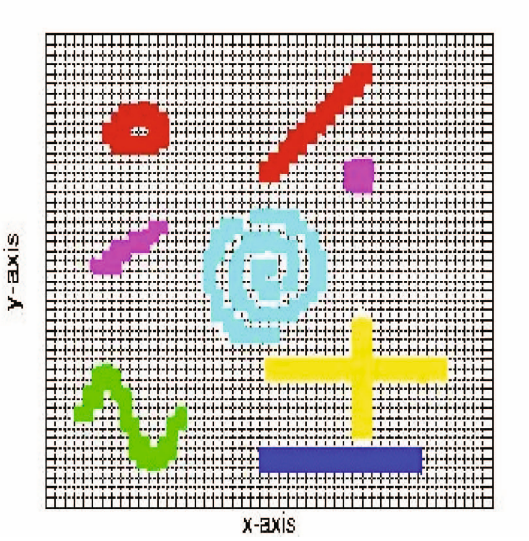

Fig. 4 Grid structure in 2-dimensional space with fixed granularity ($g = 48$) and eight arbitrary shaped clusters

While distance-based methods work with numerical attributes, grid-based methods elegantly handle attributes of mixed data types. This approach delivers arbitrary shaped clusters including shapes generated by mathematical functions as shown in Figure 4. This approach is robust with respect to outliers and noise in data. However, algorithmic parameters considerably influence the capability of identifying outliers and noise. Grid-based algorithms have been found to be faster as compared to distance- and density-based algorithms (Schikuta, 1996; Tsai and Yen, 2008; Yue et al., 2008). However, working in bounded data regions is sometimes considered as a limitation of this approach.

A trie implementation with either fixed, pre-specified granularity (g) or dynamically changing granularity is most popular data structure for maintaining the grid. Each leaf in the trie represents a d-dimensional hyper-cuboidal region (called a *cell*) containing one or more data points. The essential statistics such as sum of points, count of points etc. required for summarizing incoming stream, are maintained in each cell.

Grid-based methods use density as an indication of existence of structures in data. Density is assessed by the number of data points in a cell. Grid-based

approaches do not require distance computation and instead use density for clustering. Hence, they are shown overlapping with density-based approaches in Figure 3.

Although grid-based stream clustering algorithms do not require the user to specify the number of clusters, they require two crucial parameters viz. grid granularity and cell density threshold. Grid granularity (g) for an attribute determines the resolution of the data space and has marked influence on the quality of clustering. Numeric attributes are discretized using g, and categorical attributes are partitioned according to distinct values in their respective domains. A finely partitioned grid (large value of g) with smaller cells discovers clusters with more precise boundaries than those discovered in a grid with coarse granularity (Figure 5). Finer granularity grid allows multi-scale analysis with consequences on memory requirement (Gupta and Grossman, 2007). In a coarse granularity grid clustering quality suffers due to inclusion of larger sized data regions with grossly non-uniform distribution of data points (cells with solid dots in Figure 6).

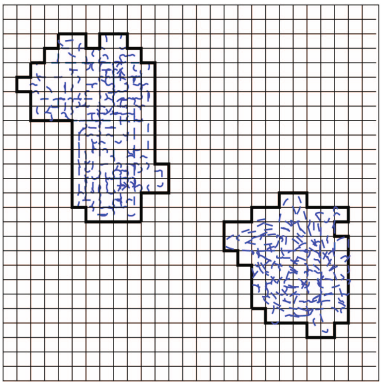

Fig. 5 *Fine granularity* captures precise boundary of clusters

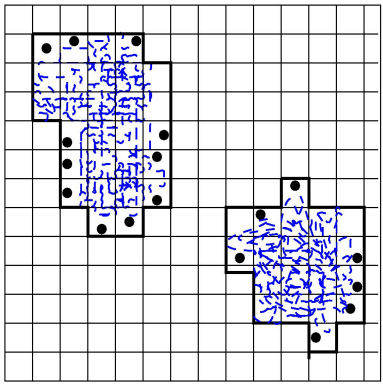

Fig. 6 *Coarse granularity* may include noisy regions at the periphery of clusters

Deciding the *right* value of g for an application is challenging for the end-user. The problem is accentuated in streaming data because of its evolving nature. Naturally, fine granularity is suitable for datasets with small clusters in close proximity whereas coarse granularity is favoured for datasets with well-separated clusters. Therefore domain expertise is critical for deciding grid granularity, in order to discover diligent clustering scheme. Use of dynamic grid precludes the need for specification of g. Prohibitive computational expense and requirement of a parameter for controlling the dimension splitting, however makes dynamic grids unattractive.

Cell density threshold (ψ) is used to discriminate between dense and non-dense cells. A cell with at least ψ data points is considered dense and is used in clustering. Since setting different values for ψ results in different clustering schemes, its value must be chosen carefully by the domain expert to avoid any ambiguity in the result. Some selected algorithms that use grid-based approaches are described below.

4.2.1 Stats-Grid Algorithm

Stats-Grid algorithm (Park and Lee, 2004) uses dynamically partitioned grid to generate arbitrary shaped clusters in streaming data. Incoming data points are inserted into the grid on the basis of spatial locality. The number of data points in a cell constitutes its support (density). When the support of a cell exceeds a predefined threshold, it is split into two cells on a dimension, selected on the basis of standard deviation. Density based splitting continues till the cell breaks down to a *unit* cell; that is, a cell whose length in each dimension is less than a predefined value. Recursive cell partitioning helps to maintain information about current trends at multiple granularity levels. To reduce the impact of historical data on current trends, cells are pruned on the basis of their support and their statistics are added back to the parent cell. At any time t, a cluster is a group of adjacent dense unit cells in a grid.

Dynamic partitioning of grid enables discovery of clusters at multiple granularity level. High density data regions yield clusters with fine boundary while relatively low dense regions result in clusters at a coarse level. On the flip side, dynamic partitioning of a grid cell may lead to repeated creation and pruning of large number of cells, which results in large memory footprint and computational overheads. Furthermore, since the size of each grid cell can be different, the cost of accessing a specific grid cell becomes high as only linear searching is possible. This degrades the performance of the method.

4.2.2 DUCstream Algorithm

DUCstream is a grid-based algorithm that treats a stream as a sequence of evenly sized chunks and uses connected component analysis for clustering (Gao et al., 2005). The algorithm deploys landmark window model to deliver clusters discovered in the data seen so far.

This algorithm uses a fixed granularity grid (FGG) as its synopsis. To begin with, distribution statistics of points in the first chunk are maintained in the grid. Subsequently, an initial representative clustering scheme is generated by applying connected component analysis on the dense cells of the first chunk. A cell is dense if its relative density (computed as $n/(m * t)$) is more than a user-defined threshold. Here n is the number of points in the cell, m is size of a chunk, and t is the number of chunks processed so far. For subsequent chunks, points are inserted in appropriate cells and distribution statistics are updated. If possible, the newly formed cells are merged with one of the

existing clusters; otherwise a new cluster is generated. After every chunk, clusters are updated by removing or merging existing clusters.

Deploying *FGG* as synopsis ensures predictable performance for processing of incoming points. The algorithm adapts to changes in the data stream by disregarding non-dense regions whose density fades over time. However, this may result in loss of emerging clusters. The algorithm requires three user-defined parameters, viz. size of the chunk, grid granularity, and cell-density threshold, choice of which may significantly influence the resultant clustering scheme.

4.2.3 Cell-tree Algorithm

Cell-Tree algorithm (Park and Lee, 2007) is an extension of the *Stats-Grid* algorithm (Park and Lee, 2004). The grid is recursively partitioned into fixed number of cells each time a point is inserted and cell statistics are distributed among them. As time elapses, these statistics are diminished by a predefined decay factor to maintain currency of clustering scheme. To achieve scalability, it introduces two novel data structures viz. *sibling-list* and *cell-tree*.

Sibling-list is used to manage grid cells in a one dimensional data space. It acts as an index for locating a specific grid cell. After a dense unit cell on one dimensional data space is created, a new sibling list for another dimension is created as a child of the grid cell. This process is recursively repeated for each dimension and it leads to a *cell-tree* with depth equal to data dimensionality. A unique path in the *cell-tree* identifies each dense unit grid cell. Clustering is performed by applying connected component analysis on dense cells.

Cell-Tree algorithm adapts to changing data distribution by partitioning high density regions or by merging adjacent cells which are decayed with time. The main draw back of this algorithm is the unpredictable per-point processing time because of dynamic partitioning of the grid. Main strength of the algorithm is that the delivered clusters have fine boundaries and capture natural shape in data space.

4.2.4 D-Stream Algorithm

D-Stream algorithm (Chen and Tu, 2007) partitions d-dimensional data space into $N = \Pi_{i=1}^{d} g_i$ density grids, where g_i is user-defined granularity of i^{th} dimension. Authors use the term 'grid' to denote d-dimensional hyper-cuboid region (*cell*). The on-line component of the algorithm maps each data point into a grid and updates statistical information stored as a characteristic vector. All grids are maintained in a data structure called *G-list*, which is implemented as a hash table. The hash table utilizes grid co-ordinates as its key and expedites look-up, updation and deletion.

A grid in *G-list* is categorized as dense, transitional, or sporadic based on its density and updation time. A grid is transitional if it is sparse and updated recently. Only dense and transitional grids are used in clustering, and sporadic

grids are pruned to ensure that *G-list* remains within memory bounds. The off-line component of the algorithm performs clustering dynamically based on the time required for transition of the nature of grids. Clusters are generated with dense grids as core, surrounded by other dense or transitional grids.

D-Stream algorithm generates high quality clusters by considering relationship between time horizon, decay factor and data density. It uses a novel strategy for controlling the decay factor and detecting outliers. Segregation of dense cells from transitional cells allows capturing of noise in the neighborhood of arbitrary shaped clusters. However, efficiency of the algorithm and and quality of the resultant clustering scheme depends on two crucial threshold parameters required for grid categorization and grid granularity. Usage of a hash table for fast accessing of grid cells requires additional memory for storage.

4.2.5 ExCC Algorithm

ExCC algorithm (Bhatnagar et al., 2013) delivers exclusive and complete clustering of data streams using fixed granularity grid (*FGG*) as synopsis. On-line component of the algorithm processes incoming data points and stores them in grid data structure based on their respective locations in data space. Off-line component of the algorithm performs connected component analysis to capture arbitrary shaped dense regions in the data space, reported as clusters. The algorithm is robust and requires no parameter other than grid granularity g. The salient feature of the algorithm is the speed-based (adaptive) pruning criterion. Thus, if speed of the stream increases (decreases), the pruning rate also increases (decreases). The pruning criterion also guarantees complete clustering, implying that all emerging clusters, howsoever small will be reported at least once in their lifetimes. This algorithm captures data drift by monitoring arrival patterns of anomalous data points, based on wait-and-watch policy.

Use of *FGG* results in constant per-point processing time for the on-line component and hence delivers predictable performance for stream processing. The algorithm is capable of detecting outliers and changes in data distribution. The algorithm delivers complete description of the cluster facilitating semantic interpretation. However, preciseness of delivered clustering scheme varies with grid granularity whose setting requires domain knowledge as mentioned earlier.

4.2.6 Other Algorithms

Several variants of *D-Stream* algorithm are designed to improve upon the functionality and overcome its weaknesses. .

The *DD-Stream* algorithm (Jia et al., 2008) accumulates points in grids, categorizes grids, and generates clusters periodically as done by the *D-Stream* (Chen and Tu, 2007) algorithm. It has an additional capability of handling data points on the grid borders by processing them periodically and absorbing them in the nearest and recent dense grid. Periodic processing balances the act of discarding border data points and immediate updating of grid boundary, which respectively affect cluster quality and efficiency of the clustering. The *MR-Stream* algorithm (Wan et al., 2009) is capable of discovering clusters at multiple resolutions whenever there is change in the underlying clustering scheme. *DENGRIS-Stream* algorithm (Amini and Wa, 2012) also works similar to *D-Stream* algorithm, but uses sliding window to maintain currency of result.

4.3 Statistical Methods Based Approach

Several stream clustering algorithms based on parametric and non-parametric statistical methods are available in literature. Clustering method based on parametric statistical techniques relies on initial sample to estimate the unknown distribution parameters. This estimate is used as approximation of true parameter value(s) for future computations.

Non-parametric statistical method does not make assumptions about underlying distribution while clustering, and commonly deploys *Density Estimation* (*DE*) approach. Given a sequence of identical, independent random variables drawn from an unknown distribution, the general density estimation problem is to reveal a density function of the underlying distribution. This probability distribution is then used to identify dense and sparse regions in a data set with local maxima of the probability distribution function taken as cluster centers (Sain, 1994). *Kernel-Density Estimation* (*KDE*) is a widely studied non-parametric *DE* method and is suited for data mining application because it does not make any assumptions about the underlying distribution. Most of the algorithms in this category use sliding window model to maintain recency of discovered clusters.

4.3.1 ICFR Algorithm

ICFR (Incremental Clustering using F-value by Regression Analysis) (Motoyoshi et al., 2004) claims to deliver more accurate clustering scheme compared to contemporary stream algorithms while treating the stream as sequence of chunks. Each chunk has a constant size over time axis. During the initialization phase, clusters are generated using similarity function applied on initial $(h-1)$ chunks where h is a user-defined parameter.

The synopsis consists of variance and co-variance matrix for each cluster and is updated after arrival of a chunk. The algorithm dynamically computes clusters over a fixed horizon and the size of synopsis is bounded by the number of points received in the horizon. For each cluster, center of gravity, variance, regression-coefficient, and F-value are computed using synopsis.

Clusters that are in close neighborhood are merged iteratively iff F-value of a newly merged cluster is bigger than F-value of each of the candidate cluster. This procedure is repeated until no more clusters can be combined and the resultant set is delivered as resultant clustering scheme. Each new chunk is processed in this way to discover set of clusters. These clusters are combined with existing valid clusters iteratively iff F-value is enhanced. Otherwise, clustering is done from scratch using all clusters formed so far using the new F-value.

However, assumptions about local regression in data make it unsuitable for real-world streaming data applications as they typically change with time. Further, the algorithm lacks a mechanism to capture evolving characteristics of a stream.

4.3.2 GMM Algorithm

GMM algorithm detects clusters using *Gaussian Mixture Model* which is a parametric method (Song and Wang, 2004). This algorithm incrementally updates the density estimates taking into account recent data and previously estimated density. This algorithm uses an *Expectation Maximization* technique for clustering and represents each cluster by its mean and covariance. Newly arrived points are merged with existing clusters by applying multivariate statistical tests for equality of covariance and mean.

Translation invariant property of covariance matrix is its main strength and benefits while determining the orientation of a cluster. However, the requirement of predefined number of clusters and assumption about multivariate normal distribution make this method impractical for real life, evolving data streams.

4.3.3 LCSS Algorithm

Song and Wang (Song and Wang, 2006) proposed an algorithm for detecting low complexity clusters by using *skewness* and *kurtosis*, in addition to mean and covariance. Multivariate skewness is a single non-negative number which characterizes the asymmetry of a probability distribution (PD) and hence represents asymmetry of clusters. Multivariate kurtosis is also a single non-negative number which is used to measure peakedness of a PD and indicates the concentration of a cluster. To generate clusters, *Expectation Maximization*

(EM) algorithm is applied and all required statistics are computed. The algorithm generates low complexity clusters and provides an accurate description of the shape of a cluster. In order to reduce the number of clusters, merging is performed in two phases . In the first phase, two clusters with comparable means and covariances are merged. Otherwise, skewness and kurtosis of the entire data in both clusters are tested against multivariate normality. If the normality is acceptable, then these two clusters are merged despite inequalities in their mean and covariance. This merging process is repeated until no more clusters can be combined.

Use of higher order statistics like multivariate skewness and kurtosis results in accurate description of the shape of clusters compared to lower order statistics. However, use of *EM* clustering technique make this algorithm computationally expensive and unsuitable for high dimensional data streams. Assumption of Gaussian Mixture Model to describe the data is somewhat incongruous in streaming environment.

4.3.4 M-Kernel Algorithm

Zhou et al. proposed *M-Kernel* algorithm for on-line estimation of probability density function with limited memory and in linear time (Zhou et al., 2003). The basic idea is to group similar data points and estimate the kernel function for the group. Each *M-Kernel* is associated with three parameters: weight, mean and bandwidth. Subsequently, the computed kernels are ranked to identify clusters.

This strategy is instrumental in keeping the memory requirement in control because if N data points in the stream are seen so far, then number of kernels (M) is much less i.e. $M << N$. The algorithm works for both landmark and sliding window models, although in the latter case, only an approximation is delivered. The algorithm has been tested for one dimensional data using *Gaussian Kernel* and requires bounded memory. Major limitation of this approach is that it processes only one dimensional data streams. This makes it impractical for use in real life streaming data applications. Further, the amount of memory used is highly sensitive to the underlying data distribution.

4.3.5 Wstream Algorithm

Wstream (Tasoulis et al., 2006) extends conventional *KDE* clustering to spatio-temporal data in stream environment using *Epanechnikow Kernel*. It uses multivariate *KDE* to handle multidimensional streams and captures dynamics of evolving streams by incrementally adjusting the kernels. The algorithm maintains a set of windows that define the clustering result at each time

point. The windows are (incrementally) moved, expanded, and contracted depending on the values of data points that join the clusters. These operations inherently take into account fading of older data by periodically computing the weight of windows and using it in the kernel function. In case a new data point arrives which does not belong to any of the existing windows (clusters), a new window is created with suitably initialized kernel parameters. Two windows that overlap considerably, are merged.

A major drawback of this approach is that with increase in the number of points more windows need to be maintained where each cluster is represented by at least one window. This makes the approach unsuitable for evolving data streams, wherein memory is a limiting factor.

4.3.6 SWEM Algorithm

SWEM (Dang et al., 2009) algorithm performs incremental and adaptive clustering of data stream based on Expectation Maximization technique (EM), using sliding window model. It is a soft clustering method, which is robust and capable of handling missing data. The algorithm works in two stages. In the first stage, *SWEM* scans incoming data points and summarizes them into a set of micro-components, where each one is characterized by some sufficient statistics. These micro-components are used in the second stage to approximate the set of global clusters for the desired time period.

Log-likelihood measure is used to evaluate the correctness of approximation in terms of set of micro components for each window. Small variation in these values for two consecutive time periods indicate no change whereas larger variation indicates change in data distribution. In the latter case, *SWEM* splits the micro component with the highest variance into two components to adapt to new distribution. Statistics of original micro component are distributed assuming stream follows normal multivariate distribution and each attribute is independent of each other. As the number of micro components maintained are fixed, splitting is followed by merging of two components that are close enough. For reducing the effect of obsolete data on recent clustering scheme, statistics are decayed using a fading factor.

Use of EM for clustering makes this algorithm computationally expensive, though stable and effective in domains despite the assumption about underlying distribution. The algorithm captures recent trends by employing a fading function. However, the algorithm is sensitive to outliers as the number of clusters maintained is fixed. Hence in presence of noise, clustering quality may deteriorate.

4.4 Discussion

We conclude this section by first stating the strengths and weaknesses of each approach. This discussion is followed by a feature-wise comparative analysis of selected algorithms.

4.4.1 Comparative Analysis of Approaches

Distance-based approach for clustering streams is simple and delivers fixed number of convex clusters, in contrast to varied number of arbitrary shaped clusters delivered by *density-based* approach. The former approach is sensitive to outliers and cannot distinguish between noise and pattern unlike the latter. As both approaches require initial phase for building synopsis for accumulating incoming data on-line, the delivered clustering scheme is biased towards initial synopsis and may take longer than desirable time to reveal true data evolution.

Accumulation of data in grid structure makes *grid-based* stream clustering approach independent of data ordering (Berkhin, 2006). This technique is also capable of discovering arbitrary shaped clusters and does not require user to specify the number of clusters. However, for high dimensional data the number of cells may increase exponentially and the grid may not fit in memory. Pruning strategies have to be necessarily employed to contain the grid in memory, resulting in occasional compromise on the quality of clustering. Grid-based approach is more suitable for spatio-temporal data as it also captures topological relationship along with the physical content of the data as per their natural ordering.

Parametric statistical approach must be used with utmost care in view of the presumed data distribution. On the other hand, *non-parametric statistical* approach for clustering is effective if data dimensionality is low and data belongs to a single distribution. Since in real life applications, data may be from a mixed distribution or may evolve with time, these approaches have limited utility. Density estimation approach overcomes some limitations and generates stable and accurate results. However, clustering results are influenced by the initial model generated using initial sample. High computational complexity of these approaches makes them unsuitable for streaming environment because of real time constraint for processing incoming data points. Typically statistics based approaches resort to batch processing.

4.4.2 Comparative Analysis of Selected Algorithms

Table 1 summarizes the algorithmic features of selected algorithms (shown in the top row) in chronological order.

CluStream and *DenStream* process incoming data on-line, and are capable of handling evolving data. *STREAM* algorithm, on the other hand, processes incoming data in batches. Thus, *STREAM* works on approximation of entire data stream from beginning to end without distinguishing between old and new data, whereas *CluStream* and *DenStream* deliver recent patterns. Use of pyramidal time frame in *CluStream* confers the capability and flexibility to explore the nature of evolution of clusters over user specified time period. *STREAM* and *CluStream* use distance-based approach for clustering and deliver convex clusters, whereas *DenStream* algorithm delivers arbitrary shaped clusters. *STREAM* algorithm minimizes sum of squared distance for clusters because of which extreme outliers with arbitrary large residuals have infinitely large influence on resulting estimates. In contrast, *CluStream* associates maximal boundary factor with each cluster to reduce this impact. *HUE-Stream* also uses distance-based clustering approach similar to *CluStream*, but maintains histograms as an additional cluster feature to capture clustering structure evolution in mixed data streams.

DUCstream, *Cell-Tree*, *DD-Stream*, *DENGRIS* and *ExCC* use *grid* structure for summarizing incoming points and require lesser per-point processing time as compared to distance-based algorithms. They use *connected component analysis* for coalescing adjacent dense cells in grid and hence generate clusters of arbitrary shape. *DUCstream*, *Cell-Tree* and *DD-Stream* use pruning and fading to handle the evolving stream. *ExCC* uses a speed-based criteria for pruning, which is purely data driven. This make *ExCC* free from any algorithmic parameter setting. Since *DD-Stream* distinguishes among dense, transitional, and sporadic cells, it captures outliers more effectively as compared to *DUCstream*, *Cell-Tree* algorithms. Similarly, *DENGRIS* differentiate between active and non-active clusters using sliding window model and discards clusters outside the specified window size. *ExCC* uses wait-and-watch policy to identify outliers in the stream. In case of severe data drift indicated by change of data space, it is capable of expanding the grid.

ICFR and *SWEM* use statistics-based approaches for generating non-overlapping clusters in data streams. Both require initial phase to set their synopses and use sliding window model for outdated data elimination. However, the delivered clustering scheme is sensitive to outliers. These algorithms are suitable for applications where data follows a specific distribution and hence have limited applicability.

5 Role of Synopsis in Stream Clustering Algorithms

Synopsis summarizes incoming data points and forms the backbone of any stream clustering algorithm. Structure of a synopsis must be chosen carefully as several functional characteristics, such as handling of outliers, type, and quality of clustering; and some operational characteristics, such as initialization, bounded memory usage, and constant per-point processing time are

Table 1 Comparison of Features of selected clustering algorithms for Streams

Algorithm Algorithm	STREAM	Clu Stream	Stat Grid	ICFR	DUC Stream	Den Stream	Cell Tree	DD Stream	SWEM	HUE Stream	DENGRIS Stream	ExCC
Features/Year	2002	2003	2004	2004	2005	2006	2007	2008	2009	2011	2012	2013
Nature of processing	Batch	Online	Online	Batch	Batch	Online	Online	Online	Online	Online	Online	Online
Pre-defined number of clusters	No	Yes	No	No	No	No	No	No	Yes	No	No	No
Initial phase	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	No	No
Support for on-demand clustering	No	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	yes
Evolution mechanism	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Clustering approach	Distance based	Distance based	Grid based	Statistical method	Grid based	Density based	Grid based	Grid based	Statistical	Distance based	Grid based	Grid based
Clustering technique	Distance based	Distance based	*CCA*	*RA*	*CCA*	Density based	*CCA*	*CCA*	*EM* based	Distance	CCA	CCA
Shape of cluster	C	C	A	E	A	A	A	A	C	C	A	A
Outlier detection	No	No	No	No	No	Yes	Yes	Yes	No	Yes	No	yes
Non-overlapping Clustering	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes

CCA: Connected Component Analysis, *RA*: Regression Analysis, *EM*: Expectation Maximization, *A*:Arbitraray, *E*:Elliptical, *C*:Convex

directly related to the synopsis used. In this section, we discuss impact of synopsis structure on these characteristics.

Mining summarized data maintained as in-core synopsis, meets important requirements of i) single scan of data, ii) processing each incoming point in real time to prevent data loss, and iii) minimizing input/output operations during clustering. Synopsis maintained by a stream clustering algorithm retains sufficient information about the data distribution so that underlying natural structures in the data can be revealed. Consequently, it plays a pivotal role in generating good quality clustering scheme and its design influences a number of functional and operational characteristics of the algorithm. Succinctly, synopsis with constant per-point processing time and bounded memory usage is critical in streaming environment.

Algorithms following statistics-based approaches for stream clustering maintain matrices/probability distribution functions etc. as synopsis, which is updated by organizing incoming data in batches. Algorithms in this category either assume a particular data distribution or use initialization phase to identify parameters/data distributions from an initial sample. It is established that this approach is computationally more expensive and works efficiently only for low dimensional data.

Grid and micro-clusters (μCs) emerge as two popular synopsis structures. Micro-clusters summarize information of incoming data points using distance computation whereas grid-based synopses discretize the entire data space to facilitate accumulation of similar points in topologically appropriate data regions. Though grid is traditionally taken to be of fixed granularity, researchers have experimented with dynamically partitioned grid. We analyze and compare below, three commonly used synopsis structures – Micro-clusters (μCs), Fixed granularity grid (FGG), and Dynamic granularity grid (DGG) – with respect to important capabilities of stream clustering algorithms. The discussion may be used as a set of guiding principles for the selection of synopsis in the design of a stream clustering algorithm.

5.1 Sensitivity of Synopsis towards Parameters

Parameter settings for a synopsis play an influential role on the quality of the resultant clustering scheme. Synopsis design with minimum number of parameters relieves the user from the responsibility of parameter specification and provides higher degree of insulation to the results.

Micro-clusters based synopsis requires at least two parameters: q (the number of μCs) and δ (the distance threshold to decide when to start a new μC). Since μCs are like pseudo-points, larger value of q results in better quality clusters. Further, high value of δ results in fewer number of clusters, each with reduced degree of similarity among members. This can possibly deteriorate the quality of the clustering scheme. On the other hand, a smaller value of δ leads to deletion of genuine but aged clusters and focuses on the recent

clustering scheme. The trade-off between quality and recency of a clustering scheme is decided by the end-user based upon application requirements.

Fixed granularity grid requires specification of grid granularity by the user. Fine granularity of grid leads to larger number of clusters than those obtained by coarse granularity. Though preciseness of clusters when seen in isolation is welcome, it looses relevance in conjunction with large number of small clusters. The trade-off between preciseness and the number of clusters is difficult to resolve even for an expert user in view of the changing data distributions with time. Another parameter commonly used by algorithms is the cell density threshold. Since only dense cells are coalesced to generate clusters, correct setting of this parameter is important for discriminating noise and clusters. In case of a dynamic grid, the user needs to specify a density threshold at which the cell is split and threshold for stopping cell partitioning. These parameters are similar to those used in fixed granularity grid, with similar influence on the resulting clustering scheme.

Thus, all three synopses require at least two parameters, which critically influence the output clustering scheme. Correct setting of these parameters presents difficulty to users. In authors' opinion, this limits usability of stream clustering algorithms for real-world applications.

5.2 Initialization of Synopsis

μCs based synopsis needs an initialization phase to determine initial set of micro-clusters for on-line processing of incoming data points. Typically, a predefined number of initial points from a stream are processed to yield initial set of μCs. After the initialization phase, incoming points in the stream are then assigned to the closest μCs in this set. One potential problem in this case is that the synopsis created during the initial phase creates a bias towards the initial set of data points. Hence these algorithms may not be able to detect data evolution quickly.

Grid-based synopsis does not need initialization because membership of a point is purely on the basis of its data values. In *FGG* and *DGG*, a data point is inserted into the appropriate hyper-cuboid on the basis of its location in the data space independent of the order of arrival. Use of pruning and fading functions ensure that data evolution is faithfully captured within reasonable time.

5.3 Ability to Capture Natural Structures in Data

Discovering natural structures in streams is the chief functionality of a clustering algorithm and it heavily depends upon the synopsis. μCs based synopsis uses distance function to assign incoming data points to a set of micro-clusters and hence leads to convex structures. Later, when the micro-clusters are combined, use of distance-based function again results in convex

shaped macro-clusters. Use of distance function at two levels can sometimes distort the natural structures in the data. Grid-based synopsis preserves topological distribution of data points in data space for both *FGG* and *DGG*. Consequently, this increases the ability of grid-based clustering approaches to capture spatial locality, and hence natural structures in the data.

5.4 Memory Usage

As the maximum number of μCs to be maintained is pre-specified and fixed (say, q), memory usage of micro-clusters based synopsis is bounded by $O(q)$. Grid synopsis has a comparatively larger memory foot print. The size of *FGG* is bounded by $O(g^d)$, where g and d are grid granularity and data dimensionality respectively. Memory requirement of *DGG* depends on the value of input parameters viz. splitting threshold for cell and the size of unit cell. Being data and input parameter driven, estimating the memory requirement in this case is complex.

In practice, memory requirement of *FGG* depends on the data distribution in the stream. Most authors of grid-based clustering algorithms have explicitly reported that the actual memory usage is much less than the theoretical upper bound of $O(g^d)$. This is because of the fact that memory is occupied only if atleast one data point exists in data region. Further, uniform distributions is very unlikely in high-dimensional space (Hinneburg and Keim, 1999; Yue et al., 2007). Pruning/fading function in grid based synopsis has dual role of maintaining recency and keeping the size of grid under control.

5.5 Per-point Processing Time

Constant per-point processing time is an important operational characteristic of the on-line component of a stream clustering algorithm. The requirement that streaming data points need to be processed quickly without any data loss makes it crucial for reducing approximation error. In μCs based approach, the complexity of the stream processing component is $O(dq)$ where d is the number of dimensions and q is the maximum number of micro-clusters. Though processing time is constant, time required for computing membership of a data point is unpredictable because each micro-cluster in the synopsis needs to be examined for ascertaining membership.

In *FGG*, traversal of exactly d dimensions for the membership of a new data point yields constant per-point processing time of $O(d)$, which makes grid structure quite attractive for handling fast data streams. In *DGG*, each dense dimension is recursively partitioned until it is of the size of unit cell. Splitting is done on-line while adding data points to the grid. Because of its non-predictable per-point processing time, this approach is non-reliable for bursty data stream.

5.6 Sensitivity to Data Ordering

Clustering algorithms using μCs as synopsis are sensitive to the order in which data is processed because insertion of a data point in a μC updates its center. Continuous change of cluster centers affects the future memberships. Change in cluster centers with time results into overlapping clusters that fail to meet the requirement of hard or exclusive clustering.

On the other hand, accumulation of data points in grid structure is according to its value and preserves topological distribution in the data space. This makes grid-based clustering insensitive to data ordering unlike micro-cluster based approaches. As cells in a grid are independent units and a point is placed in exactly one cell, the grid-based synopsis leads to exclusive clustering.

5.7 Managing Mixed Attributes

Micro-cluster based approach can handle numeric attributes naturally because of distance computation for similarity assessment. Clustering of categorical data streams however requires specialized handling. In order to work with mixed attributes, two alternatives are possible. First, categorical attributes are converted to numeric values before numeric distance measures are applied. This may be semantically inappropriate in some applications. The second approach is to discretize numeric attributes and apply categorical clustering method. This alternative is also not flawless as it is a known fact that discretization leads to loss of information. Further, real time transformation of data to conform to distance function incurs overheads that are undesirable and unacceptable in stream processing. Thus μCs based synopsis are not recommended for mixed data streams.

In contrast, grid structure works well with mixed attributes types. It is naturally favourable for handling mixed attributes as attribute space is quantized for processing data. For numeric data granularity is specified by the user, while for categorical data granularity is same as the size of the domain.

5.8 Handling of Outliers

Dynamic nature of data streams often leads to intermingling of outliers and structures in an unpredictable manner. Hence, a functional mechanism to distinguish between outliers and emerging structures is highly desirable in stream clustering algorithms.

On-line outlier detection is not feasible in μCs-based synopsis because of the underlying data structure and the similarity function used for membership. Since a fixed number of micro-clusters is maintained, there is a possibility of an outlier getting absorbed in one of the existing cluster centers. To overcome this problem, minimum distance threshold is used for absorbing a

point to an existing μC. If the distance exceeds the threshold, then a new μC is created with the current point as the only member by replacing the oldest μC. In case of unexpected disturbance in the data generation process, more than one outliers may displace genuine but aged clusters in the clustering scheme. This problem can be overcome by placing outliers in temporary μCs which are observed periodically before reporting them as genuine outliers.

Grid-based synopsis facilitates reporting an outlier on-the-fly if it lies outside the data space. All points in sparse cells are reported as noise. An alternative mechanism is to observe the arrival patterns of points in its constituent cell in the grid before reporting them as anomalous points. Thus outlier handling is more elegant in grid synopsis.

5.9 Capturing Data Evolution

μC-based approach handles data evolution by replacing existing μCs by new ones to capture changes in data characteristics. Even if changes are short lived, existing established μCs are replaced because of the constraint of maintaining fixed number of structures. This may result in loss of genuine but no-so-recent clusters or emerging clusters. This loss is clearly undesirable.

Grid-based synopsis handles data evolution more generously. New hypercuboids are inserted into the grid structure to capture changes in data characteristics and stale (old) cells are pruned to remove obsolete information from the grid. As each cell in the grid is an independent entity, insertion and removal of cells does not undesirably perturb others.

5.10 Summary

From the above discussion, it is clear that selection of synopsis is an important issue in design of stream clustering algorithm. Improper selection would hamper the achievement of goals of a clustering algorithm. Table 2 summarizes functional and operational characteristics of three synopses compared above. It is apparent from the table that grid structure addresses most of the issues of clustering data streams and is less restrictive and more versatile as compared to micro-cluster based synopsis in terms of both operational and functional characteristics of a stream clustering algorithm. However, dynamic grid with non-predictable per point processing time does not fulfill the main requirement for clustering data streams.

Table 2 Comparison of synopses

Characteristics (Type)/Synopsis	μC	FGG	DG
Detection of Inherent, Natural Patterns (F)	No	Yes	Yes
Sensitivity to Data Ordering (F)	Yes	No	No
Hard / Exclusive Clusters (F)	No	Yes	Yes
Outlier Detection (F)	Yes	Yes	Yes
Data Evolution (F)	Yes	Yes	Yes
Initialization required (O)	Yes	No	No
Parameters for Synopsis (O)	Two	One	Two
Per-Point Processing Time (O)	$O(dq)$	$O(d)$	$O(dT)$
Memory Usage (O)	$O(q)$	$O(g^d)$	$O(\lambda^d)$

F: Functional Characteristic, O: Operational Characteristic, μC: Micro-cluster, FGG: Fixed Granularity Grid, DG: Dynamic Grid, d: No. of dimensions, q: No. of micro-clusters, g: Grid granualrity, T: Time for accessing and splitting a dimension, λ: smallest pre-requisite unit length of the dimension

6 Further Issues and Challenges for Stream Clustering

So far we have objectively compared various approaches and algorithms with respect to issues and features. We present below few observations and propose specific directions that may help the research community to enhance the utility of stream clustering algorithms in practical data mining applications.

6.1 Weak Experimental Evaluation

Although substantial research has been carried out leading to the development of several effective stream clustering algorithms, experimentation is somewhat limited in scope and unimaginative in most works. Different implementations of an algorithm could lead to strikingly different results for various datasets and parameters. With numerous competing algorithms claiming improvement over the previously proposed algorithms, solid experimental evaluation is highly desirable. In particular following issues emerge from the existing literature on stream clustering algorithms.

1. Though experimental studies carried out in (Aggarwal et al., 2003; Bhatnagar et al., 2013; Cao et al., 2006; Luhr and Lazarescu, 2009; Park and Lee, 2007) are extensive, they are not oriented towards real-world applications. Careful study of the experimental analysis of these papers reveals that the applications for which stream clustering algorithms are evaluated are somewhat limited, giving rise to doubts whether it is realistic to apply these algorithms in applications, such as weather

monitoring, stock trading, telecommunication, web-traffic monitoring etc., as mentioned in these papers[2].

2. Non-availability of benchmark datasets, is a prominent reason for this situation. Most algorithms convincingly establish proposed functionality and validity of results using synthetic datasets (Aggarwal et al., 2003; Bhatnagar et al., 2013; Cao et al., 2006; Chen and Tu, 2007; Jia et al., 2008; Park and Lee, 2004, 2007). Intrusion detection data set (KDD Cup) and Forest cover data set are the most popular public data in stream clustering research because of their sizes. Both these datasets have mixed attributes, of which the numerical attributes are conveniently selected for most experimental work. While KDD cup data can be used for demonstrating cluster evolution by some stretch of imagination, Forest cover data set is not appropriate. There is little scope for testing data evolution in a meaningful manner on this data set. Absence of comparative analysis of algorithms using benchmark datasets diminishes both the utility and their applicability to real-world problems.
3. Inadequacy of the experiments also arises because of the unavailability of implementation of the published algorithms. Comparative analysis of existing algorithms with a proposed algorithm is possible and reliable only when it is conducted using the original implementation (of an existing algorithm). However, too few publications use original implementation of the competing algorithms.
4. Our third apprehension is regarding the commonly used *purity* metric used as evidence of cluster quality. Purity of a cluster is computed as

$$P_i = \frac{\rho_i^{dom}}{\rho_i} \times 100\%, \tag{1}$$

where ρ_i is the number of points in i^{th} cluster, and ρ_i^{dom} is the number of points with dominant class labels. Average purity of the clustering scheme at different time horizons has been reported in many algorithms (Aggarwal et al., 2003; Bhatnagar et al., 2013; Cao et al., 2006; Park and Lee, 2007). Since purity of clustering scheme increases with number of clusters, maximum purity of 100% attained when each record is treated as one cluster. As an example, in *DenStream* vs. *CluStream* cluster quality experiments, *DenStream* is claimed to be always better. However neither of the experiments mention the number of clusters obtained. This limitation of the metric can be somewhat reduced by using a weighted sum of individual cluster purity which is computed as

$$\mathcal{P} = \frac{1}{k}\sum_{i=1}^{k} \frac{P_i * \rho_i}{N} \tag{2}$$

[2] The only documented case study, where authors applied stream clustering, is weather monitoring with the restriction of processing data in chunks (Motoyoshi et al., 2004)

where N is total points in clustering scheme with k clusters.

Sensitivity of the cluster quality to the synopsis parameters is inadequately investigated in most of the works. A systematic study of the relationship between the parameters and data characteristics is necessary to determine the suitability of the proposed algorithm in different scenarios.

Similar situation was faced by the Frequent Itemset Mining (FIM) community, about a decade ago. The zoo of algorithms with competing and (sometimes) contradicting claims, motivated workshops in ICDM'03 (FIMI, 2003) and ICDM'04 (FIMI, 2004). The highlight of the workshops was code submission, along with the paper describing the algorithm and a detailed performance study by the authors on publicly provided datasets. The submissions were tested independently by the members of the program committee on test datasets which were made public until after the submission deadline. This unique exercise, resulted in a useful repository of datasets and algorithm implementations (Goethals, 2013). Though the number of published stream clustering algorithms is much less than FIM algorithms, proactive step in same direction will consolidate the research and benefit the community.

6.2 Usability

Our second observation is related to the usability of stream clustering algorithms. Although the survey by de Andrade Silva et al. (2013) does refer to several emerging applications, these laboratory experiments do not instil confidence in the end-user to select an appropriate algorithm for the task at hand.

As is evident from the discussion in Section 4.4, a number of functional and operational characteristics of stream clustering algorithm are determined by the synopsis structure used in the design. Thus, if a user desires functional characteristics X and operational characteristics Y, it may not be possible to meet the requirements by any single algorithm. For instance, if a user desires arbitrary shaped, fixed number of clusters with outlier detection, then neither *CluStream* nor *DenStream* can satisfy all three requirements.

To improve usability, it should be possible for the user to articulate functional and operational requirements for the application. Table 3 shows some exemplar applications, which typically demand clustering requirements shown in column I. For the specified requirements, an intelligent mechanism should be able to assemble on-the-fly an algorithm by identifying the most appropriate synopsis structure and the clustering method. One such proposal for assembling of a stream clustering algorithm has been proposed in (Bhatnagar et al., 2009). Its component-based architecture empowers the end-user to choose the features of the algorithm given the functional and operational requirements and assembles the algorithm to maximum advantage. There is a need to look further into the matter and design intelligent

interfaces so as to encourage the use of stream clustering algorithms for real-world problems. Work in this direction serves to empower the end-user and ultimately encourage penetration of KDD technology to a wider audience in the long term.

Table 3 Example applications with specific clustering requirements

Clustering requirement	Applications	
Nature of clustering	Hard	Spread of illness, Network data monitoring Stock monitoring
	Fuzzy	Image analysis, Web mining
Type of data processed	Numeric	Stock monitoring, Tracking meteorological data
	Mixed	Network data monitoring, Web mining
Capturing data evolution	Important	Stock monitoring, Tracking meteorological data, Spread of illness, Network data monitoring
Completeness of clustering	Required	Patient monitoring, Stock monitoring
	Not-required	Tracking meteorological data, Network data monitoring, Web content mining
Nature of stream processed	Uniform	Sensor monitoring, Tracking meteorological data
	Bursty	Network data monitoring, Web usage mining

6.3 Change Modeling

We strongly feel that one interesting and useful derivative of stream clustering algorithm is the modeling of changes in clustering schemes over temporal dimension. Often, the temporal changes in the model are more interesting than the model itself. Spreading of disease, change in buying patterns of customers, change in preferences etc. require analysis of the temporal changes in the clustering schemes.

In the recent past, changes in the clustering schemes have been examined in (Kaur et al., 2009; Kifer et al., 2004; Spillopoulou et al., 2006; Spinosa et al., 2007; Tasoulis et al., 2007; Udommanetanakit et al., 2007). But these are ad hoc and isolated works that need to be consolidated and formalized so as to design a 'change model' for the clustering scheme. The 'change model' would take as input the cluster features (centroid, density, size, shape, rate of growth etc.) that needs to be monitored for the changes and to quantify the change over the user-defined time horizon. By consolidating the changes in all the clusters in the clustering scheme, it is possible to infer the causal factors responsible for changes in the underlying data generation process. This derivative of stream clustering will find useful applications in management of health care systems, education systems, financial systems, e-governance etc. This, in our opinion, is an important direction for future research.

7 Conclusion

In this primer, we have presented alternative approaches used for clustering data streams, starting from the first documented proposal. This chapter does not merely enumerate the works in the light of basic requirements for clustering stream but attempts to abstract their key features with respect to the synopsis structure and discusses their strengths and weaknesses. Further, we indicate algorithm complexity where possible.

Detailed analysis of the features of the algorithms reveals that synopsis plays a pivotal role in generating good quality clustering scheme and its design influences a number of functional and operational characteristics of an algorithm. Three commonly used synopsis – micro-clusters based, fixed granularity grid and dynamic granularity grid structure – have been analyzed. It is found that there is no universal synopsis that can meet all application requirements and one is preferred over others under certain conditions (there is no silver bullet!).

At the end, we present a few inadequacies and propose directions of research to make stream clustering algorithms more useful for data mining applications. We hope that this work will serve as a good starting point and a useful reference for researchers working on the development of new stream clustering algorithms.

References

Abadi, D.J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., Zdonik, S.B.: Aurora: a new model and architecture for data stream management. Springer Journal on Very Large Databases 12(2), 120–139 (2003)

Ackermann, M.R., Lammersen, C., Märten, M., Raupach, C., Sohler, C., Swierkot, K.: Streamkm++: A clustering algorithm for data streams. In: The 2010 SIAM Workshop on Algorithm Engineering and Experiments, Texas, January 16, pp. 173–187 (2010), doi:10.1137/1.9781611972900

Aggarwal, C.C. (ed.): Data Streams: Models and Algorithms. Springer Science+Business Media (2007) ISBN: 978-0-387-28759-1

Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: The 2003 International Conference on Very Large Data Bases (VLDB), Germany, September 9-12, pp. 81–92 (2003) ISBN: 0-12-722442-4

Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for on-demand classification of evolving data streams. IEEE Transaction on Knowledge and Data Engineering 18(5), 577–589 (2006)

Aggarwal, C.C., Han, J., Yu, P.S.: A framework for projected clustering of high dimensional data streams. In: The 2004 International Conference on Very large Databases (VLDB), Canada, August 31-September 3, pp. 853–863 (2004) ISBN: 0-12-088469-0

Aggarwal, C.C., Yu, P.S.: A framework for clustering massive text and categorical data streams. In: The 2006 SIAM International Conference on Data Mining, Maryland, April 20-22, pp. 479–483 (2006) ISBN: 978-0-89871-611-5

Akodjènou-Jeannin, M.-I., Salamatian, K., Gallinari, P.: Flexible grid-based clustering. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 350–357. Springer, Heidelberg (2007)

Amini, A., Teh, Y.W., Saybani, M.R., Aghabozorgi, S.R., Yazdi, S.: A study of density-grid based clustering algorithms on data streams. In: The 2011 IEEE International Conference on Fuzzy Systems and Knowledge Discovery, China, July 26-28, pp. 1652–1656 (2011) ISBN: 978-1-61284-180-9

Amini, A., Wa, T.Y.: Dengris-stream: A density-grid based clustering algorithm for evolving data streams over sliding window. In: The 2012 International Conference on Data Mining and Computer Engineering, Thailand, December 21-22, pp. 206–211 (2012)

Amini, A., Weh, T.Y., Saboohi, H.: On density-based data streams clustering algorithms: A survey. Springer Journal of Computer Science and Technology 29(1), 116–141 (2014)

Arasu, Babcock, Babu, Cieslewicz, Datar, Ito, Motwani, R., Srivastava, and Widom: Stream: The stanford data stream management system. Technical Report 2004-20, The Stanford InfoLab (2004)

Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: The 2002 ACM Symposium on Principles of Database Systems, Wisconsin, June 3-5, pp. 1–58113 (2002) ISBN: 1-58113-507-6

Baraldi, A., Blonda, P.: A survey of fuzzy clustering algorithms for pattern recognition- part i and ii. IEEE Transactions on Systems, Man and Cybernetics 29(6), 778–801 (1999)

Barbára, D.: Requirements of clustering data streams. ACM SIGKDD Explorations 3(2), 23–27 (2002)

Barbára, D., Chen, P.: Tracking clusters in evolving data sets. In: The 2001 FLAIRS Special Track on Knowledge Discovery and Data Mining, Florida, May 18-20, pp. 239–243 (2001) ISBN: 1-57735-133-9

Berkhin, P.: A survey of clustering data mining techniques. In: Springer Grouping Multidimensional Data - Recent Advances in Clustering, pp. 25–71. Springer (2006)

Bhatnagar, V., Kaur, S., Chakravarthy, S.: Clustering data streams using grid-based synopsis. Springer Journal on Knowledge and Information System 41(1), 127–152 (2014)

Bhatnagar, V., Kaur, S., Mignet, L.: A parameterized framework for stream clustering algorithms. IGI International Journal for Data Warehousing and Mining 5(1), 36–56 (2009)

Braverman, V., Meyerson, A., Ostrovsky, R., Roytman, A., Shindler, M., Tagiku, B.: Streaming k-means on well-clusterable data. In: The 2011 ACM-SIAM Symposium on Discrete Algorithms, California, January 23-25, pp. 26–40. SIAM (2011), doi:10.1137/1.9781611973082.3

Cao, F., Ester, M., Qian, W., Zhou, A.: Density-based clustering over an evolving data stream with noise. In: The 2006 SIAM International Conference on Data Mining, USA, April 20-22, pp. 326–337 (2006), doi:10.1137/1.9781611972764.29

Chakravarthy,S., Jiang, Q.: Stream Data Processing: A Quality of Service Perspective. Springer (2009) ISBN 978-0-387-71002-0

Charikar, M., Callaghan, L.O., Panigrahy, R.: Better streaming algorithms for clustering problems. In: The 2003 ACM Symposium on Theory of Computing, California, June 9-11, pp. 30–38 (2003), doi:10.1145/780542.780548

Chen, Y., Tu, L.: Density-based clustering for real-time stream data. In: The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, August 12-15, pp. 133–142 (2007), doi:10.1145/1281192.1281210

Coppi, R., Gil, M.A., Kiers, H.A.L. (eds.): Data Analysis with Fuzzy Clustering Methods, vol. 51(1). Elsevier (2006)

Cormode, G., Muthukrishnan, S.: What's hot and what's not: Tracking most frequent items dynamically. In: The 2003 ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, San Diego, June 9-12, pp. 296–306 (2003), doi:10.1145/1061318.1061325

Dang, X.H., Lee, V.C.S., Ng, W.K., Ong, K.-L.: Incremental and adaptive clustering stream data over sliding window. In: The 2009 International Conference on Database and Expert Systems Applications, Austria, August 31-September 4, pp. 660–674 (2009), doi:10.1007/978-3-642-03573-9-55

de Andrade Silva, J., Faria, E.R., Barros, R.C., Hruschka, E.R., de Carvalho, A.C.P.L.F., Gama, J.: Data stream clustering: A survey. ACM Computing Surveys 46(1), 1–31 (2013)

Domingos, P., Hulten, G.: Mining High-Speed Data Streams. In: The 2000 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Maryland, August 20-23, pp. 71–80 (2000), doi:10.1145/347090.347107

Dong, G., Han, J., Lakshmanan, L.V., Pei, J., Wang, H., Yu, P.S.: Online mining of changes from data streams: Research problems and preliminary results. In: The 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams, San Diego, CA, June 8 (2003)

Eiben, A., Smith, J.: Introduction to Evolutionary Computing, 2nd edn. Natural Computing. Springer (2007)

Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: The 1996 AAAI International Conference on Knowledge Discovery and Data Mining, Oregon, August 2-4, pp. 226–231 (1996)

Fan, W., Huang, Y., Wang, H., Yu, P.S.: Active mining of data streams. In: The 2004 SIAM International Conference on Data Mining, Florida, April 22-24, pp. 457–461 (2004), doi:10.1137/1.9781611972740.46

FIMI, ICDM Workshop on Frequent Itemset Mining Implementations, FIMI 2003 (2003)

FIMI, ICDM Workshop on Frequent Itemset Mining Implementations, FIMI 2004 (2004)

Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Mining data streams: A review. ACM SIGMOD Record 34(2), 18–26 (2005)

Gama, J. (ed.): Knowledge Discovery From Data Streams. Chapman and Hall/CRC Press (2010) ISBN: 978-1-4398-2611-9

Gao, J., Li, J., Zhang, Z., Tan, P.-N.: An incremental data stream clustering algorithm based on dense units detection. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 420–425. Springer, Heidelberg (2005)

Garofalakis, M., Gehrke,J.,Rastogi, R.: Querying and mining data streams: you only get one look a tutorial. In: The 2002 ACM SIGMOD International Conference on Management of Data, Medison, USA, June 02-06, p. 635 (2002)

Giannella, C., Han, J., Pei, J., Yan, X., Yu, P.: Mining frequent patterns in data streams at multiple time granularities. In: Kargupta, H., Joshi, A., Sivakumar, K., Yesha, Y. (eds.) Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT Press (2003)

Goethals, B.: Frequent itemset mining implementation repository, `http://fimi.ua.ac.be/` (last retrieved in July 2013)

Guha, S., Mishra, N., Motwani, R., O'Callaghan, L.: Clustering data streams. In: The 2000 IEEE Annual Symposium on Foundation of Computer Science, California, November 12-14, pp. 359–366 (2000) ISBN: 0-7695-0850-2

Guha, S., Mishra, N., Motwani, R., O'Callaghan, L.: Streaming-data algorithms for high-quality clustering. In: The 2002 IEEE International Conference on Data Engineering, California, February 26-March 1, pp. 685–694 (2002), doi:10.1109/ICDE.2002.994785

Gupta, C., Grossman, R.L.: Genic: A single-pass generalized incremental algorithm for clustering. In: The 2004 SIAM International Conference on Data Mining, Florida, April 22-24, pp. 147–153 (2004), doi:10.1137/1.9781611972740.14

Gupta, C., Grossman, R.L.: Outlier Detection with Streaming Dyadic Decomposition. In: The 2007 Industrial Conference on Data Mining, Germany, July 14-18, pp. 77–91 (2007) ISBN: 978-3-540-73434-5

Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann (2006) ISBN 1-55860-901-6

He, Z., Xu, X., Deng, S., Huang, J.Z.: Clustering Categorical Data Streams. Computing Research Repository, abs/cs/0412058 (2004)

Hinneburg, A., Keim, D.A.: Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In: Proceedings of the 25th International Conference on Very Large Databases, Scotland, September 7-10, pp. 506–517 (1999) ISBN: 1-55860-615-7

Hirsh, H.: Data Mining Research: Current Status and Future Opportunities. Wiley Periodicals 1(2), 104–107 (2008), doi:10.1002/sam.10003

Jia, C., Tan, C., Yong, A.: A grid and density-based clustering algorithm for processing data stream. In: The 2008 IEEE International Conference on Genetic and Evolutionary Computing, USA, September 25-28, pp. 517–521 (2008), doi:10.1109/WGEC.2008.32

Kaur, S., Bhatnagar, V., Mehta, S., Kapoor, S.: Categorizing concepts for detecting drifts in stream. In: The 2009 International Conference on Management of Data, Mysore, December 9-12, pp. 201–209 (2009)

Kifer, D., David, S.B., Gehrke, J.: Detecting change in data streams. In: The 2004 International Conference on Very Large Data Bases, Canada, August 29-September 3, pp. 180–191 (2004)

Kim, Y.S., Mitra, S.: Integrated adaptive fuzzy clustering (iafc) algorithm. In: The 1993 IEEE International Conference on Fuzzy System, San Francisco, March 28-April 1, vol. 2, pp. 1264–1268 (1993), doi:10.1109/FUZZY.1993.327574

Kranen, P., Assent, I., Baldauf, C., Seidl, T.: The clustree: Indexing micro-clusters for anytime stream mining. Springer Journal on Knowledge and Information Systems 29(2), 249–272 (2010)

Li, Y., Gopalan, R.P.: Clustering transactional data streams. In: Proceedings of Australian Conference on Artificial Intelligence, Australia, December 4-8, pp. 1069–1073 (2006), doi:10.1007/11941439-124

Lu, Y.S., Sun, Y., Xu, G., Liu, G.: A grid-based clustering algorithm for high-dimensional data streams. In: Li, X., Wang, S., Dong, Z.Y. (eds.) ADMA 2005. LNCS (LNAI), vol. 3584, pp. 824–831. Springer, Heidelberg (2005)

Luhr, S., Lazarescu, M.: Incremental clustering of dynamic data streams using connectivity-based representative points. Elsevier Journal on Data and Knowledge Engineering 68(1), 1–27 (2009)

Mahdiraji, A.R.: Clustering data stream: A survey of algorithms. IOS Knowledge-Based and Intelligent Engineering Systems 13(2), 39–44 (2009)

Meesuksabai, W., Kangkachit, T., Waiyamai, K.: Hue-stream: Evolution-based clustering technique for heterogeneous data streams. In: The 2011 International Conference on Advanced Data Mining and Applications, China, December 17-19, pp. 27–40 (2011), doi:10.1007//978-3-642-25856-5-3

Motoyoshi, M., Miura, T., Shioya, I.: Clustering stream data by regression analysis. In: The 2004 ACSW of Australasian Workshop on Data Mining and Web Intelligence, New Zealand, pp. 115–120 (January 2004)

Park, N.H., Lee, W.S.: Statistical grid-based clustering over data streams. ACM SIGMOD Record 33(1), 32–37 (2004)

Park, N.H., Lee, W.S.: Cell trees: An adaptive synopsis structure for clustering multi-dimensional on-line data streams. Springer Journal of Data and Knowledge Engineering 63(2), 528–549 (2007)

Ruiz, C., Menasalvas, E., Spiliopoulou, M.: C-denstream: Using domain knowledge on a data stream. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 287–301. Springer, Heidelberg (2009), doi:10.1007/978-3-642-04747-3-23

Sain, S.R.: Adaptive Kernel Density Estimation. PhD thesis, Rice University (1994)

Schikuta, E.: Grid-clustering: An efficient hierarchical clustering method for very large datasets. In: The 1996 IEEE International Conference on Pattern Recognition, UK, August 23-26, pp. 101–105 (1996)

Solo, A.M.G.: Tutorial on fuzzy logic theory and applications in data mining. In: The 2009 World Congress in Computer Science, Computer Engineering and Applied Computing, USA, July 14-17 (2008)

Song, M., Wang, H.: Incremental estimation of gaussian mixture models for online data stream clustering. In: The 2004 International Conference on Bioinformatics and Its Applications, USA, December 16-19 (2004)

Song, M., Wang, H.: Detecting low complexity clusters by skewness and kurtosis in data stream clustering. In: The 2006 International Symposium on Artficial Intelligence and Maths, January 4-6, pp. 1–8 (2006)

Spillopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: Monic: Modeling and monitoring cluster transitions. In: The 2006 ACM International Conference on Knowledge Discovery and Data Mining, August 20-23, pp. 706–711 (2006), doi:10.1145/1150402.1150491

Spinosa, E.J., Carvalho, A.P., Gama, J.: Olindda: A cluster-based approach for detecting novelty and concept drift in data streams. In: The 2007 ACM Symposium on Applied Computing, March 11-15, pp. 448–452 (2007), doi:10.1145/1244002.1244107

Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education (2006)

Tang, L., Tang, C.-J., Duan, L., Li, C., Jiang, Y.-X., Zeng, C.-Q., Zhu, J.: Movstream: an efficient algorithm for monitoring clusters in evolving data streams. In: The 2008 IEEE International Conference on Granular Computing, China, August 26-28, pp. 582–587 (2008), doi:10.1109/GRC.2008.4664715

Tasoulis, D.K., Adams, N.M., Hand, D.J.: Unsupervised clustering in streaming data. In: The 2006 IEEE International Workshop on Mining Evolving and Streaming Data (ICDM), China, December 18-22, pp. 638–642 (2006), doi:10.1109/ICDMW.2006.165

Tasoulis, D.K., Ross, G.J., Adams, N.M.: Visualising the cluster structure of data streams. In: The 2007 International Conference on Intelligent Data Analysis, Slovenia, September 6-8, pp. 81–92 (2007)

Tsai, C.-F., Yen, C.-C.: G-TREACLE: A new grid-based and tree-alike pattern clustering technique for large databases. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 739–748. Springer, Heidelberg (2008)

Udommanetanakit, K., Rakthanmanon, T., Waiyamai, K.: E-stream: Evolution-based technique for stream clustering. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 605–615. Springer, Heidelberg (2007)

Wan, L., Ng, W.K., Dang, X.H., Yu, P.S., Zhang, K.: Density-based Clustering of Data Streams at Multiple Resolutions. ACM Transaction on Knowledge Discovery in Data 3(3), 1–28 (2009)

Yue, S., Wei, M., Li, Y., Wang, X.: Ordering grids to identify the clustering structure. In: The 2007 International Symposium on Neural Networks, China, June 3-7, pp. 612–619 (2007), doi:10.1007/978-3-540-72393-6-73

Yue, S., Wei, M., Wang, J.-S., Wang, H.: A general grid-clustering approach. Elsevier Pattern Recognition Letters 29(9), 1372–1384 (2008)

Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In: The 1996 ACM International Conference on Management of Data, Canada, June 4-6, pp. 103–114 (1996)

Zhou, A., Cai, Z., Wei, L., Qian, W.: M-kernel merging: Towards density estimation over data streams. In: The 2003 IEEE International Conference on Database Systems for Advanced Applications, March 26-28, pp. 285–292 (2003)

Cross Language Duplicate Record Detection in Big Data

Ahmed H. Yousef

Abstract. The importance of data accuracy and quality has increased with the explosion of data size. This factor is crucial to ensure the success of any cross-enterprise integration applications, business intelligence or data mining solutions. Detecting duplicate data that represent the same real world object more than once in a certain dataset is the first step to ensure the data accuracy. This operation becomes more complicated when the same object name (person, city) is represented in multiple natural languages due to several factors including spelling, typographical and pronunciation variation, dialects and special vowel and consonant distinction and other linguistic characteristics. Therefore, it is difficult to decide whether or not two syntactic values (names) are alternative designation of the same semantic entity. Up to authors' knowledge, the previously proposed duplicate record detection (DRD) algorithms and frameworks support only single language duplicate record detection, or at most bilingual. In this paper, two available tools of duplicate record detection are compared. Then, a generic cross language based duplicate record detection solution architecture is proposed, designed and implemented to support the wide range variations of several languages. The proposed system design uses a dictionary based on phonetic algorithms and support different indexing/blocking techniques to allow fast processing. The framework proposes the use of several proximity matching algorithms, performance evaluation metrics and classifiers to suit the diversity in several languages names matching. The framework is implemented and verified empirically in several case studies. Several Experiments are executed to compare the advantages and disadvantages of the proposed system compared to other tool. Results showed that the proposed system has substantial improvements compared to the well-known tools.

Keywords: Duplicate Record Detection, Cross Language Systems, entity matching, data cleaning, Big Data.

Ahmed H. Yousef
Ain Shams University
Computers and Systems Engineering Department, Cairo, Egypt
e-mail: ahassan@eng.asu.edu.eg

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_5

1 Introduction

Business intelligence and data mining projects has many applications in several domains. In the health sector, information retrieved from linked data is used to improve health policies with census data. Data aggregation is also used in crime and terror detection. In the higher education sector, business intelligence projects include the examples of aggregating scholar data from citation databases and digital libraries, aggregating students' data participating in eLearning and mobile based learning initiatives with their data in management information systems (Mohamed, 2008, El-Hadidi, 2008, Hussein et al., 2009, Elyamany and Yousef, 2013). All these applications are characterized by the huge volume, variety and velocity of data.

Billions of rows datasets are found in social networks. Data types and structures become more complex; with an increasing volume of unstructured data (80-90% of the data in existence is unstructured). The velocity of new data creation increased dramatically as well. It is necessary to ensure that this data is clean and consistent before it is converted to meaningful information, knowledge or wisdom using business intelligence, data mining or data science.

On the personal level, you always try to ensure that you have only one record for a friend at your mobile phone to minimize duplication. The existence of an application to detect this duplicate record will be useful either on the mobile or on the cloud. The same is needed for YouTube, it detects that the same video is uploaded several times to minimize storage requirements. In a social network or a professional network with hundreds of millions of records, several accounts for a person or an entity lead to a high probability of fake pages and duplicate accounts existence.

With the increasing number of users of social network sites every day, data explosion continues and grows with a globalization effect. These social networks (including Facebook, Google+ and twitter) attracted new users from all over the world. Professional social networks like LinkedIn are used by professionals to communicate to each other in multicultural projects. The globalization effect of these networks enabled users of Egypt, for example, to search for their friends or colleagues located in certain city in Sweden. At the same time, some organization analyzed professional social networks recently as a business intelligence tool to identify trends and gather statistics about certain groups of people (Yousef, 2012). These kinds of big data problems require new tools/technologies to store and manage these types of data that is available in several languages in order to achieve the business objectives.

Each natural language has its own specific variations to write down their entities names. For example, French person names are characterized by the existence of double first names. Non-European languages like Arabic contain complex names that consist of several words like Abdel Hamid, Aboul Wafa. Due to several factors including spelling and pronunciation variation, dialects and special vowel and consonant distinction and other linguistic characteristics, names can be

written in multiple format and transliterations although they are referring to the same entity. For example, in Arabic language, the names transliterations "Abdel Gabbar", "Abd Al Jabbar" and "Abd El Gabbar" are all equivalents. Even in the native Arabic language, "عبدالإله" ,"عبد الإله","عبد الاله" are equivalents.

The invention of a smart software program that takes all these variations into consideration is crucial in the era of duplicate record detection of big data. It is also important in information retrieval domain. Assume a user in Japan searching for his German colleague named Jürgen Voß. It will not be nice to instruct the Japanese person to use online German keyboard or to ask the German person to write his name in English. With the use of English language as the global one, the Japanese person will search Facebook using the English transliteration of the person name. i.e. Jurgen Voss. It is expected that Facebook takes into consideration these special character found in the German language and its English equivalent. Facebook should find the German person in a smart way. The following table shows some special characters in the German language and its English equivalents.

Table 1 Special characters in the German language and its English equivalents

German	English
ä	a, e
ö	o
ü	u
ß	ss

The same situation appears in French, Swedish, Danish and other European languages with more letters as shown in the following table.

Table 2 Special letters in the European languages

Language	Special Characters
French	à, â, ç, é, è, ê, ë, î, ï, ô,œ, ù, û, ü, ÿ
Spanish	á,é,í,ñ,ó,ú,ü
Swedish	ä,å,é,ö
Danish	å,æ,é,ø
Turkish	ç,ğ,ı,İ,ö,ş,ü
Romanian	ă,â,î,ş,□,ţ,□
Polish	ą,ć,ę,ł,ń,ó,ś,ź,ż
Icelandic	Á,æ,ð,é,í,ó,ö,þ,ú, ý

The situation becomes more difficult with Greek, Russian where the entire alphabet symbols are different than the English language. For non-European

language including Japanese, Chinese, Arabic, Persian and Hebrew, the situation becomes much more difficult.

Data like person names are not usually defined in a consistent way across different data sources. Data quality is compromised by many factors including: data entry errors, missing integrity constraints and multiple conventions for recording information.

The majority of names and their transliterations consist of one syllable. In some languages, the full name consists of several syllables (words). For example, in Dutch language, the name *van Basten*, *van der Sar* represent only the last names of two known football players. The last name is composed of several syllables. *Van* here is considered a prefix. Some names from several languages have postfixes as well. This includes *Ibrahimović and Bigović* with the postfix *"ović"*. The Russian names have many prefixes for male names as well. One of them is *"vsky"* as found in *Alexandrovsky* and *Tchaikovsky.* Another one is *"ov"* as found in *dimitrov* and *sharapov.* The prefix *"ova"* is used for Russian female names like *Miliukova and Sharapova*.

Arabic names can be composed of more than one syllable. These names have either prefixes like (*Abdel Rahman, Abd El Aziz, Abou El Hassan, Abou El Magd*) or postfixes like (*Seif El Din, Hossam El Din*). A list of used spelling variants of Arabic names prefixes in Arabic language, transliterated into English language, includes: *Abd, Abd Al, Abd El, Abdel, Abdol, Abdul, Abo, Abo El, Aboel, Abou, Abou El, Abu, Al, El*. Postfixes include: *El Din, El Deen, Allah* and others. In Arabic language, pronunciation and position of the character inside the string determines its presentation, for example, the character "ا"can be represented as (" أ , إ , ا ") based on pronouncing. Composite words such "ابو الفتوح" can be saved as a two words string or single word with no space character between them. The nature of the data entry process does not control these issues due to the lack of a unified standardized system. The problem of data representation becomes more difficult when transliteration for data is represented. This includes representing Arabic names in English, where the same name can have multiple transliteration representation. For example, the name ("عبد الرحمن") is equivalent to ("*Abd El rahman*", "*Abdul Rahman*" and many other transliterations).

Motivations

Detecting duplicate records in data sources that contain millions of records of data in several languages is a difficult problem. The currently available duplicate record detection tools have limitations to detect name variation in English, French, German, Dutch, Greek and/or Arabic. There is a need to have generic solution architecture that supports adding new language extensions (language specific similarity functions and phonetic algorithms) and machine based dictionaries. This architecture should be scalable enough to support the era of big data.

Contribution

The contributions of this chapter are significant for a number of reasons. Firstly, the paper compares two of the state of the art duplicate record detection tools and

then proposes a generic enhanced framework for cross language duplicate record detection based on rules and dictionaries. Furthermore, the paper proposes a strategy to automatically build names dictionary using enhanced localized Soundex algorithms. It compares the current available tools and its detailed components including similarity functions, classification algorithms, dictionary building components and blocking techniques. The effectiveness and efficiency metrics including accuracy, precision and reduction ratio are compared. Results show the power of the new proposed framework and highlight the advantages and disadvantages of the current available tools.

Chapter Outline
The remaining parts of this chapter are organized as follows: Section 2 presents an overview of the basic duplicate record detection background. Section 3 presents related work and mentions the different quality and complexity measures. Section 4 demonstrates the proposed framework, solution architecture and methodology. Section 5 represents the results of applying several experiments on the current available tools and the implemented framework. Analysis and discussion are also presented. Finally, section 6 concludes the chapter.

2 Duplicate Record Detection Overview

Big data practitioners consistently report that 80% of the effort involved in dealing with data is cleaning it up in the first place. Duplicate record detection is one of the data cleaning processes. It is the process of identifying records that have multiple representations of the same real-world objects. Sometimes duplicate records are caused by misspelling during data entry, in other cases the duplicated records are resulted from a database integration process.

Hence real-world data collections are exposed to be noisy, contaminated, incomplete and incorrectly formatted while being saved in database, data cleaning and standardization is a crucial preprocessing stage. In a data cleaning and standardization step, data is unified, normalized and standardized to be converted into a well-defined form. This step is done because original data may be recorded or captured in various, possibly obsolete formats. Data items of names and addresses are especially important to make sure that no misleading or redundant information is introduced (e.g. duplicate records). Names are often reported differently by the same person depending upon the organization they are in contact with, resulting in missing middle names or even swapped name parts.

Often, in the real world, entities may have two or more representations even in the same database. Usually, data entry operators add new records for a person when they fail to find the requested record due to misspelling mistake. With the fact that there are several valid spelling variations for persons' names, especially in non-English languages, the problem of detecting these extra records is recognized. Person name records contain errors that make duplicate record detection a difficult task. Errors are introduced as the result of transcription errors, in-complete information, lack of standard formats, or any combination of these factors.

The name matching and duplicate records detection problems become more difficult if person names are represented in one language in a record and in another language in another record. The new problem of cross language duplicate record detection (CLDRD) is a new one and is similar to the Cross Language Entity Linking that is defined in 2011 by (Paul McNamee, 2011). A new test collection is used to evaluate cross-language entity linking performance in twenty-one languages including Arabic and English (Paul McNamee, 2011). CLDRD supports the process of finding duplicate related records written in different languages using an automated system. This concept is used further in cross language information retrieval (Amor-Tijani, 2008).

The aim of duplicate records detection and record linkages is to match and aggregate all records related to the same entity (e.g. people, organizations or objects) (Winkler, 2006), (Goiser and Christen, 2006). Several frameworks were proposed to achieve these tasks (Köpcke and Rahm, 2010). Because duplicate record detection can be classified a special case of the record linking problem, it shares the same historical information. Computer-assisted data linkage goes back as far as the 1950s. Fellegi and Sunter put the mathematical foundation of probabilistic data linkage in 1969 (Fellegi, 1969).

What makes name matching a challenging problem is the fact that real-world data quality is low in most cases. Name matching can be viewed as related to the similarity search (wild card search). This chapter focuses on person entities, when the identifier is the person name.

In order to clean a database and remove duplicate records, many researches have been developed. The general steps for DRD are cleaning and standardization, indexing/blocking, record pair comparison and similarity vector classification.

The cost of duplicate record detection is affected by the number of generated pair of records to be compared. Computational effort of comparing records increases quadratically as database is getting larger. Many indexing/blocking techniques have been developed in order to minimize number of generated pair of records. Blocking splits the dataset into non-overlapping blocks. The main purpose is to reduce number of records that will be com-pared with conjunction of maximum number of correlated records within the same block such that only records within each block are compared with each other. In duplicate record detection the maximum number of possible comparisons N_c is defined as:

$$N_c = N*(N-1). \quad (1)$$

where N is the number of records in a dataset. Several indexing techniques exist including traditional blocking and sorted neighborhood blocking.

There are several similarity functions that can be used in duplicate record detection (Christen, 2006). They are used to compare the names in two records and return a similarity value. The similarity value ranges from 0 which means completely different to 1 which means completely equal. The similarity functions include numeric percentage and absolute comparisons for numeric fields. Special similarity functions that support range tolerance are used for age, date and time. Key-Difference similarity functions are used with telephone numbers and postal

codes. For string variables including person names, a plenty of similarity function exists. This includes exact string comparison, truncated string comparisons and approximate string comparison. These approximate string comparisons include Winkler, Jaro, Bag distance, Damerau-Levenshtein (Levenshtein, 1966), Smith-Waterman and many others.

Each pair of records will be classified as duplicates and non-duplicates according to the similarity value which is calculated by a similarity function. Computer based systems can classify high similarity value (larger than certain threshold) as duplicates and low value (lower than another certain threshold) as non-duplicates.

The record pairs with similarity values that are between the two thresholds are classified as possible duplicate. Therefore, a clerical review process is required where these pairs are manually assessed and classified into duplicates or non-duplicates.

For a data source A, the set of the ordered record pair resulting from cross joining the data source to itself AxA is the union of three disjoint sets, M, U and P (Christen and Goiser, 2007). The first set M is the matched set where the two records from A are equivalent. The second set U is the unmatched set where the two records from A are not equivalent. The third set P is the possible matched set. In the case that a record pair is assigned to P, a domain expert should manually examine this pair to judge if the record can be moved to either M or U.

There are many applications of computer-based name matching algorithms including duplicate records detection, record linkage and database searching that solve variations in spelling, caused for example by transcription errors. The success of such algorithms is measured by the degree to which they can overcome discrepancies in the spelling of names. In some cases it is not easy to determine whether a name variation is a different spelling of the same name or a different name altogether.

Spelling variations can include misplaced letters due to typographical errors, substituted letters (as in Mohamed and Mohamad), additional letters (such as Mohamadi), or omissions (as with Mohamed and Mohammed). This type of variations in writing names doesn't affect the phonetic structure of the name. These variations mainly arise from misreading or mishearing, by either a human or an automated device. Phonetic variations appear when the phonemes of the name are modified, e.g. through mishearing, the structure of the name is substantially altered.

Alternate first names problem appear in western languages. Also, a person name may change during the course of his life, usually when his marital state changes from single to married. Double names occur in some cases where names are composed of two syllable but both are not always shown. For example, a double name such as Philips-Martin may be given in full, as Philips or as Martin. In Arabic language, on the other side, such names such as Abdel-Hamid may be written Abdul Hamid or Abd El Hameed. Some of the difficulties associated with person names in Middle East languages including Arabic language were addressed in (Shaalan and Raza, 2007).

Monge and Elkan (Alvaro Monge and Elkan, 1996) proposed a token-based metrics for matching text fields based on atomic strings. An atomic string is a sequence of alphanumeric characters delimited by punctuation characters. Two atomic strings match if they are equal or if one is the prefix of the other. In (Yousef, 2013), Yousef proposed a weighted atomic token function to suit Arabic language and compared it to Levenshtein edit-distance algorithm. Although the traditional atomic token does not take into consideration the order of the two strings, the weighted atomic token takes it into account. Therefore, the performance was better than the classic Levenshtein edit-distance algorithm.

2.1 Phonetic Name Matching Algorithms

There are several phonetic name matching algorithms including the popular Russell Soundex (Russell, 1918, 1922) and Metaphone algorithms that are designed for use with English names. The ambiguity of the Metaphone algorithm in some words limited its use. The Henry Code is adapted for the French language while the Daitch-Mokotoff Coding method is adapted for Slavic and German spellings of Jewish names. The Arabic version of the Soundex algorithm is found in (Aqeel, 2006) and modified in (Koujan, 2008). Its approach is to use Soundex of conflating similar sounding consonants. However a special version of soundex for Arabic person names is proposed in (Yousef, 2013). This enhanced Arabic Combined Soundex Algorithm solved the limitation of the standard soundex algorithm with Arabic names that are composed of more than one word (syllable) like (Abdel Aziz, Abdel Rahman, Aboul Hassan, Essam El Din). These phonetic algorithms can be used as a name matching method. These algorithms convert each name to a code, which can be used to identify equivalent names.

2.2 Quality of the Duplicate Record Detection Techniques

The quality of record linking techniques can be measured using the confusion matrix as discussed in (Christen and Goiser, 2007). The confusion matrix compares actual matched (M) and non-matched (U) records (according to the domain expert) to the machine matched (M') and non-matched records (U'). Well known measures include true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP).

The measurement of accuracy, precision and recall are usually expressed as a percentage or proportion as follows (Christen and Goiser, 2007):

$$\text{Accuracy} = (TP+ TN) / (TP+FP+TN+FN). \quad (2)$$

$$\text{Precision} = TP / (TP+FP) \quad (3)$$

$$\text{Recall} = TP / (TP+FN) \quad (4)$$

Because the number of negatives TN is very large compared to the number of records in the comparison space, it is widely accepted that quality measures that

depend on TN (like accuracy) will not be very useful because TN will dominate the formula (Christen and Goiser, 2007).

3 Related Work

Many studies and frameworks have been presented to solve duplicate record detection problem, some of them were about presenting a complete framework or about developing enhanced technique for one of duplicate record detection stages. (Christen, 2006) Provides background on record linkage methods that can be used in combining data from different sources. The deduplication is a similar problem to record linking when the source and destination are the same. A thorough analysis of the literature on duplicate record detection is presented in (Elmagarmid et al., 2007) where similarity metrics and several duplicate detection algorithms are discussed.

The evaluation metrics of duplicate record detection system can be measured from two points of view. The first metric is complexity which is measured based on the number of generated record pairs and the reduction ratio RR. The second is the quality of the DRD results which can be measured by calculating the positive predictive value (precision) and true positive rate (Recall). These values are calculated using true positives, false positives, true negatives and false negatives.

In order to perform a join between two relations without a common key, it is needed to determine whether two specific tuples, i.e. field values are equivalent. Entity matching frameworks provide several methods as well as their combination to effectively solve different matching tasks. In (Köpcke and Rahm, 2010), eleven proposed frameworks for entity matching are compared and analyzed. The study stressed the diversity of requirements needed in such frameworks including high effectiveness, efficiency, generality and low manual effort.

There are three basic types of duplicate record detection strategies: deterministic, probabilistic and modern (Herzog et al., 2007). The deterministic approach can be applied only if high quality precise unique entity identifiers are available in all the data sets to be linked. At this level, the problem of duplication detection at the entity level becomes trivial. A simple database self-join is all that is required. However, in most cases no unique keys are shared by all records in the dataset, and more sophisticated duplication detection techniques need to be applied. In probabilistic linkage, the process is based on the equivalence of some existing common attributes be-tween records in the dataset. The probabilistic approach is found to be more reliable, consistent and provides more cost effective results. The modern approaches include approximate string comparisons and the application of the expectation maximization (EM) algorithm (Winkler, 2006).

Many Other techniques are explored including machine learning, information retrieval and database research. Some frameworks utilize training data to find an entity matching strategy to solve a given match task in a semi-automatic way. The quality of the computer duplicate detection process is found to be higher than the manual process (done by hands of humans) (Boussy, 1992).

Many algorithms are proposed that depends on machine learning approaches. Efficient and accurate classification of record pairs into matches and non-matches is considered one of the major challenges of duplicate record detection. Traditional classification is based on manually-set thresholds or on statistical procedures. More recently developed classification methods are based on supervised learning techniques. They therefore require training data, which is often not available in real world situations or has to be prepared manually by a time-consuming process. Several unsupervised record pair classification techniques are presented in (Christen, 2008b, Christen, 2008a). The first is based on a nearest-neighbor classifier, while the second improves a Support Vector Machine (SVM) classifier by iteratively adding more examples into the training sets.

The problem of record matching in the Web database scenario is addressed in (Weifeng et al., 2010). Several techniques to cope with the problem of string matching that allow errors are presented in (Navarro, 2001). Many fast growing areas such as information retrieval and computational biology depend on these techniques.

Machine transliteration techniques are discussed in (Al-onaizan, 2002, Knight, 1997, Amor-Tijani, 2008) for Arabic and Japanese languages. Finite state machines are used with training of spelling-based model. Statistical methods are used for automatically learning a transliteration model from samples of name pairs in two languages in (AbdulJaleel, 2003a, AbdulJaleel, 2003b). Machine translation could be extended later from text to speech as found in (Jiampojamarn, 2010). Co-training algorithms with unlabeled English-Chinese and English-Arabic bilingual text is used in (Ma and Pennsylvania, 2008). A system for Cross Linguistic Name Matching in English and Arabic is implemented in (Freeman et al., 2006, Paul McNamee, 2011). The system augmented the classic Levenshtein edit-distance algorithm with character equivalency classes.

With Big data tools like R, duplicate records are identified by several objects in several scenarios. For example, the read objects read a transaction data file from disk and creates a transactions object with the option to remove duplicates. However, the small variations of spelling will not be recognized by such systems.

Several tools are used for duplicate record detection in English language. These tools include TAILOR, Big-Match and Febrl. They use different techniques to identify any identical entities from one or more data sources in the same language. TAILOR (Elfeky et al., 2002) is a flexible toolbox which allows the users to apply different duplicate detection methods on the data sets. BigMatch is the duplicate detection program used by the US Census Bureau (Yancey, 2002). If the sizes of the datasets are large, online record linking can be used (Dey et al., 2011). Febrl (Christen, 2008c) provides data structures that allow efficient handling of very large data sets. Febrl includes a new probabilistic approach for improved data cleaning and standardization that support parallelization (Christen et al., 2004). The results of a survey of Febrl users are discussed in (Christen, 2009). However, FEBRL is an open source framework that has some usability and configuration limitations. It is needed to install FEBRL on a local machine and configure the operating system and prerequisite software to match FEBRL platform requirements.

Up to our knowledge and experiments with the current available aforementioned tools, these tools do not sup-port neither cross language duplicate record detection nor transliterations. Moreover, some of them do not sup-port the Unicode system. Also they are not aware with the structure and semantic of non-English languages names and their characteristics. Therefore, many researchers exerted a lot of efforts to support native languages and transliterations. For example, many tools are developed to support Arabic language (El-Shishtawy, 2013, Yousef, 2013, Higazy et al., 2013). These tools used different algorithms for duplicate record detection. For example, (Higazy et al., 2013) proposed and implemented a tool (DRD) for Arabic Language duplicate record detection. They used a sample for scholars' data saved in Arabic language. They found that the true positive rate (Recall) for the machine has been has been improved substantially (from 66% to 94.7%), when an Arabic Language extension is developed. They proposed and implemented nested blocking based on two stages. They found that number of comparisons is reduced substantially without scarifying the quality of duplicate records combination.

(Yousef, 2013) proposed a SQL wildcards based search as a way for name blocking/indexing. Then iterative relax condition process is used to solve the over-blocking problem when the number of words in records is different. The use of weighted atomic token is adopted to suit Arabic Language. The use of subject matter experts verified dictionaries that are based on compound soundex algorithm is proposed to solve the bilingual duplicate record detection problem in Arabic.

The record duplicate record detection and record linking problems are extended in several ways. The problem of carrying out the detection or linkage computation without full data exchange between the data sources has been called private record linkage and discussed in (Yakout, 2009). The problem to identify persons from evidence is the primary goal of forensic analysis (Srinivasan, 2008). Machine translation in query translation based cross language information access is studied in (Dan Wu, 2012). Speech-to-speech machine translation is another extension that can be achieved using grapheme-to-phoneme conversion (Jiampojamarn, 2010). The problem is extended in another way to match duplicate videos and other multimedia files including image and audio files. This increases the need to have high performance record linking and duplication detection (Kim, 2010).

In cross language information retrieval, it is found that a combination of static translation resources plus transliteration provides a successful solution. As normal bilingual dictionaries cannot be used for person names, these person names should be transliterated because they are considered out of vocabulary (OOV) words. A simple statistical technique to train English to Arabic transliteration model from pairs of names is presented in (AbdulJaleel, 2003a, AbdulJaleel, 2003b). Additional information and relations about the entities being matched could be extracted from the web to enhance the quality of linking.

4 Methodology

In order to detect duplicate records that represent names in different languages with different alphabets, like English and Arabic or Chinese and French, a

framework and solution architecture is proposed. The framework represents an extended solution, designed for cross language duplicate record detection and based on the merge of the software and solution architecture of two previous researches (Yousef, 2013, Higazy et al., 2013). The new proposed framework is named CLDRD and its components will be described in the next subsections and compared to Febrl according to available options. Several experiments are executed to compare CLDRD to Febrl.

The framework can be summarized as follows: a language detection algorithm is used to know the languages found in the dataset. Then, a dictionary for each detected language is built using the data records in the dataset. This dictionary is used as an interface for transliteration and other duplicate detection procedures. In the following subsections, the proposed framework will be presented. The details of the preprocessing stage will be de-scribed. Then, the dictionary building process will be described. The record comparison process will be mentioned and finally the quality metrics evaluation will be presented.

4.1 The Proposed Duplicate Record Detection Framework

The architecture of the proposed system is shown in figure 1 which outlines the general steps involved in the cross language duplicate record detection. The data cleaning and standardization process is used because real-world databases contain always dirty and noisy, incomplete and incorrectly formatted information. The main task of data cleaning and standardization process is the conversion of the raw input data into well defined, consistent forms, as well as the resolution of inconsistencies in the way information is represented.

In this research, a web-based duplicate record detection framework is designed and implemented to overcome the missing features and capabilities in the currently available frameworks. The proposed framework provides black box web based duplicate record detection service with no need for additional configurations or installations on the client side machine. As shown in Figure 1, the overall architecture for the proposed framework, the subject matter expert request the service and start to specify the required information about: data source, language extensions, indexing/blocking options and other parameters for the duplicate record detection process.

The DRD process can be performed with the built in rules that have been built based on training data. The additional value for building this framework as a web-based is to build an accumulative standardization rules based on the human interaction through the web interface, which gives the system the ability to improve its behavior through user's experience. After reviewing the frequently added rules and testing it using training data, the system takes a decision if this rules should be added to the language extension or not.

The framework proposes an enhanced approach to perform DRD over dataset contains English or Arabic data, general cleaning and standardization rules are applied for both languages. In case of Arabic data, a special treatment is performed to cover the wide range of typographical variation in Arabic. This is done by having an Arabic adjustment extension to take care of the special features of Arabic names. The DRD complexity is reduced by performing a new implementation for

indexing/blocking step. The following sub-sections discuss the implemented methodology in details. In Figure 1 the blue blocks/items refers to user interaction process, other blocks represent the system default values/functions and processes.

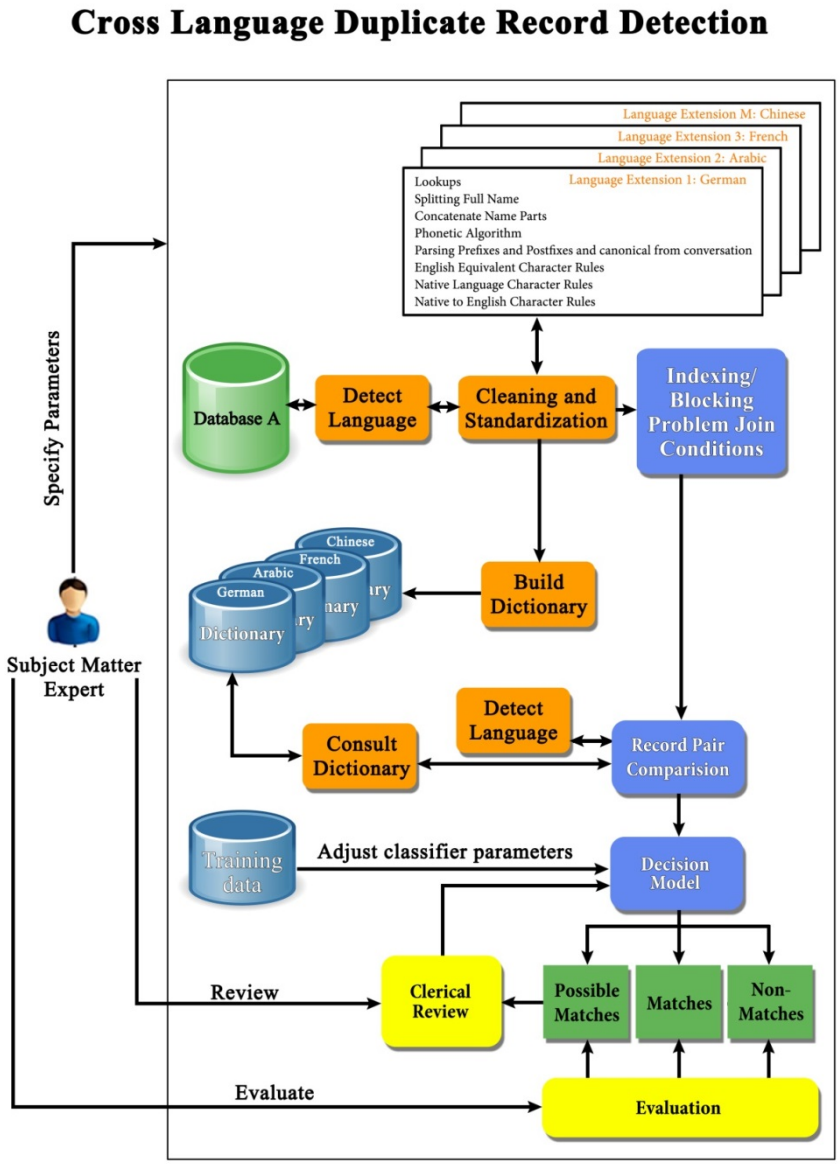

Fig. 1 The proposed framework for Cross Language Duplicate Record Detection

The second step ('Indexing/Blocking') applies the problem domain join conditions to eliminate clear un-matched records and then generates pairs of candidate records. These records are compared in detail using approximate string comparisons and similarity functions, which take (typographical) variations into account and generates the similarity weight vectors. Then, the decision model is used to classify the compared candidate record pairs according to their weight vectors into matches, non-matches, and possible matches.

Clerical review process is used to manually assess the possible matched pairs and classify them into matches or non-matches. Clerical review is the process of human oversight to decide the final duplicate record detection status of possible matched pairs. The person undertaking clerical review usually has access to additional data which enable them to resolve the status or applies human intuition and common sense to take decision based on available data. Measuring and evaluating the quality and the complexity of the duplicate record detection project is the final step that calculates the efficiency and effectiveness of the cross language duplicate record detection machine.

4.2 Pre-processing: Data Cleaning and Standardization

The current used techniques that perform cleaning and standardization do not cover all areas of typographical variations. Therefore, the data cleaning and standardization process is designed to depend on the installed language extensions. During this process, data is unified, normalized and standardized. These steps improve the quality of the in-flow data and make the data comparable and more usable. It simplifies the recognition of names and detecting their language which is an important step to recognize typographic variants. The pre-processing step is done on several levels including character level normalization, splitting and parsing, converting the combined names into canonical format and using lookups.

The cleaning and standardization process include built in character normalization stage that removes separators like additional spaces, hyphens, underscores, commas, dots, slashes and other special characters from the full names. This includes also converting every uppercase letter to a lowercase letter.

4.3 Language Extensions

For non-English languages, standardizing names through character normalization is more difficult and involves several steps. These steps are defined as services from bottom to top, where a top level service can depend on a lower service and call it. For example, the full name splitting service depends on the parsing service that is aware of names prefixes and postfixes. In the next subsections, these services are presented.

4.3.1 Character Standardization Rule

The language extension defines several types of character standardization rules. The first type is the native language to English character rules. Table 1, mentioned earlier, shows a sample of the basic applied rules for German language. The second type of rules is the native language character rules. For example, a rule can be de-fined to instruct the CLDRD machine that the set of Arabic characters (أ ، إ ، آ ، ا) are equivalent to the Arabic character (ا). The same rules can be defined for other languages for Arabic language. The third type of rules is the English Equivalent character rules which can be used in Arabic to make G and J interchangeable. For example, the names "Gamal" and "Jamal" are both equivalent.

Training data can be used to build a basic standardization rules. The typographical variations are recognized by the subject matter expert and he can then define a set of equivalent values for each case. Linguistics and subject matter expert (SME) can set also any additional rules to normalize data. For example: the SME can edit rules to remove titles, support middle names abbreviations and support title prefixes like (prof., Dr.) as a string level standardization.

4.3.2 Name Parsing and Unification (Canonical Form Conversion)

Names with prefixes and postfixes should be parsed and converted to a canonical form. For example, with an ordinary word splitter parser, a full name like “Abdel Rahman Mohamad” or "Marco Van Basten" are split each by the parser to three words and appears as if it consists of three names. The Arabic language extension and the Dutch language extension define canonical form aware name parsing process. This process uses the pre-stored prefixes table to reorganize “Abdel Rahman” as a single first name and "Van Basten" as a composite last name. The last step here is the unifying process which unifies the variants of “Abdel Rahman” including “Abd El Rahman”, “Abdul Rahman”, “Abd Al Rahman” to a single unified canonical form. In the pro-posed framework, the SME has the ability to create a standard form to represent input data that matches some condition such all (AbdAl%) will be replaced by (Abd Al%). A custom rule is defined to replace data strings starting with certain prefix by another one.

4.3.3 Splitting and Reordering

If the data contains name fields in a full name format, the full names are split into separate names representing first name, middle initials and last name. For example, John M. Stewart is converted to three names: John, M., Stewart.

In some applications and languages including English and French, names are represented in a format where last names appear first. In other language including Arabic, first names appear first. Changing the order of the names represented in a language to match the transliterated names representing another language is an important step to align the names.

4.3.4 User Defined Lookup

Some fields need a special treatment to enhance data quality. In a lack of shared lookup tables, some records may refer to the same value which is represented in different ways like (København, Copenhagen). Although these values are identical, it will cause two different blocks to be generated, thus a wrong insertion for candidate records will occur. Subject matter expert can define a lookup to solve this issue. An example of this lookup can solve the globalization effect of having the same city with spelling variations between different languages. The following table shows an example for using a lookup table that maps a group of different values for certain city into a single one.

Table 3 Sample lookup rules for cities

City	Equivalent City
København	Copenhagen
Copenhagen	Copenhagen
CPH	Copenhagen

4.4 Building the Phonetic Based Dictionaries

After cleaning and standardizing the dataset, the language of each record is detected and a dictionary will be built for each non-English alphabet language relating names to its equivalent transliteration. This is done before starting the record comparison process. This dictionary will contain a record for each non-English Character and the corresponding English Equivalents. It will contain also the list of all non-English names and their English transliterated equivalent.

Because names from different language are nearly pronounced with the same pronunciation and share the same phonetic attributes, phonetic algorithms described in previous sections could be used to build dictionaries from the records in the dataset that have the same phonetic code for name fields. This code can be used as a join condition. Soundex technique matches similar phonetic variants of names.

The soundex code for every non-English name and its equivalent English transliterated name record in the dataset is generated according to the algorithm defined in the language extension. For example, this technique is used with the Arabic compound Soundex algorithm to create the same code of Arabic names. For example both the Arabic name "محمد" and its transliteration "Mohamed" will have "M530" as a soundex code.

4.5 Indexing/Blocking

Each duplicate record detection problem is associated with some problem domain indexing/blocking conditions. These conditions identify which record pairs are possible candidates according to their similarity in certain fields. These fields are usually additional fields, other than the name fields. Field matching is used also as

a blocking scheme that minimizes the number of record pairs to be compared later. Indexing/blocking is responsible for reducing the number of generated pairs of records by preventing the comparison of record pairs that will certainly causes a false result. The nature of data determines which technique can be used to perform this task efficiently. Records that are not satisfying the condition will be assumed true negatives and not considered as matches or possible matches.

Febrl uses several types of blocking including full index, blocking index, sorting index, q-gram index, canopy clustering index, string map index, suffix array index, big match index and deduplication index. It is worth mentioning here that both Febrl and CLDRD have options of no blocking, traditional and sorted neighborhood blocking with the same implementation of these types of blocking. The nested blocking feature of the CLDRD is unique and is used to enhance the reduction ratio, computation time and performance without affecting the accuracy.

4.6 Record Pair Comparison

In the CLDRD proposed framework, Jaro-Winkler string matching function is selected and used because it considers the number of matched characters and the number of transportation needed regardless of the length of compared two strings. In the future, many other string matching functions will be implemented in the CLDRD to be comparable with Febrl. Febrl uses many approximate string comparisons. Each filed is compared using the similarity functions and a weight vector is produced for each record pair.

4.7 Classification Function

Each weight vector representing a pair of records will be classified into duplicates, non-duplicate and possible duplicates, based on the calculated similarity values. The proposed framework uses training data to set the upper and lower thresholds. Users can modify this value and set the suitable threshold value according to nature of the data. The classifier used in CLDRD (Higazy et al., 2013, Yousef, 2013) was based on Fellegi and Sunter classifier. The available implemented classifiers in FEBRL include Fellegi and Sunter classifier, K-means clustering, Farthest first clustering, optimal threshold classifier, Support vector machine classifier, Two-step classifier and True match status for supervised classification (Christen, 2008b, Christen, 2008a).

4.8 Quality Evaluation of the Cross Languages Duplicate Record Detection

In the duplicate record detection problem, the self-join between two records (a, b) from the source dataset A will result in a record pair. The machine can identify that the resultant record pair represent an exact match and should be put in the Match set (M). If the record pair is not considered a match, it will be a member of the non-matched set (U). If the record pair cannot be classified as a match or

non-match, it will be a member of the possible match set (P). Then, the clerical reviewer should decide which classification is more realistic. It is clear that the smaller number of records in P the better for the clerical reviewer.

Assume that we have the subject matter experts evaluation results presented in the following table for a four records dataset.

Table 4 Example of Matching Results after being evaluated by the subject matter experts

Record ID	Record ID	Machine Similarity	Machine Classification	SME Evaluation	Reasons
A1	A2	0.9	Accept	Accept	
A1	A3	0.7	Accept	Not Accept	R1
A1	A4	0.5	Accept	Not Accept	R2
A2	A3	0.2	Not Accept	Accept	R3
A2	A4	0.95	Accept	Not Accept	
A3	A4	0.3	Not Accept	Not Accept	

Record Pairs shown in the second, third and fourth rows present differences between the machine result and SME opinion in the classification process. The reasons of these differences should be recorded and converted to rules to improve the results of machine classification. The reasons may be as follows: R1: The machine did not recognize compound names which need special localized phonetic soundex for each language, R2: The machine swapped the first name and last name, which is not accepted in certain languages, R3: The machine did not find dictionary entries for certain names when one of the records is a transliteration of the second. These reasons and comments from SME that explain the differences between the machine results and SME results are converted to rules and added to the language extension.

The data is then classified in the confusion matrix. The metrics of TP, TN, FP and FN are then counted. Other metrics will be computed, including Accuracy, Precision and Recall. It is worth mentioning here is that the mismatch between the machine and subject matter expert is an opportunity for improvement if it is converted to a used defined rule and added to the language extension.

4.9 Future Aspects: Moving towards Big Data

Porting the CLDRD application to Big Data can be done easily using the following steps. Available data should be imported and distributed to many servers using Hadoop and HBase. Use of HBase as a column oriented database built over HDFS (Hadoop distributed file system). HBase is selected because it supported structured data with millions of rows (sparse tables) and random, real time read/write access. HBase is the open source version of Google's BigTable and supports the storage of petabytes of data across thousands of commodity servers. Machine learning algorithms used in classification can be converted to Mahout which is

used as a scalable machine learning and data mining library for Hadoop. Then, the Web based DRD Applications can be converted to the cloud using the software as a service model (SaaS). The system can be reengineered also to support service oriented architecture to allow integration with applications.

5 Results and Discussions

The features of the proposed framework are compared to Febrl. Then, several experiments are designed to test both tools, verify the new proposed framework and compare its results to FEBRL. The following table compares the features of CLDRD and Febrl.

Table 5 Features comparison of CLDRD and Febrl

Features	CLDRD	Febrl
Unicode support	Yes	No
Language detection algorithm	Yes	No
Number of Similarity Functions	3	17+
Number of Classifiers	1	7
Number of Blocking techniques	4	9+
Clerical Reviews Tool	Yes	No
Dictionary building and searching	Yes	No
Metrics Evaluation (TP, Accuracy, Precision, RR)	Yes	No
Lookup	Yes	Yes
Display record pair comparison results with details	Enhanced	Partial
Display classifier inputs, outputs to trace classifier.	Yes	Yes

In CLDRD, the similarities functions are limited to Winkler, Jaro and customized phonetic algorithms for each language, while Febrl support much more similarities functions including Q-gram, Positional Q-gram, Skip-gram, Edit distance, Bag distance, Damerau-Levenshtein, Smith-Waterman, Syllable alignment, Sequence match, Editex approximate, Longest common sub-string, Ontology longest common sequence, Compression based approximate, Token-set approximate string comparison and phonetic algorithm including Soundex.

Regarding classifiers, CLDRD supports only Fellegi and Sunter classifier, while Febrl supports many other classifiers including K-means clustering, Farthest first clustering, optimal threshold classifier, Support vector machine classifier, Two-step classifier and True match status for supervised classification.

The blocking techniques used in CLDRD are limited to traditional, SNH, no blocking and nested, while Febrl supports other indexing/blocking techniques including full index, blocking index, sorting index, q-gram index, canopy clustering index, string map index, suffix array index, big match index and deduplication index.

In order to test the effectiveness of the proposed framework and compare it with FEBRL, a dataset is used and made available on the web in http:// goo.gl/BTYXf9. The author encourages researchers to use it with their own tools for results comparison with Febrl and CLDRD.

This data set is based on the original Febrl dataset named dataset_A_10,000. This original dataset has been generated artificially using the Febrl data set generator (as available in the dsgen directory in the Febrl distribution). It contains names, addresses and other personal information that are based on randomly selected entries from Australian white pages (telephone books). Some fields were randomly generated using lookup tables and predefined formulas. French, German and Arabic records are inserted with their English transliterations to test cross language duplicate record detection. Data cleaning is done then by removing records with missing information and minimizing number of fields to given_name, last_name, age and state. The resultant dataset has 7,709 records.

5.1 Experiment 1: Comparing the CLDRD to FEBRL

The first experiment aims to detect cross language duplicate records from the dataset mentioned above. The dataset contains records including names from English, Arabic, French, German languages.

For each language, a language extension is designed, implemented and installed to the CLDRD system. The language extension contains the definition of the used phonetic algorithm of the language, the character rules, prefixes, postfixes and lookup. Some of these rules are shown in the following table:

Table 6 Sample Rules of Language Extensions for French, German and Arabic Languages

Features	French	German	Arabic
Native to English Equivalence Character Rules	ç = c é = e	ü = u ß =ss ä = e ö = o	ا = A ب = B ,…etc
Native Equivalence Character Rules	k = c	ä = e	أ = ا ة = ه ي = ى =ئ
English Equivalence Character Rules	None	J = Y	'G' = 'J'
Phonetic Algorithms	Henry Code	DaitchMokotoff Soundex	Abdel = Abdul Arabic Combined Soundex
Parsing Prefixes and Postfixes and canonical form conversion	None	None	Abdel, Abdul, Abd El, El, Aboul, Abu El, Abo El, Eldin, El deen
Concatenate name parts (Last name first)	Yes	No	No
Additional Lookups (Example for City = Cairo)	Caire	Kairo	Al Qahirra القاهرة

In this experiment, the classifier used in both CLDRD and FEBRL is K-means classifiers with 'Euclidean' as a distance measure and 'Min/max' centroid initialization.

For the English records of the dataset, the resultant output weight vectors between CLDRD and FEBRL were identical. For the French records, German records and Arabic records, the similarity weight vector results of CLDRD were better than FEBRL. We would like here to highlight the output of four records to show the differences.

It is clear that the CLDRD has higher value of similarity compared to FEBRL. The first reason behind these values is that the duplicate record detection machine has the ability to match English character to its equivalent in a local language. For example, the user can define rules for characters ü, ß in German language with their equivalent characters u, ss in English language, respectively. The second reason is the use of dictionary to compare names from different alphabets, as that found of the last record of the table above which matches an Arabic name to its English transliteration.

5.2 Experiment 2: Comparing Blocking Techniques in FEBRL and CLDRD

In this experiment, both Febrl and CLDRD blocking capabilities are compared. The number of input records is 7,709. If no blocking is used, more than 59,420,972 (about 60 millions) of comparison operations are needed. Using traditional blocking for both Age and state fields, both software has only 292,010 comparison operations. Because both tools have the same implementations of blocking techniques, all results were the same.

Nested blocking is available only in CLDRD. Although (Higazy et al., 2013) claimed less number of operations when nested blocking is used, this is not achieved without the use of additional field as an index. For example, using sorting for the given name field is used as an additional indexing in nested blocking. When CLDRD is used with nested blocking, the number of comparisons decreased to 96,012 comparison operation, with reduction about 67% of traditional blocking. More reduction can be achieved with the use of indexing on the last name as well.

The number of true positives for both blocking types was comparable. 3,278 true positives in traditional blocking and 3,267 in nested blocking. This highlights the possibility of adding more fields to indexing/blocking to reduce number of comparison operations. When the subject matter expert define the use of the indexing/blocking as a problem domain join conditions, the block to be searched is narrowed and the performance is enhanced without losing a lot of true positives. This means that most eliminated candidate pairs were certainly true negatives.

Two communication points here worth mentioning. The first communication with CLDRD team indicated the use of given_name as an additional index in nested blocking, which makes the comparison with traditional blocking somehow biased. Communication with the Febrl team indicated that Febrl does not support currently the nested blocking (having traditional blocking for one field with sorted blocking for another field). However, the open source characteristics of Febrl allow free modification by extending the code. Also, the user could define various indexes in different project files, get several sets of weight vectors and then combine them.

6 Conclusion

In this chapter, two frameworks of duplicate record detection (CLDRD and Febrl) are compared from their capabilities of supporting cross language duplicate record detection. It is found that CLDRD efficiently supports true cross language duplicate record detection and nested blocking features that affects the time and performance of duplicate record detection that suits big data. Febrl has many advanced options in similarity functions and classifiers with little support to cross language duplicate record detection and moderate blocking options. Based on CLDRD, A generic framework for cross languages duplicate record detection was proposed and implemented partially. The proposed framework used language extensions and language specific phonetic algorithms to building baseline dictionaries from existing data and use them in record comparisons.

References

Abduljaleel, N.L., Leah, S.: English to Arabic Transliteration for Information Retrieval: A Statistical Approach (2003a)

Abduljaleel, N.L., Leah, S.: Statistical transliteration for English-Arabic cross language information retrieval. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM, pp. 139–146 (2003b)

AL-Onaizan, Y., Knight, K.: Machine Transliteration of Names in Arabic Text. In: ACL Workshop on Comp. Approaches to Semitic Languages (2002)

Monge, A., Elkan, C.: The field matching problem: Algorithms and applications. In: Second International Conference on Knowledge Discovery and Data Mining (1996)

Amor-Tijani, G.: Enhanced english-arabic cross-language information retrieval. George Washington University (2008)

Aqeel, S., Beitzel, S., Jensen, E., Grossman, D., Frieder, O.: On the Development of Name Search Techniques for Arabic. Journal of the American Society of Information Science and Technology 57(6) (2006)

Boussy, C.A.: A comparison of hand and computer-linked records. University of Miami (1992)

Christen, P.: A Comparison of Personal Name Matching: Techniques and Practical Issues. In: Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops 2006, pp. 290–294 (December 2006)

Christen, P.: Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2008a)

Christen, P.: Automatic Training Example Selection for Scalable Unsupervised Record Linkage. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 511–518. Springer, Heidelberg (2008b)

Christen, P.: Febrl: a freely available record linkage system with a graphical user interface. In: Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management, vol. 80. Australian Computer Society, Inc., Wollongong (2008c)

Christen, P.: Development and user experiences of an open source data cleaning, deduplication and record linkage system. SIGKDD Explor. Newsl. 11, 39–48 (2009)

Christen, P., Churches, T., Hegland, M.: Febrl – A Parallel Open Source Data Linkage System. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 638–647. Springer, Heidelberg (2004)

Christen, P., Goiser, K.: Quality and Complexity Measures for Data Linkage and Deduplication. In: Guillet, F., Hamilton, H. (eds.) Quality Measures in Data Mining. SCI, vol. 43, pp. 127–151. Springer, Heidelberg (2007)

Dan Wu, D.H.: Exploring the further integration of machine translation in English-Chinese cross language information access. Program: Electronic Library and Information Systems 46(4), 429–457 (2012)

Dey, D., Mookerjee, V.S., Dengpan, L.: Efficient Techniques for Online Record Linkage. IEEE Transactions on Knowledge and Data Engineering 23(3), 373–387 (2011)

El-Hadidi, M., Anis, H., El-Akabawi, S., Fahmy, A., Salem, M., Tantawy, A., El-Rafie, A., Saleh, M., El-Ahmady, T., Abdel-Moniem, I., Hassan, A., Saad, A., Fahim, H., Gharieb, T., Sharawy, M., Abdel-Fattah, K., Salem, M.A.: Quantifying the ICT Needs of Academic Institutes Using the Service Category-Stakeholder Matrix Approach. In: ITI 6th International Conference on Information & Communications Technology, ICICT 2008, pp. 107–113. IEEE (2008)

El-Shishtawy, T.: A Hybrid Algorithm for Matching Arabic Names. arXiv preprint arXiv:1309.5657 (2013)

Elfeky, M.G., Verykios, V.S., Elmagarmid, A.K.: TAILOR: a record linkage toolbox. In: Proceedings of the 18th International Conference on Data Engineering, vol. 2002, pp. 17–28 (2002)

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)

Elyamany, H.F., Yousef, A.H.: A Mobile-Quiz Application in Egypt. In: The 4th IEEE International E Learning Conference, Bahrain, May 7-9 (2013a)

Fellegi, I.P., Sunter, A.B.: A Theory for Record Linkage. Journal of the American Statistical Association 64, 1183–1210 (1969)

Freeman, A.T., Condon, S.L., Ackerman, C.M.: Cross linguistic name matching in English and Arabic: a “one to many mapping” extension of the Levenshtein edit distance algorithm. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York (2006)

Goiser, K., Christen, P.: Towards automated record linkage. In: Proceedings of the Fifth Australasian Conference on Data Mining and Analytics, vol. 61. Australian Computer Society, Inc., Australia (2006)

Herzog, T.N., Scheuren, F.J., Winkler, W.E., Herzog, T., Scheuren, F., Winkler, W.: Record Linkage – Methodology. Springer, New York (2007)

Higazy, A., El Tobely, T., Yousef, A.H., Sarhan, A.: Web-based Arabic/English duplicate record detection with nested blocking technique. In: 2013 8th International Conference on Computer Engineering & Systems (ICCES), November 26-28, pp. 313–318 (2013)

Hussein, A.S., Mohammed, A.H., El-Tobeily, T.E., Sheirah, M.A.: e-Learning in the Egyptian Public Universities:Overview and Future Prospective. In: ICT-Learn 2009 Conference, Human and Technology Development Foundation (2009)

Jiampojamarn, S.: Grapheme-to-phoneme conversion and its application to transliteration. Doctor of Philosophy, University of Alberta (2010)

Kim, H.-S.: High Performance Record Linking. Doctor of Philosophy, The Pennsylvania State University (2010)

Knight, K.G., Jonathan: Machine Transliteration. Computational Linguistics (1997)

Köpcke, H., Rahm, E.: Frameworks for entity matching: A comparison. Data & Knowledge Engineering 69(2), 197–210 (2010)

Koujan, T.: Arabic Soundex (2008), `http://www.codeproject.com/Articles/26880/Arabic-Soundex`

Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707–710 (1966)

Ma, X., Pennsylvania, U.O.: Improving Named Entity Recognition with Co-training and Unlabeled Bilingual Data, University of Pennsylvania (2008)

Mohamed, K.A., Hassan, A.: Web usage mining analysis of federated search tools for Egyptian scholars. Program: electronic library and information systems 42(4), 418–435 (2008)

Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys (CSUR) 33(1), 31–88 (2001)

Mcnamee, P., Mayfield, J., Lawrie, D., Oard, D., Doermann, D.: Cross Language Entity Linking. In: IJCNLP: International Joint Conference on Natural Language Processing (2011)

Russell, R.C.: Russell Index U.S. Patent 1,261,167 (1918), `http://patft.uspto.gov/netahtml/srchnum.htm`

Russell, R.C.: Russell Index U.S. Patent 1,435,663 (1922), `http://patft.uspto.gov/netahtml/srchnum.htm`

Shaalan, K., Raza, H.: Person name entity recognition for Arabic. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Association for Computational Linguistics, Prague (2007)

Srinivasan, H.: Machine learning for person identification with applications in forensic document analysis. Doctor of Philosophy (Ph.D.), State University of New York at Buffalo (2008)

Weifeng, S., Jiying, W., Lochovsky, F.H.: Record Matching over Query Results from Multiple Web Databases. IEEE Transactions on Knowledge and Data Engineering 22(4), 578–589 (2010)

Winkler, W.E.: Overview of record linkage and current research directions. Bureau of the Census (2006)

Yakout, M.A., Mikhail, J.: Elmagarmid, AHMED. 2009. Efficient private record linkage. In: IEEE 25th International Conference on Data Engineering, ICDE 2009, pp. 1283–1286. IEEE (2009)

Yancey, W.E.: Bigmatch: A Program for Extracting Probable Matches from a Large File for Record Linkage. Statistical Research Report Series RRC2002/01. US Bureau of the Census, Washington, D.C. (2002)

Yousef, A.H.: Cross-Language Personal Name Mapping. International Journal of Computational Linguistics Research 4(4), 172–192 (2013)

Yousef, A.H., Tantawy, R.Y., Farouk, Z., Mohamed, S.: Using Professional Social Networking as an Innovative Method for Data Extraction, The ICT Alumni Index Case Study. In: 1st International Conference on Innovation & Entrepreneurship. Technology Innovation and Entrepreneurship Center, Smart Village (2012)

A Novel Hybridized Rough Set and Improved Harmony Search Based Feature Selection for Protein Sequence Classification

M. Bagyamathi and H. Hannah Inbarani

Abstract. The progress in bio-informatics and biotechnology area has generated a big amount of sequence data that requires a detailed analysis. Recent advances in future generation sequencing technologies have resulted in a tremendous raise in the rate of that protein sequence data are being obtained. Big Data analysis is a clear bottleneck in many applications, especially in the field of bio-informatics, because of the complexity of the data that needs to be analyzed. Protein sequence analysis is a significant problem in functional genomics. Proteins play an essential role in organisms as they perform many important tasks in their cells. In general, protein sequences are exhibited by feature vectors. A major problem of protein dataset is the complexity of its analysis due to their enormous number of features. Feature selection techniques are capable of dealing with this high dimensional space of features. In this chapter, the new feature selection algorithm that combines the Improved Harmony Search algorithm with Rough Set theory for Protein sequences is proposed to successfully tackle the big data problems. An Improved harmony search (IHS) algorithm is a comparatively new population based meta-heuristic optimization algorithm. This approach imitates the music improvisation process, where each musician improvises their instrument's pitch by seeking for a perfect state of harmony and it overcomes the limitations of traditional harmony search (HS) algorithm. An Improved Harmony Search hybridized with Rough Set Quick Reduct for faster and better search capabilities. The feature vectors are extracted from protein sequence database, based on amino

M. Bagyamathi
Department of Computer Science, Gonzaga College of Arts and Science for Women, Krishnagiri, Tamil Nadu, India
e-mail: bagyaarul@gmail.com

H. Hannah Inbarani
Department of Computer Science, Periyar University, Salem, Tamil Nadu, India
e-mail: hhinba@gmail.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_6

acid composition and K-mer patterns or K-tuples and then feature selection is carried out from the extracted feature vectors. The proposed algorithm is compared with the two prominent algorithms, Rough Set Quick Reduct and Rough Set based PSO Quick Reduct. The experiments are carried out on protein primary single sequence data sets that are derived from PDB on SCOP classification, based on the structural class predictions such as all α, all β, all α+β and all α / β. The feature subset of the protein sequences predicted by both existing and proposed algorithms are analyzed with the decision tree classification algorithms.

Keywords: Data Mining, Big Data Analysis, Bioinformatics, Feature Selection, Protein Sequence, Rough Set, Particle Swarm Optimization, Harmony Search, Protein sequence classification.

1 Introduction

Big data is a term that describes the exponential growth and availability of data, both structured and unstructured form. Many factors contribute to the increase in data volume. In under a year, proteomics and genomics technologies will enable individual laboratories to generate terabyte or even petabyte scales of data at a reasonable cost. However, the computational infrastructure that is required to maintain and process these large-scale data sets, and relate these data patterns to other biologically relevant information A primary goal for biological researchers is to construct predictive models for the big data set that can be computationally demanding (Schadt et al., 2010).

Due to the growth of molecular biology technologies and techniques, more large-scale biological data sets are becoming available. Identifying biologically-useful information from these data sets has become a significant challenge. Computational biology aims to address this challenge (Wei. 2010). Many problems in computational biology, e.g., protein function prediction, sub cellular localization prediction, protein-protein interaction, protein secondary structure prediction, etc., can be formulated as sequence classification tasks (Wong and Shatkay. 2013), where the amino acid sequence of a protein is used to classify the protein in functional and localization classes.

Proteins play a fundamental role in all living organisms and are involved in a variety of molecular functions and biological processes. Proteins are the most essential and versatile macromolecules of life, and the knowledge of their functions is an essential link in the development of new drugs, better crops, and even the development of synthetic biochemical such as biofuels (Nemati et al., 2009). Traditionally, computational prediction methods use features that are derived from protein sequence, protein structure or protein interaction networks predict function (Rost et al. 2003; Rentzsch and Orengo. 2009).

In the past few decades, major advances in the field of molecular biology, tied to advances in genomic and proteomic technologies, have led to a fiery increase in the biological information generated by the scientific community. Although the number of proteins with known sequence has grown exponentially in the last few

years, due to rapid advances in genome sequencing technology, the number of proteins with known structure and function has grown at a substantially lower rate (Freitas and de Carvalho. 2007). Thus, characterizing the function of proteins is an important goal in proteomics research (Wong and Shatkay. 2013).

Proteins are composed of one or more chains of amino acids and show several levels of structure. The primary structure is defined by the sequence of amino acids, while the secondary structure is defined by local, repetitive spatial arrangements such as helix, strand, and coil. The 3D structure of proteins is uniquely determined by their amino acid sequences. The tertiary structure is defined by how the chain folds into a three dimensional configuration. The assumption is that the primary structure of a protein codes for all higher level structures and associated functions. In fact, according to their chain folding pattern, proteins are usually folded into four structural classes such as all α, all β, all α + β and all α / β (Cao et al. 2006). In this chapter, the features are extracted from protein primary sequence, based on amino acid composition and K-mer patterns, or K-grams or K-tuples (Chandran. 2008).

Protein sequence data contain inherent dependencies between their constituent elements. Given a protein sequence $x = x_0,..., x_{n-1}$ over the amino acid alphabet, the dependencies between neighboring elements can be modeled by generating all the adjoining sub-sequences of a certain length K, x_i-K, ..., x_i-1, $i = K,..., n$, called K-grams, or sequence motifs. Because the protein sequence motifs may have variable lengths, generating the K-grams can be done by sliding a window of length K over the sequence x, for various values of K. Exploiting dependencies in the data increases the richness of the representation. However, the fixed or variable length K-gram representations, used for protein sequence classification, usually result in prohibitively high-dimensional input spaces, for large values of K (Caragea et al. 2011).

Models such as Principal Component Analysis, Latent Dirichlet Allocation and Probabilistic Latent Semantic Analysis are extensively used to perform dimensionality reduction. Unfortunately, for very high dimensional data, with hundreds of thousands of dimensions, processing data instances into feature vectors at runtime, using these models, is computationally expensive, due to inference at runtime in most of the cases. A less expensive approach to dimensionality reduction is feature selection, which reduces the number of features by selecting a subset of the available features based on some chosen criteria (Guyon and Elisseeff. 2003; Fleuret. 2004). In particular, feature selection by average mutual information selects the top features that have the highest average mutual information with the class.

Feature Selection (FS) is an important part of knowledge discovery. FS is used to improve the classification accuracy and reduce the computational time of classification algorithms (Chandrasekhar et al. 2012). FS is divided into the supervised and unsupervised categories. When class labels of the data are known, supervised feature selection can be applied, otherwise the unsupervised feature selection is appropriate. The Supervised Feature Selection methods assess a range of feature subsets using an evaluation function or metric to choose only those

features which relate to the decision classes of the data under consideration (Mitra et al. 2002; Jothi and Inbarani. 2012). In terms of feature selection methods, they fall into the filter and wrapper categories. In filter model, features are evaluated based on the general characteristics of the data without relying on any mining algorithms. On the contrary, wrapper model requires one mining algorithm and utilizes its performance to determine the goodness of feature sets (Fu et al. 2006). The selection of relevant features is important in both the cases. Hence a rough set based feature selection method is applied to the proposed work.

Rough set theory (Pawlak. 1993), provides a mathematical tool that can be used for both feature selection and knowledge discovery. It helps us to find out the minimal attribute sets called 'reducts' to classify objects without deterioration of classification quality. The idea of reducts has encouraged many researchers in studying the effectiveness of rough set theory in a number of real world domains, including medicine, pharmacology, control systems, fault-diagnosis, text categorization (Pawlak. 2002). The proposed work is based on rough set based operations for feature reduction and then it is compared with an existing rough set based feature selection algorithms. The major purpose of the proposed algorithm is to increase the effectiveness of feature selection methods by the optimal reduct.

Optimization is the process of selecting the most excellent element from a set of available alternatives under certain conditions. This process can be solved by minimizing or maximizing the objective or cost function of the problem. In each iteration of the optimization process, selecting the values from within an acceptable set is done systematically until the minimum or maximum result is reached or when the stopping condition is met. Optimization techniques are used on a daily basis for industrial planning, resource allocation, econometric problems, scheduling, decision making, engineering, computer science applications. Research in the optimization field is very active and new optimization methods are being developed regularly (Alia and Mandava. 2011). One of the novel and powerful meta-heuristic algorithm that has been successfully utilized in a wide range of optimization problems is Harmony Search (HS) algorithm.

Geem et al., developed a New Harmony search (HS) meta-heuristic algorithm that was conceptualized using a musical process of searching for a perfect state of harmony (Geem et al. 2001). This harmony in music is analogous to find the optimality in an optimization process. In music improvisation process, the musician plays different notes of different musical instrument and find the best combination of frequency for best tune. Similarly, in HS method also the best combination of available solutions is selected and the objective function is optimized. The HS method had been successfully applied to a diverse range of problems – structural analysis, mechanical component design, water distribution network, medical imaging, games and many others. HS algorithm has many advantages over other meta-heuristic algorithms (Seok and Geem. 2005; Kattan et al. 2010): (a) HS algorithm imposes fewer mathematical requirements and does not require initial value settings of the decision variables. (b) As the HS algorithm uses stochastic random searches, derivative information is also unnecessary. (c) The HS algorithm generates a new vector, after considering all of the existing

vectors. These features increase the flexibility of the HS algorithm and produce better solutions. On the basis of HSA, (Mahdavi et al. 2007) developed a new algorithm called Improved Harmony Search Algorithm (IHSA). In this algorithm, few drawbacks of the HS method have been removed by modifying the algorithm.

The rest of the chapter is structured as Sections 2 to 7. Section 2 describes a review of various feature selection algorithms and Protein Sequences. Section 3 explains the proposed framework of this study. Section 4 describes about the basic principles of Rough Set Theory. Section 5 explains about the feature extraction method from protein sequences. Section 6 describes the existing and proposed feature selection algorithms using rough set Quick Reduct. The experimental analysis with the results and discussion were described in Section 7 and the chapter concludes with a discussion on the interpretation and highlights the possibility of future work in this area.

2 Related Work

During the last decade, application of feature selection techniques in bio-informatics has become a real prerequisite for model building. In particular, the high dimensional nature of many modeling tasks in bio-informatics, going from sequence analysis over microarray analysis to spectral analysis and literature mining has given rise to a wealth of feature selection techniques being presented in the field (Blum and Dorigo. 2004).

In this review, the application of feature selection techniques is focussed, which do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of variables, hence offering the advantage of interpretability by a domain expert (Saeys et al. 2007). While the feature selection can be applied to both supervised and unsupervised learning, this chapter concentrates on the problem of supervised learning (classification), where the class labels are known in advance. In recent decade, Population-based algorithms have been attracting an increased attention due to their powerful search capabilities.

For the particular problem of feature selection, population-based methods aim to produce better or fitter future generations that contain more informative subsets of features (Al-Ani and Khushaba. 2012). Famous population-based FS approaches are based on the genetic algorithm (GA) (Siedlecki and Sklanskyl.1989), simulated annealing (SA), particle swarm optimization (PSO) (Wang et al. 2007) and ant colony optimization (ACO) (Aghdam et al. 2008; Basiri et al. 2008; Nemati et al. 2008). Feature selection, as a preprocessing step to bio-informatics data, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving the result comprehensibility (Xie et al. 2010). There are numerous studies in Feature selection as reported in Table 1 focusing on the bio-informatics dataset using various feature selection methods, hybridized feature selection methods and classification techniques.

Table 1 Related work for this study

Authors	Purpose	Technique
Wang et al. (2007)	Feature selection	This work applies Genetic Algorithm, Particle Swarm Optimization (PSO), PSO and Rough Set-based Feature Selection (PSORSFS) for the feature selection process.
Chandran. (2008)	Feature Extraction and Feature Selection	In this study, the Enhanced Quick Reduct Feature Selection (EQRFS) algorithm using Fuzzy-Rough set is proposed for protein primary sequences.
Nemati et al. (2009)	Feature selection and classification	This study integrates ACO-GA feature selection algorithms with classification techniques for protein function prediction.
Peng et al. (2010)	Feature selection and classification	This study integrates filter and wrapper methods into a sequential search procedure for feature selection for biomedical data using SVM classification.
Xie et al. (2010)	Feature selection and classification	In this study, IFSFFS (Improved F-score and Sequential Forward Floating Search) for feature selection is proposed.
Gu et al. (2010)	Feature selection and classification	In this study, amino acid pair compositions with different spaces are used to construct feature sets. The binary Particle Swarm Optimization and Ensemble Classifier are applied to extract a feature subset.
Pedergnana et al. (2012)	Feature Extraction, Feature Selection and Classification	In this paper, a novel supervised feature selection technique, which is based on genetic algorithms (GAs). GAs are used along with the Random Forest (RF) classifier for finding the most relevant subset of the input features for classification.
Inbarani and Banu. (2012)	Feature selection and clustering	In this work, the following FS methods are applied to gene expression data set. Unsupervised Quick Reduct (USQR) Unsupervised Relative Reduct (USRR) Unsupervised PSO based Relative Reduct (USPSO-RR).

Table 1 (*continued*)

Authors	Purpose	Technique
Inbarani et al. (2012)	**Feature selection and clustering**	**In this study, the following FS methods are applied for gene expression dataset using Unsupervised Quick Reduct (USQR) Unsupervised Relative Reduct (USRR) Unsupervised PSO based Quick Reduct (USPSO-QR).**
Hor et al. (2012)	Feature selection and classification	In this paper, a modified sequential backward feature selection method is adopted and build SVM models with various subsets of features.
Inbarani et al. (2013)	Feature Selection	In this study, a novel supervised feature selection using the Tolerance Rough Set - PSO based Quick Reduct (STRSPSO-QR) and Tolerance Rough Set - PSO based Relative Reduct (STRSPSO-RR), is proposed.
Azar et al. (2013)	Feature Selection	This paper aims to identify the important features to assess the fetal heart rate and are selected by using Unsupervised Particle Swarm Optimization (PSO) based Relative Reduct .
Azar and Hassanien. (2014)	Feature selection and classification	In this paper, linguistic hedges Neuro-fuzzy classifier with selected features (LHNFCSF) is presented for dimensionality reduction, feature selection and classification.
Azar. (2014)	Feature selection	In this paper, Neuro-fuzzy feature selection approach based on linguistic hedges is applied for medical diagnosis.
Inbarani et al. (2014a)	Feature selection and classification	In this study, supervised methods such as a rough set based PSO Quick Reduct and PSO Relative Reduct were applied for medical data diagnosis.
Inbarani et al. (2014b)	Feature selection and classification	In this work, the features are selected by using unsupervised particle swarm optimization PSO - based relative reduct (US-PSO-RR) technique for fetal rate.

Table 1 (*continued*)

Authors	Purpose	Technique
Lin et al (2010)	Feature Selection	In this work, forward selection (addition) with the complete set of PseAAC is proposed to find a good small feature set and classified using SVM.
Alia and Mandava. (2011)	Optimization	In this work, a population based Harmony Search Algorithm is proposed for Optimization techniques.
Mahdavi et al. (2007)	Optimization	In this study, an Improved Harmony Search Algorithm is proposed for Optimization problems.

3 The Proposed Framework

There are several strategies available for classifying the protein sequences. The proposed model shown in Fig. 1 predicts the optimal number of features that improves the classification performance. In this study, the protein primary sequences are collected from Protein Data Bank (PDB) in fasta format (Cao et al. 2006). The fasta sequence file is used as input data to the PseAAC-builder, a Web server, that generates the protein feature space using amino acid composition and amino acid K- tuples or K- mer patterns (Du et al. 2012).

The generated features are shown in Table 2. The data in table 2 are real valued, but the rough set theory best in dealing with discrete values. Hence the real valued data are to be discretized. The discretized values are the actual extracted feature set of this study. In this study, the rough set based feature selection algorithms such as Rough Set Quick Reduct, Particle Swarm Optimization and Improved Harmony Search algorithms were used to select the feature subsets. In the last step, the feature subset predicted by the various feature selection algorithms are evaluated with classification techniques using the WEKA tool (Hall et al. 2009). The most important elements that construct the proposed framework is discussed in the following subsections.

3.1 Protein Primary Sequence

Protein sequences are consecutive amino acid residues, and it is represented with an alphabet A of size |A| = 20. Many feature extraction methods have been developed in the past several years. Typically, these methods can be classified into two categories. One is based solely on amino acid composition. The other one is an extension of the atomic length from only one amino acid to a K amino acid tuple, where K is an integer and larger than one, which can also be referred as 'K-tuple', such as 2-tuple in (Park and Kanehisa. 2003). In the first step of feature

generation, 20 amino acid features are adopted as our initial representative features, which is simple and effective. Along with the 20 features, K-tuple features are introduced. The high computational cost of K-tuple prediction caused by large feature amount, the Rough Set based feature selection methods are introduced to reveal most relevant K-tuples and eliminating the irrelevant ones. In this study, the supervised rough set based feature selection algorithms such as Quick Reduct, PSO Quick Reduct, and Improved Harmony Search Quick Reduct were applied and the protein sequence feature vectors are classified to their respective structural classes.

3.2 PseAAC-builder

Pseudo amino acid composition (PseAAC) is an algorithm that could convert a protein sequence into a digital vector that could be processed by data mining algorithms. The design of PseAAC incorporated the sequence order information to improve the conventional amino acid compositions. The application of pseudo amino acid composition is very common, including almost every branch of computational proteomics (Du et al. 2012).

3.3 Amino Acid Composition

In amino acid composition prediction model, each protein sequence *i* in the dataset of size *N* is represented by an input vector $\vec{x}_i$ of 20 dimensions and a location label y_i, for i = 1, . . . ,N. The prediction procedure can be understood within a 20 dimensional space and each protein sequence represents a point in it. Then these points must be classified to their corresponding labels.

Naturally, amino acid composition (AAC-I in short) to be considered as amino acid residue occurrence times.

$$x_{ij} = count_i(j), \text{ for } i = 1, \ldots, N \text{ and } j = 1, \ldots, 20 \tag{19}$$

where x_{ij} is the j^{th} element of $\vec{x}_i$, and $count_i(j)$ denotes the number of times that amino acid j occurs in protein sequence i.

For normalization purpose, the following equation is always satisfied for any protein sequence whether it is longer or shorter than others.

$$\sum_{j=1}^{20} x_{ij} = 1 \tag{20}$$

So the amino acid composition to be considered as amino acid residue occurrence probability (AAC-II in short)

$$x_{ij} = \frac{count_i(j)}{\sum_{j=1}^{20} count_i(j)} \tag{21}$$

It is reported that better performance could be obtained by normalizing each $|\overrightarrow{x}_i|$ to $|\overrightarrow{a}_i|$ (Nemati et al., 2009), where $|\overrightarrow{a}_i| = 1$ for i = 1, · · ·,N. So each $\overrightarrow{a}_i$, will be the unit length vector in 20 dimensional Euclidean space. The following relation (AAC III in short) between $\overrightarrow{x}_i$ and $\overrightarrow{a}_i$ can be easily proven.

$$a_{ij} = \sqrt{x_{ij}}, \text{ for i = 1, . . . ,N and j = 1, . . . , 20} \quad (22)$$

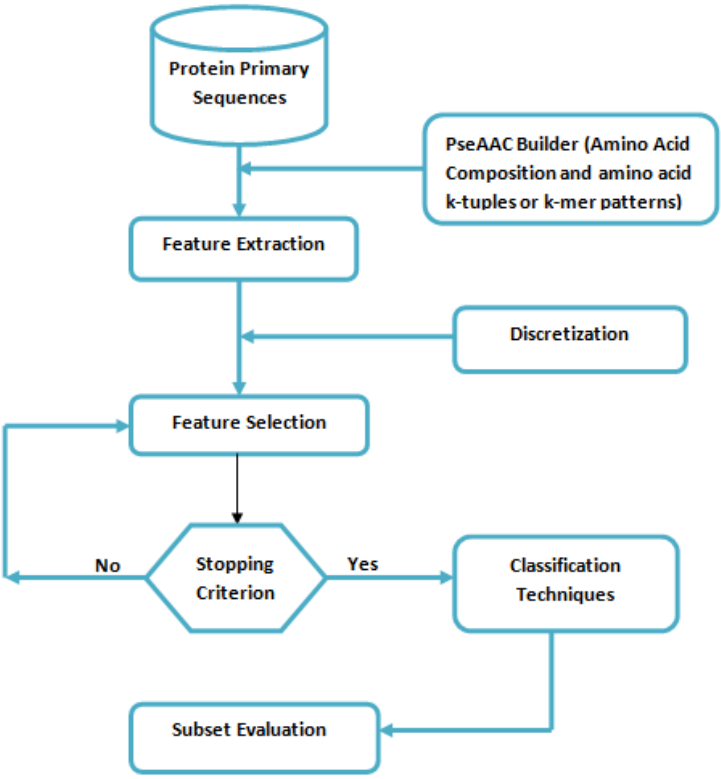

Fig. 1 The Proposed Framework

3.4 K- Tuple Subsequence

It should be pointed out that all of the prediction algorithms based on amino acid composition do not take the sequence order effect into account. To improve the prediction accuracy, it is necessary to incorporate some other information. Intuitively, the amino acid tuple is used to partially represent the sequence order. For example, the sequences "AIC" and "CIA" have the same representation of the 20 amino acid features. But if 2-tuple features are used, "AIC" is represented by "AI" and "IC", and "CIA" is represented by "CI" and "IA".

In this study, ACC-II defined in Eq. (21) is adopted and accordingly modify it to be the K-tuple feature vector for each protein sequence. It should be noted that the dimensionality of K-tuple space increases exponentially with K. So if K is assigned to an arbitrary number, such as 10 or larger, the dimensionality of feature space will be $2010 \approx 1013$. It is too large a feature space for learning.

Many researchers used one method of feature extraction among the following two different strategies. (a) Without dimension reduction: the prediction can be based on full K-tuple space without any dimension reduction, and the maximum value of K is set to 5. As a result, at most $20^5 = 3.2\times10^6$ features are extracted. (b) With the dimension reduction: when the proteins are represented in a high dimensional space, the occurrences of many K-tuples will be very scarce. Some K-tuples just occur only once or even never occur in the dataset. Thus, lots of them must be irrelevant to the protein classification process since they are too sparse.

Motivated by this phenomenon, in this chapter, the feature selection techniques are adopted to filter the K-tuple feature set. Therefore, only the relevant tuples are selected as best subset in order to reveal the better classification result.

Table 2 Discretized features of 2-tuple protein sequences

Objects	A	C	D	.	.	.	Y	AA	AC	AD	.	.	.	YY	Class
1	5	1	6	.	.	.	3	5	8	10	.	.	.	8	1
2	6	2	6	.	.	.	2	5	7	9	.	.	.	8	1
3	13	0	6	.	.	.	0	5	9	12	.	.	.	9	2
4	9	1	10	.	.	.	1	10	3	14	.	.	.	11	2
5	13	0	6	.	.	.	0	5	9	12	.	.	.	9	3
6	9	1	10	.	.	.	1	10	3	15	.	.	.	10	3
7	13	0	6	.	.	.	0	5	9	12	.	.	.	9	4
8	9	1	10	.	.	.	1	10	3	14	.	.	.	12	4

3.5 Discretization

Many Machine Learning algorithms are known to produce better models by discretizing continuous attributes (Kotsiantis and Kanellopoulos. 2006). The Rough Set theory and the classification techniques such as Naive Bayes requires the estimation of probabilities and the continuous attributes are not easy to handle, as they often take too many different values for a direct estimation of frequencies (Ferrandiz and Boullé. 2005). As a result, a large number of protein sequence features can only be applied to datasets composed entirely of nominal variables.

However, a very large proportion of real data sets include continuous variables: that is variables at the interval or ratio level. One solution to this problem is to partition numeric variables into a number of sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is usually termed discretization.

3.6 Protein Classification

Proteins generally form a compact and complex three-dimensional structure. The sequence of amino acids that comprise a protein is called its primary structure. Out of the approximately 30,000 proteins found in humans, only a few have been adequately described.

Many of them exhibit large similarities, both in structure and function, and are naturally viewed as members of the same group. There are many approaches to the classification of proteins. In this study, the most common structural classification systems in computational molecular biology are Structural Classification of Proteins (SCOP). SCOP is a hierarchical structure-based classification of all proteins in PDB (http://www.rcsb.org).

4 Basics of Rough Set Theory

Rough Set Theory (RST) has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information (Mitra et al. 2002). Over the past ten years, RST has become a topic of great interest to researchers and has been applied to many domains. Rough Set based Attribute Reduction (RSAR) provides a filter based tool by which knowledge may be extracted from a domain in a concise way; retaining the information content whilst reducing the amount of knowledge involved (Chouchoulas and Shen. 2001). Central to RSAR is the concept of indiscernibility. The basic concepts of rough set are introduced in the discussion in the rest of the chapter.

Let $I= (U,AU\{d\})$ be an information system, where U is the universe with a non-empty set of finite objects, A is a non-empty finite set of conditional attributes, and d is the decision attribute (decision table),$\forall a \in A$, there is a corresponding function f_a: $U \rightarrow V_a$, where V_a is the set of values of a (Velayutham and Thangavel. 2011). If $P \subseteq A$, there is an associated equivalence relation:

$$IND(P) = \{(x,y) \in UXU \mid \forall a \in P, \ \ f_a(x) = f_a(y)\} \tag{1}$$

The partition of U generated by $IND(P)$ is denoted U/P. If $(x,y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted as $[x]_P$. Let $X \subseteq U$,the P-lower approximation $\underline{P}X$ and P-upper approximation $\bar{P}X$ of set X can be defined as:

$$\underline{P}X = \{x \in U \mid [x]_P \subseteq X\} \quad (2)$$

$$\overline{P}X = \{x \in U \mid [x]_P \cap X \neq \varphi\} \quad (3)$$

Let $P, Q \subseteq A$be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{x \in U|Q} \underline{P}X \quad (4)$$

$$NEG_P(Q) = U - \bigcup_{x \in U|Q} \overline{P}X \quad (5)$$

$$BND_P(Q) = \bigcup_{x \in U|Q} \overline{P}X - \bigcup_{x \in U|Q} \overline{P}X \quad (6)$$

The positive region of the partition U/Q with respect to P, $\mathrm{POS_P(Q)}$, is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P. Q depends on P in a degree k $(0 \leq k \leq 1)$ denoted by $P \Rightarrow_{\mathrm{k}} \mathrm{Q}$

$$K = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \quad (7)$$

Where P is a set of all conditional attributes, Q is the decision attributes, and $\gamma_{\mathrm{P}}(\mathrm{Q})$ is the quality of classification. If k=1, Q depends totally on P; if $0 < k < 1$, Q depends partially on P; and if k=0 then Q does not depend on P. The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original (Pawlak. 1993). The set of all reducts is defined as:

$$Red(C) = \{R \subseteq C \mid \gamma_R(D) = \gamma_C(D), \forall B \subset R, \gamma_B(D) \neq \gamma_C(D)\} \quad (8)$$

A dataset may have many attribute reducts. The set of all optimal reducts is:

$$Red(C)_{min} = \left\{ R \in Red \mid \forall R' \in Red, |R| \leq \left|R'\right| \right\} \quad (9)$$

5 Feature Extraction

Protein sequences are consecutive amino acid residues, that can be regarded as text strings with an alphabet A of size |A| = 20. Many feature extraction methods have been developed in the past several years. Typically, these methods can be classified into two categories. One is based on amino acid composition (Chandran. 2008). The other one is an extension of the atomic length from only one amino acid to a K amino acid tuple, where K is an integer and larger than one, that can be further referred as 'K-tuple', such as 2-tuple in (Park and Kanehisa. 2003).

In this chapter, the features are extracted from protein primary sequence, based on both amino acid composition and K-mer patterns or K-tuples (Chandran. 2008). In Rough set method, the decision table is constructed for dimensionality

reduction, which consists of conditional attributes and decision attributes, A = (U, A U {d}) (Pawlak. 1993). The features extracted from protein primary sequence are considered as conditional attributes. In this chapter, conditional attributes set A consists of K-mer patterns or K-tuples of compositional values of the 20 amino acid in protein primary sequences. The four structural classes such as all α, all β, all α + β and all α / β are considered as decision attribute d as shown in Table 2.

Table 3 Decision Table (amino acid composition of 2-tuple feature vector)

Objects	**A**	**C**	**D**	**.**	**.**	**.**	**Y**	**AA**	**AC**	**AD**	**.**	**.**	**.**	**YY**	**Class**
1	5.42	1.81	6.02	.	.	.	2.71	5.42	8.13	9.64	.	.	.	8.13	1
2	6.15	1.54	6.15	.	.	.	1.54	5.38	6.92	9.23	.	.	.	7.69	1
3	12.5	0	6.25	.	.	.	0	4.69	8.59	11.72	.	.	.	8.59	2
4	8.57	1.43	10	.	.	.	1.43	10	2.86	14.29	.	.	.	11.43	2
5	12.5	0	6.25	.	.	.	0	4.69	8.59	11.72	.	.	.	8.59	3
6	8.96	1.49	10.45	.	.	.	1.49	10.45	2.99	14.93	.	.	.	10.45	3
7	12.5	0	6.25	.	.	.	0	4.69	8.59	11.72	.	.	.	8.59	4
8	8.7	1.45	10.14	.	.	.	1.45	10.14	2.9	14.49	.	.	.	11.59	4

The protein feature vector constructed using amino acid composition that represents a simple sequence that is widely used in prediction of various structural aspects. When K=1, the features are constructed from 20 amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y which are denoted as A1, A2,…, A19, and A20, and the number of occurrences of Ai in the sequence that is denoted as xi, the composition vector is defined as (x1/L, x2/L,…, xi/L), where L is the length of the sequence (Chen et al. 2006; Shi et al. 2007). However, the composition vector is insufficient to represent a sequence, since it only counts the frequencies of individual amino acids. Therefore, along with the 1-tuple feature set, 2-tuple features (when K=2) are constructed to generate the frequencies of amino acid pairs (dipeptides) provide more information since they reflect interaction between local amino acid pairs. Based on the frequency of collocation of amino acid pairs in the sequence, all dipeptides in the sequence can be counted. Since there are 400 possible dipeptides (AA, AC, AD,…, YY), a feature vector of that size is used to represent the occurrence of these pairs in the sequence (Gu et al. 2010). As a result, a total of 400 + 20 + 1= 421 features (420 conditional attributes and 1 decision attributes) is represented as in Table 2.

6 Feature Selection

Feature selection is one of the preprocessing work in Data Mining Techniques, which extracts the most relevant features from the very huge databases without affecting its originality. In this chapter, the rough set based supervised feature selection techniques for the protein primary sequences are proposed.

6.1 Rough Set Quick Reduct (RSQR)

The Rough Set based Quick Reduct (RSQR) algorithm presented in Algorithm 1 attempts to calculate a reduct without exhaustively generating all possible subsets (Jensen and Shen. 2004). It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset (Inbarani et al., 2014a).

According to the QUICKREDUCT algorithm, the dependency degree of each protein feature vector is calculated and the best candidate is chosen. However, it is not guaranteed to find a minimal feature set as its too greedy. Using the dependency degree to discriminate between candidates may lead the search down a non-minimal path. Moreover, in some cases, QUICKREDUCT algorithm cannot find a feature reduct that satisfies the accurate result that is the feature subset discovered may contain irrelevant features. The protein classification accuracy may be degraded when designing a classifier using the feature subset with irrelevant features (Velayutham and Thangavel. 2011).

6.2 Rough Set Particle Swarm Optimization (RSPSO)

Particle swarm optimization is a new population-based heuristic method discovered through simulation of social models of bird flocking, fish schooling, and swarming to find optimal solutions to the non-linear numeric problems. It was first introduced in 1995 by social-psychologist Eberhart and Kennedy (Kennedy and Eberhart. 1995). Particle swarm optimization is an efficient, simple, and an effective global optimization algorithm that can solve discontinuous, multimodal, and non-convex problems. PSO can therefore also be used on optimization problems that are partially irregular, noisy, change over time, etc. PSO is initialized with a population of random solutions, called particles (Shi and Eberhart. 1998).

Each particle is treated as a point in an S-dimensional space. The i^{th} particle is represented as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iS})$. The best previous position Pbest, the position giving the best fitness value) of any particle is recorded and represented as $Pi = (p_{i1}, p_{i2}, \ldots, p_{iS})$. The index of the best particle among all the particles in the population is represented by the symbol gbest. The rate of the position change (velocity) for particle i is represented as $Vi = (v_{i1}, v_{i2}, \ldots, v_{iS})$. The articles are manipulated according to the following equation:

$$Vid = w * vid + c1 * rand( ) *(pid -xid) + c2 * rand( ) *(pgd - xid) \quad (10)$$

$$xid = xid + vid \quad (11)$$

where d = 1, 2, . . ., S, w is the inertia weight, it is a positive linear function of time changing according to the generation iteration. A suitable selection of the inertia weight provides a balance between global and local exploration, and results in less iteration on average to find a sufficiently optimal solution. The acceleration constants c1 and c2 in Eq. (10) represent the weighting of the stochastic acceleration terms that pull each particle toward p_{best} and g_{best} positions (Wang et al., 2007).

Algorithm 1. QUICKREDUCT (C,D)

Input: C, the set of all conditional features; D, the set of decision features;
Output: Reduct R

(1) $R \leftarrow \{\}$
(2) do
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
(5) if $\gamma_{R \cup \{x\}}(D) > \gamma_{T}(D)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) Until $\gamma_{R}(D) > \gamma_{C}(D)$
(9) Return R

Particle velocities on each dimension are limited to a maximum velocity V_{max}. It determines how large steps through the solution space each particle is allowed to take. If V_{max} is too small, particles may not survey sufficiently beyond locally good regions. They could become intent in local optima. On the other hand, if V_{max} is too high particles might fly past good solutions. Eq. (11) is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best experience (position) and the group's best experience. Then the particle flies toward a new position according to Eq. (11). The performance of each particle is measured according to a pre-defined fitness function (Velayutham and Thangavel. 2011, Chen et al. 2012). The Pseudocode of PSO based Quick Reduct algorithm (Inbarani et al. 2014a) is given in Algorithm 2.

6.3 Harmony Search Algorithm

Harmony search (HS) is a relatively new population-based meta heuristic optimization algorithm, that imitates the music improvisation process where the musicians improvise their instrument's pitch by searching for a perfect state of harmony. It was able to attract many researchers to develop HS based solutions for many optimization problems (Geem and Choi. 2007; Degertekin. 2008).

Algorithm 2. Rough Set PSOQR(C,D)

Input: *C, the set of all conditional features;*
D, the set of decision features.
Output: Reduct R

Step 1: Initialize X with random position and V_i with random velocity
$\forall$ *: $X_i \leftarrow$ random Position();*
$V_i \leftarrow$ random Velocity();
Fit ← 0; globalbest←Fit;
Gbest← X_1; Pbest(1) ← X_1
For i =1. . .S
pbest(i)=X_i
Fitness (i)=0
End For
Step 2 : While Fit!=1 //Stopping Criterion
For i =1. . .S //for each particle
Compute fitness of feature subset of X_i
R ← Feature subset of X_i (1's of X_i)
$\forall x \in (C - R)$
$\gamma_{R\cup\{x\}}(D) = \frac{|POS_{R\cup\{x\}}(D)|}{|U|}$
Fitness(i)=$\gamma_{R\cup\{x\}}(D)$ $\forall X \subset R, \gamma_X(D) \neq \gamma_C(D)$
Fit = Fitness(i)
End For
Step 3: Compute best fitness
For i=1:S
If (Fitness(i) > globalbest)
globalbest ← Fitness(i);
gbest←X_i; getReduct(X_i)
Exit
End if
End For
UpdateVelocity(); //Update Velocity V_i's of X_i's
UpdatePosition(); //Update position of X_i's
//Continue with the next iteration
End {while}
Output Reduct R

HS imitates the natural phenomenon of musician's behavior when they co-operate the pitches of their instruments together to achieve a fantastic harmony as measured by aesthetic standards. This musicians' prolonged and intense process led them to the perfect state. It is a very successful meta heuristic algorithm that can explore the search space of a given data in parallel optimization environment, where each solution (harmony) vector is generated by intelligently exploring and exploiting a search space (Geem. 2009). It has many features that make it as a preferable technique not only as standalone algorithm, but also to be combined with other meta heuristic algorithms. The steps in the Harmony Search procedure are as follows (Geem et al. 2001):

Step 1. Initialize the problem and algorithm parameters.
Step 2. Initialize the harmony memory.
Step 3. Improvise a new harmony.
Step 4. Update the harmony memory.
Step 5. Check the stopping criterion.
These steps are briefly described in the following subsections.

Step 1: Initialize the Problem and Algorithm Parameters
The proposed Supervised RS-IHS algorithm is shown in Algorithm 3. In this approach, rough set based lower approximation is used for computing the dependency of conditional attributes on decision attribute, discussed in section 3. The rough set based objective function is defined as follows:

$$Max\ f(x) = \frac{|POS_{R\cup\{x\}}(D)|}{|U|} \tag{12}$$

The other parameters of the IHS algorithm such as harmony memory size (HMS), harmony memory considering rate (HMCR ϵ [0,1]), pitch adjusting rate (PAR ϵ [0,1]), and number of improvisations (NI) are also initialized in this step.

Step 2: Initialize the Harmony Memory
The harmony memory (HM) is a matrix of solutions with a size of HMS, where each harmony memory vector represents one solution as can be seen in Eq. 13. In this step, the solutions are randomly constructed and rearranged in a reversed order to HM, based on their fitness function values as $f(x^1) \leq f(x^2).....\leq f(x^{HMS})$ (Alia and Mandava. 2011)

$$HM = \left(\begin{array}{cccc|c} x_1^1 & x_2^1 & \ldots & x_n^1 & f(x^1) \\ x_1^2 & x_2^2 & \ldots & x_n^2 & f(x^2) \\ . & . & & . & . \\ . & . & & . & . \\ . & . & & . & . \\ x^{HMS}{}_1 & x^{HMS}{}_2 & \ldots & x^{HMS}{}_n & f(x^{HMS}) \end{array}\right) \tag{13}$$

For applying the proposed methods, each harmony vector in the HM is represented as binary bit strings of length *n*, where *n* is the total number of conditional attributes. This is the same representation as that used for PSO and GA-based feature selection. Therefore, each position in the harmony vector is an attribute subset.

a	b	c	d
1	0	1	0

For example, if a, b, c and d are attributes and if the selected random harmony vector is (1, 0, 1, 0), then the attribute subset is (a, c).

Step 3: Improvise a New Harmony

A new harmony vector $x' = x'_1, x'_2, \ldots, x'_n$ is generated based on three rules: (1) memory consideration, (2) pitch adjustment and (3) random selection. Generating a new harmony is called "improvisation" (Mahdavi et al., 2007).

In the memory consideration, the values of the new harmony vector are randomly inherited from the historical values stored in HM with a probability of *HMCR*. The HMCR, which varies between 0 and 1, is the rate of choosing one value from the historical values stored in the HM, while (1 – HMCR) is the rate of randomly selecting one value of the possible range of values. This cumulative step ensures that good harmonies are considered as the elements of New Harmony vectors (Alia and Mandava. 2011).

$$x'_i \leftarrow \begin{cases} x'_i \in \{x_i^1, x_i^2 \ldots x_i^{HMS}\} \text{ with probability } HMCR \\ x'_i \in X_i \quad \text{with probability } (1 - HMCR) \end{cases} \tag{14}$$

For example, a HMCR of 0.95 indicates that the HS algorithm will choose the decision variable value from historically stored values in the HM with 90% probability or from the entire possible range with a (100–90) % probability. Every component obtained by the memory consideration is examined to determine whether it should be pitch-adjusted (Navi. 2013).

This operation uses the PAR parameter, which is the rate of pitch adjustment as follows:

$$x'_i \leftarrow \begin{cases} \text{Yes (Adjusting Pitch)} & \text{with probability PAR} \\ \text{No (doing nothing)} & \text{with Probability (1 –PAR)} \end{cases} \tag{15}$$

The value of (1 – PAR) sets the rate of doing nothing. If a generated random number $rand() \in [0, 1]$ is within the probability of PAR then, the new decision variable (x'_i) will be adjusted based on the following equation:

$$(x'_i) = (x'_i) \pm rand() * bw \tag{16}$$

Here, *bw* is an arbitrary distance bandwidth used to improve the performance of HS and *rand()* is a function that generates a *random number* $\in$ *[0, 1]*. Actually, *bw* determines the amount of movement or changes that may have occurred to the components of the new vector.

In this Step, memory consideration, pitch adjustment or random selection is applied to each variable of the New Harmony vector in turn. Consequently, it explores more solutions in the search space and improves the searching abilities (Geem. 2009).

Step 4: Update the Harmony Memory

For each new value of harmony, the value of the objective function *f(x'),* is calculated. If the New Harmony vector is better than the worst harmony in the HM, the New Harmony is included in the HM and the existing worst harmony is excluded from the HM.

Step 5: Check the Stopping Criterion

If the stopping criterion (maximum number of improvisations) is satisfied, computation is terminated. Otherwise, Steps 3 and 4 are repeated. Finally the best harmony memory vector is selected and is considered as the best solution to the problem.

6.4 Rough Set Improved Harmony Search (RSIHS)

This method HSA is developed by Mahdavi et al. 2007 (Geem. 2006; Mahdavi et al. 2007). In HSA , HMCR, PAR, bw, but PAR and bw are very important parameters in fine-tuning of optimized solution vectors. The traditional HS algorithm uses a fixed value for both PAR and bw. In the HS method, PAR and bw values are adjusted in Step 1 and cannot be changed during new generations (Al-Betar et al. 2010). The main drawback of this method is that the numbers of iterations increases to find an optimal solution. To improve the performance of the HS algorithm and to eliminate the drawbacks that lies with fixed values of PAR and bw, IHSA uses variables PAR and bw in improvisation step (Step 3) (Chakraborty et al., 2009). PAR and bw change dynamically with generation number and expressed as follows:

$$PAR(gn) = PAR_{min} + \frac{PAR_{max} - PAR_{min}}{NI} * gn \tag{17}$$

Where, *PAR(gn)* = Pitch Adjusting Rate for each generation PAR_{min} = Minimum Pitch Adjusting Rate, PAR_{max} = Maximum Pitch Adjusting Rate, *NI* = Number of Improvisations and *gn* = Generation Number

Algorithm 3. Rough Set IHSQR(C,D)

Input : C, the set of conditional attributes; D, the set of decision attributes
Output : Best Reduct (feature)

Step 1: Define the fitness function,f(X)
Initialize the variables HMS=10 // Harmony Memory Size (population)
HMCR = 0.95 // Harmony Memory Consideration Rate (for improvisation)
NI = 100 // Maximum number of Iterations,
PVB //Possible value bound of X
PAR_{min},PAR_{max},bw_{min}, bw_{max},// Pitch Adjusting Rate & bandwidth Є(0 to 1)
fit =0;
X_{old}= X_1; bestfit =X_1;bestreduct={};

Step 2: Initialize Harmony Memory, HM = (X_1,X_2,….X_{HMS})
For i = 1 to HMS //for each harmony
∀ : X_i // X_i is the i^{th}harmony vector of HM
//Compute fitness of feature subset of X_i
R←Feature subset of X_i (1's of X_i)
$\forall x \in (C - R)$
$\gamma_{R\cup\{x\}}(D) = \frac{|POS_{R\cup\{x\}}(D)|}{|U|}$
$f(X_i) = \gamma_{R\cup\{x\}}(D) \quad \forall X \subset R, \gamma_X(D) \neq \gamma_C(D)$
if $f(X_i)$ > fit
$fit \leftarrow f(X_i)$
$X_{old} \leftarrow X_i$
End if
End for

Step 3: Improvise new Harmony Memory
While iter ≤ NI or fit ==1// Stopping Criterion
for j=1,2,…NVAR
∀ :X_{old} (j) // x is the variable of X
Update Pitch Adjusting Rate();
Update bandwidth();
if rand ( ) ≤ HMCR // rand Є [0,1]
// construct the new harmony Xnewfrom the best harmony vector Xold

$X_{new} \leftarrow X_{old}$; // assigning the best harmony to the new harmony

if rand ( ) ≤ PAR // rand ∈ [0,1]

$X_{new}(j) = X_{new}(j) \pm rand() * bw$

end if

else

//choose a random value of variable Xnew

$X_{new}(j)=PVB_{lower} + rand( ) * (PVB_{upper} – PVB_{lower})$

end if

end for

Step 4: Update the new Harmony Memory

Compute fitness function for New Harmony X_{new} as defined in Step 2.

if $f(X_{new}) \geq f(X_{old})$

// Accept and replace the old harmony vector with new harmony.

$X_{old} \leftarrow X_{new}$;

if $f(X_{new}) > fit$

$fit \leftarrow f(X_{new})$;

$bestfit \leftarrow X_{new}$;

End if

Exit

end if

//continue with the next iteration

end while

bestreduct ← feature subset of bestfit // Reduced feature subset: 1's of bestfit

$$bw(gn) = bw_{max} * exp(c * gn); \quad c = ln [(bw_{min} / bw_{max})] / NI \quad (18)$$

Where, $bw(gn)$ = Bandwidth for each generation bw_{min} = Minimum bandwidth bw_{max} = Maximum bandwidth.

The Pseudocode of the Improved HS using Rough Set based Quick Reduct is given in Algorithm 3.

7 Experimental Analysis

7.1 Data Source

In this chapter, the protein primary sequence dataset is derived from PDB (http://www.rcsb.org/pdb) on SCOP classification. The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences (Chinnasamy et al., 2004). The data set consists of sequences with 7623 of all α, 10672 of all β, 11048 of all α + β and 11961 of all α / β (Cao et al., 2006). Among one thousand sequences with combinations of all α , all β, all α + β and all α / β, each 250 sequences are taken for this study.

7.2 Results and Discussion

7.2.1 Results

A sequence of experiments was conducted to show the efficacy of the proposed feature selection algorithm. All experiments have been run on a machine with 3.0 GHz CPU and 2 GB of RAM. The proposed Improved Harmony Search hybridized with Rough Set Quick Reduct feature selection algorithm is implemented in Matlab 2012a. The operating system is Windows Vista. For experimental studies, 1000 sequences were extracted from the SCOP classification of Protein Data Bank. The following section describes the implementation results of this study.

Table 4 Number of features selected by feature selection algorithms

Protein Data Set	***Number of features extracted using K-tuple sequences***			***Number of features selected***		
	K	***Number of Conditional features***	***Number of decision features***	***RSQR***	***PSOQR***	***Improved HSQR***
1000 objects	1	20^1	20	1	14	12	11
	2	20^2	420	1	280	217	32

The feature subset length and the classification quality are the two criteria that is considered to assess the performance of algorithms. Comparing the first criterion, the number of selected features shown in Table 3, the proposed algorithm Improved Harmony Search Quick Reduct outperforms the Quick Reduct and PSO Quick Reduct algorithms in selecting smaller subset of features in both the tuples which is compared in Fig. 2.

Next, the other criterion, predictive accuracy is compared. The PSO Quick Reduct and the proposed Improved HS Quick Reduct algorithms revealed best accuracy than the Unreduced Set (all features) and Quick Reduct algorithms. The predictive accuracy results of the existing and proposed algorithms of 1-tuple and 2-tuples compared in Table 4 and 5. Figs. 4 – 8 show the predictive accuracy for each of the feature selection algorithms considered in this work.

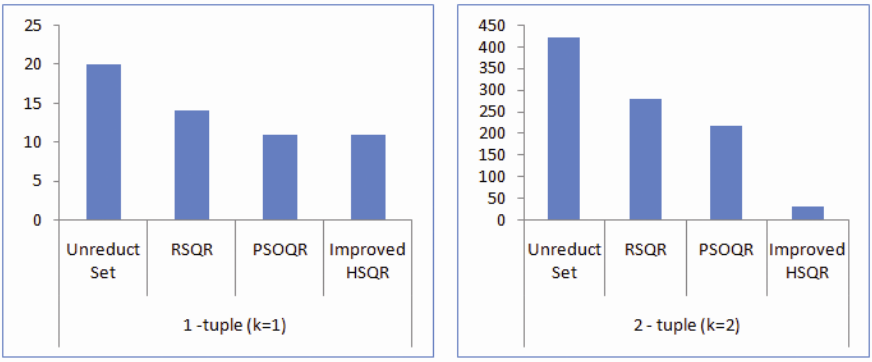

Fig. 2 Reduced Feature Set of 1-tuple and 2-tuples

Table 5 Classification accuracy of 1-tuple (K=1) protein sequence features

Classification Method	*Predictive Accuracy(%)*			
	Unreduct Set	*Quick Reduct*	*PSO QR*	*Improved HSQR*
IBK	89.7	90.3	91.1	91.1
Kstar	89.5	90.1	90.6	90.0
Randomforest	88.4	76.7	86.1	88.5
J48	79.9	82.3	81.6	81.3
JRip	74.3	75.0	76.2	78.3

7.2.2 Discussion

Experimental results show that the use of irrelevant features hurts classification accuracy and Feature Selection technique is used to reduce redundancy in the information provided by the selected features. Using only a small subset of selected features, the proposed Improved HS Quick Reduct algorithms obtained better classification accuracy than the existing algorithms compared in this study.

Table 6 Classification accuracy of 2-tuple (K=2) protein sequence features

Classification Method	*Predictive Accuracy (%)*			
	Unreduct Set	*Quick Reduct*	*PSO QR*	*Improved HSQR*
IBK	83.5	51.3	86.7	91.3
Kstar	86.0	50.5	86.7	90.8
Randomforest	86.7	80.4	90.1	91.7
J48	78.9	80.6	88.6	90.5
JRip	76.2	78.7	89.3	91.8

To compare the performance of the above feature selection algorithms, classification techniques such as IBK, Kstar, Randomforest, J48 and JRip are applied in this work. The selected feature subset of protein sequences is used as the input of the classifiers. All experiments were carried out using a ten-fold cross validation approach.

The Fig. 3 demonstrates the performance of feature subset selected by each feature selection approaches using IBK classifier. The Improved HSQR algorithm outperforms the Quick Reduct algorithm. The predictive accuracy of the unreduct set (all features) and the PSO QR algorithm is slightly lesser than the proposed algorithm.

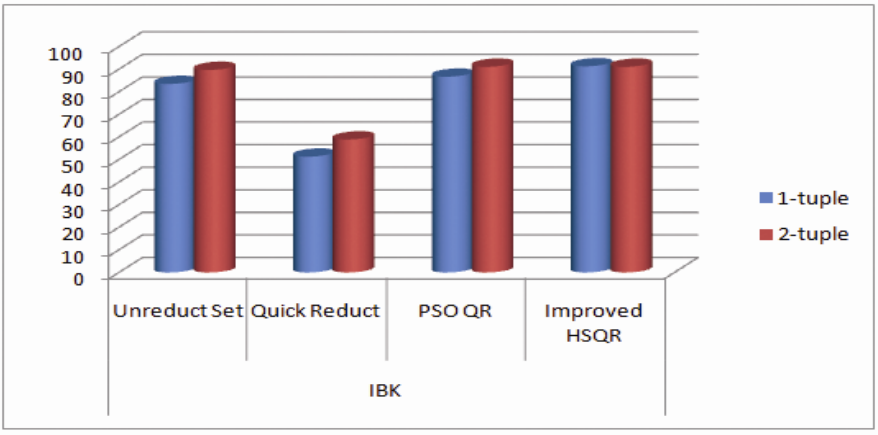

Fig. 3 Predictive Accuracy of IBK Classifier

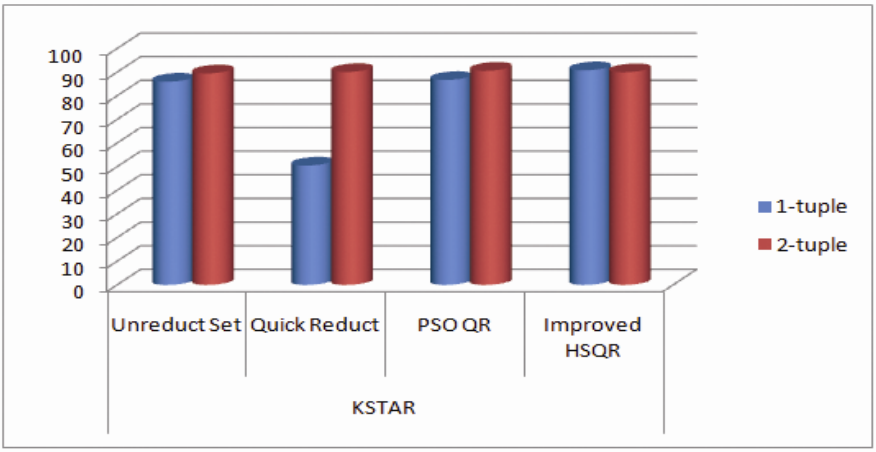

Fig. 4 Predictive Accuracy of KSTAR Classifier

The Fig. 4 show the predictive accuracy of the KStar Classifier which is used to evaluate the performance of the feature selection algorithms. It shows the highest accuracy for all the feature selection algorithms and also to the unreduct set of protein feature set. In both the 1-tuple and 2-tuples feature set the Improved HSQR algorithm outperforms the other algorithms, but the Quick Reduct algorithm shows the poor performance in 1-tuple set.

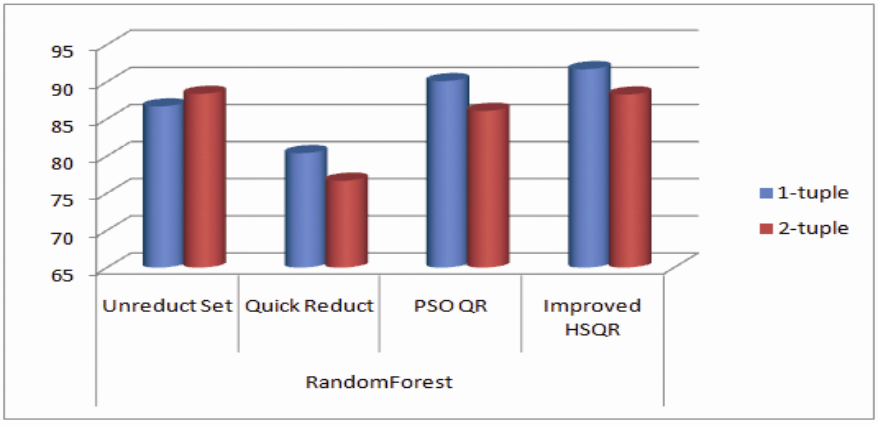

Fig. 5 Predictive Accuracy of RandomForest Classifier

The Performance evaluation of the feature selection algorithms is implemented with Randomforest classifier which can be visualized in Fig.5. The proposed algorithm, Improved HSQR outperforms all the feature selection algorithms of this study at high rates with highest predictive accuracy.

The Fig. 6 demonstrates the performance of feature subset selected by each feature selection approaches using J48 classifier. The Improved HSQR algorithm outperforms all the feature selection algorithms with highest predictive accuracy. In the 2-tuple feature set, the Quick Reduct algorithm outperforms all the feature selection algorithms with its high predictive accuracy.

The feature selection algorithms with its feature subsets are evaluated by the JRip classifier which can be visualized in Fig.7. The proposed algorithm, Improved HSQR outperforms all the feature selection algorithms of this study at high rates with a highest predictive accuracy in both 1-tuple and 2-tuple feature sets.

The results strongly suggest that the proposed method can assist in solving the high-dimensionality problem, and accurately classifies the protein sequences to its corresponding structures and can be very useful for predicting the function of the protein. If Rough set based feature selection methods like QR and PSOQR consumes lot of processing time, IHSQR method is the best method applied to reduce the time needed for executing the algorithm.

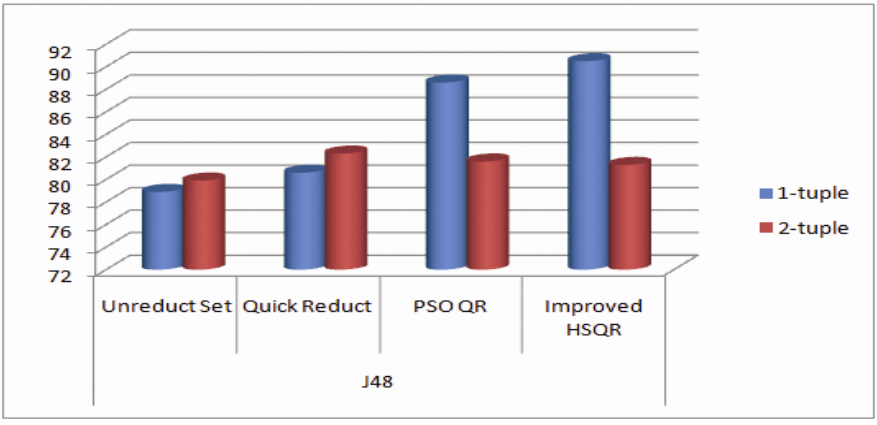

Fig. 6 Predictive Accuracy of J48 Classifier

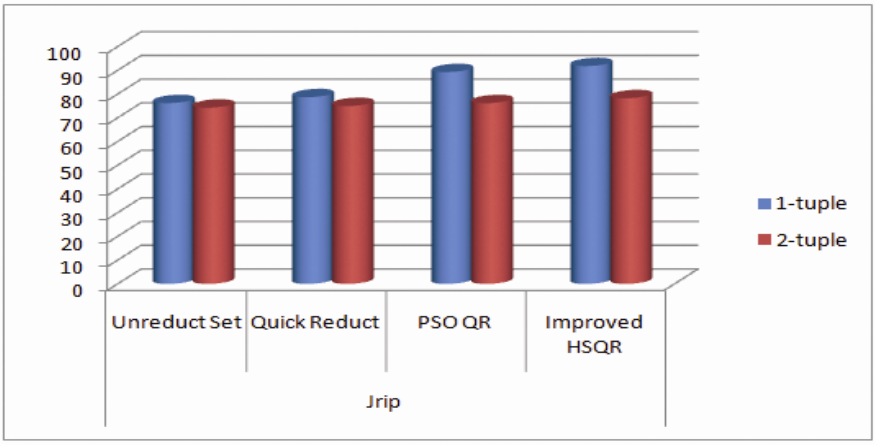

Fig. 7 Predictive Accuracy of JRIP Classifier

8 Conclusion and Future Work

This chapter introduced a hybridized approach to combine the strengths of Rough Set Theory (RST) and Improved Harmony Search Algorithm to solve the high dimensionality problem. The IHS algorithm has a number of advantages over conventional approaches. Experimental comparative studies show that the proposed approach can reveal best solutions, quickly converges, strong search capability in the problem space and can capably find minimal reduct in very big databases. The experimental results show that how the meta-heuristics approaches increases the predictive accuracy for bio-informatics datasets. The proposed methods are compared with an existing rough set based supervised algorithms and classification accuracy measures are used to evaluate the performance of the proposed approaches. Hence the analysis section clearly proved the effectiveness of Harmony Search and RST based approaches for Protein Sequence classification for predicting their structures and function in the future. As a future work, this model can be applied for proteomics and Genomic tasks such as predicting protein function, protein-protein interaction, DNA Sequence Analysis and etc., in the field of bio-informatics. This model can also be extended to hybridize advanced swarm intelligence techniques.

References

Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Application of ant colony optimization for feature selection in text categorization. In: Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2008), Hong Kong, June 1-6, pp. 2867–2873 (2008)

Al-Ani, A., Khushaba, R.N.: A Population Based Feature Subset Selection Algorithm Guided by Fuzzy Feature Dependency. In: Hassanien, A.E., Salem, A.-B.M., Ramadan, R., Kim, T.-h. (eds.) AMLTA 2012. Communications in Computer and Information Science, vol. 322, pp. 430–438. Springer, Heidelberg (2012)

Al-Betar, M., Khader, A., Liao, I.: A harmony search with multi-pitch adjusting rate for the university course timetabling. In: Geem, Z.W. (ed.) Recent Advances in Harmony Search Algorithm. SCI, vol. 270, pp. 147–161. Springer, Heidelberg (2010)

Alia, O.M., Mandava, R.: The variants of the harmony search algorithm: an Overview. Artificial Intelligence Review 36(1), 49–68 (2011)

Azar, A.T.: Neuro-fuzzy feature selection approach based on linguistic hedges for medical diagnosis. International Journal of Modelling, Identification and Control (IJMIC) 22(3) (forthcoming, 2014)

Azar, A.T., Hassanien, A.E.: Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier. Soft Computing (2014), doi:10.1007/s00500-014-1327-4.

Azar, A.T., Banu, P.K.N., Inbarani, H.H.: PSORR - An Unsupervised Feature Selection Technique for Fetal Heart Rate. In: 5th International Conference on Modelling, Identification and Control (ICMIC 2013), Egypt, August 31-September 1-2, pp. 60–65 (2013)

Basiri, M.E., Ghasem-Aghaee, N., Aghdam, M.H.: Using ant colony optimization-based selected features for predicting post-synaptic activity in proteins. In: Marchiori, E., Moore, J.H. (eds.) EvoBIO 2008. LNCS, vol. 4973, pp. 12–23. Springer, Heidelberg (2008)

Blum, C., Dorigo, M.: The hyper-cube framework for ant colony optimization. IEEE Transaction on Systems, Man, and Cybernetics – Part B 34(2), 1161–1172 (2004)

Caragea, C., Silvescu, A., Mitra, P.: Protein sequence classification using feature hashing. In: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, November 12-15. Proteome Science 2012, vol. 14, p. S14 (2011), doi:10.1186/1477-5956-10-S1-S14.

Cao, Y., Liu, S., Zhang, L., Qin, J., Wang, J., Tang, K.: Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7(1), 20 (2006), doi:10.1186/1471-2105-7-20.

Chakraborty, P., Roy, G.G., Das, S., Jain, D., Abraham, A.: An improved harmony search algorithm with differential mutation operator. Fundamenta Informaticae 95(4), 1–26 (2009), doi:10.3233/FI-2009-181.

Chandran, C.P.: Feature Selection from Protein Primary Sequence Database using Enhanced Quick Reduct Fuzzy-Rough Set. In: Proceedings of International Conference on Granular Computing, GrC 2008, Hangzhou, China, August 26-28, pp. 111–114 (2008), doi:10.1109/GRC.2008.4664758

Chandrasekhar, T., Thangavel, K., Sathishkumar, E.N.: Verdict Accuracy of Quick Reduct Algorithm using Clustering and Classification Techniques for Gene Expression Data. IJCSI International Journal of Computer Science Issues 9(1), 357–363 (2012)

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y.: Using pseudo amino acid composition and support vector machine to predict protein structural class. Journal of Theoretical Biology 243(3), 444–448 (2006)

Chen, L.F., Su, C.T., Chen, K.H., Wang, P.C.: Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. International Journal of Neural Computing and Applications 21(8), 2087–2096 (2012)

Chinnasamy, A., Sung, W.K., Mittal, A.: Protein Structure and Fold Prediction Using Tree-Augmented Bayesian Classifier. Journal of Bioinformatics and Computational Biology 3(4), 803–819 (2005)

Chouchoulas, A., Shen, Q.: Rough set-aided keyword reduction for text categorization. An International Journal of Applied Artificial Intelligence 15(9), 843–873 (2001), doi:10.1080/088395101753210773

Degertekin, S.O.: Optimum design of steel frames using harmony search algorithm. Structural and Multidisciplinary Optimization 36(4), 393–401 (2008)

Du, P., Wang, X., Xu, C., Gao, Y.: PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical Biochemistry 425(2), 117–119 (2012)

Ferrandiz, S., Boullé, M.: Multivariate Discretization by Recursive Supervised Bipartition of Graph. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), vol. 3587, pp. 253–264. Springer, Heidelberg (2005)

Fleuret, F.: Fast Binary Feature Selection with Conditional Mutual Information. Journal of Machine Learning Research 5(1), 1531–1555 (2004)

Freitas, A.A., de Carvalho, A.C.P.L.F.: A tutorial on hierarchical classification with applications in bioinformatics. Research and Trends in Data Mining Technologies and Applications 99(7), 175–208 (2007)

Fu, X., Tan, F., Wang, H., Zhang, Y.Q., Harrison, R.R.: Feature similarity based redundancy reduction for gene selection. In: Proceedings of the International Conference on Data Mining, Las Vegas, NV, USA, June 26-29, pp. 357–360 (2006)

Geem, Z.W., Kim, J.H., Loganathan, G.V.: A New Heuristic Optimization Algorithm: Harmony Search. Simulation 76(2), 60–68 (2001), doi:10.1177/003754970107600201

Geem, Z.W.: Improved harmony search from ensemble of music players. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 86–93. Springer, Heidelberg (2006)

Geem, Z.W., Choi, J.-Y.: Music composition using harmony search algorithm. In: Giacobini, M. (ed.) EvoWorkshops 2007. LNCS, vol. 4448, pp. 593–600. Springer, Heidelberg (2007)

Geem, Z.W.: Particle-swarm harmony search for water network design. Engineering Optimization 41(4), 297–311 (2009)

Gu, Q., Ding, Y., Jiang, X., Zhang, T.: Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. Amino Acids 38(4), 975–983 (2010)

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3(1), 1157–1182 (2003)

Hall, M., Frank, E., Holmes, G., Pfahringer, G., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)

Hor, C., Yang, C., Yang, Z., Tseng, C.: Prediction of Protein Essentiality by the Support Vector Machine with Statistical Tests. In: Proceedings of 11th International Conference on Machine Learning and Applications, USA, vol. 1(1), pp. 96–101 (2012), doi:10.1109/ICMLA.2012.25

Inbarani, H.H., Banu, P.K.N., Andrews, S.: Unsupervised hybrid PSO - quick reduct approach for feature reduction. In: Proceedings of International Conference on Recent Trends in Information Technology, ICRTIT 2012, April 19-21, pp. 11–16 (2012), doi:10.1109/ICRTIT.2012.6206775

Inbarani, H.H., Banu, P.K.N.: Unsupervised hybrid PSO – relative reduct approach for feature reduction. In: Proceedings of International Conference on Pattern Recognition, Informatics and Medical Engineering, Salem, Tamil Nadu, India, March 21-23, pp. 103–108 (2012), doi:10.1109/ICPRIME.2012.6208295

Inbarani, H.H., Jothi, G., Azar, A.T.: Hybrid Tolerance-PSO Based Supervised Feature Selection For Digital Mammogram Images. International Journal of Fuzzy System Applications (IJFSA) 3(4), 15–30 (2013)

Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Computer Methods and Programs in Biomedicine 113(1), 175–185 (2014a)

Inbarani, H.H., Banu, P.K.N., Azar, A.T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Computing and Applications (2014b), doi:10.1007/s00521-014-1552-x.

Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: rough and fuzzy-rough based approaches. IEEE Transactions on Knowledge and Data Engineering 16(12), 1457–1471 (2004)

Jothi, G., Inbarani, H.H.: Soft set based quick reduct approach for unsupervised feature selection. In: Proceedings of International Conference on Advanced Communication Control and Computing Technologies, Tamil Nadu, India, August 23-25, pp. 277–281. IEEE (2012)

Kattan, A., Abdullah, R., Salam, R.A.: Harmony search based supervised training of artificial neural networks. In: Proceedings of International Conference on Intelligent Systems, Modeling and Simulation (ISMS 2010), Liverpool, England, pp. 105–110 (2010), doi:10.1109/ISMS.2010.31

Kennedy, J., Eberhart, R.C.: A new optimizer using particle swarm theory. In: Proceedings of 6th International Symposium on Micro Machine and Human Science, Nagoya, pp. 39–43 (1995), doi:10.1109/MHS.1995.494215

Kotsiantis, S., Kanellopoulos, D.: Discretization Techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering 32(1), 47–58 (2006)

Lin, H., Ding, H., Guo, F., Huang, J.: Prediction of subcellular location of mycobacterial protein using feature selection techniques. Molecular Diversity 14(4), 667–671 (2010)

Mahdavi, M., Fesanghary, M., Damangir, E.: An improved harmony search algorithm for solving optimization problems. Applied Mathematics and Computation 188(2), 1567–1579 (2007)

Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)

Navi, S.P.: Using Harmony Clustering for Haplotype Reconstruction from SNP fragments. International Journal of Bio-Science and Bio-Technology 5(5), 223–232 (2013)

Nemati, S., Boostani, R., Jazi, M.D.: A novel text-independent speaker verification system using ant colony optimization algorithm. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008 2008. LNCS, vol. 5099, pp. 421–429. Springer, Heidelberg (2008)

Nemati, S., Basiri, M.E., Ghasem-Aghaee, N., Aghdam, M.H.: A novel ACO–GA hybrid algorithm for feature selection in protein function prediction. Expert Systems with Applications 36(10), 12086–12094 (2009)

Park, K.J., Kanehisa, M.: Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19(13), 1656–1663 (2003)

Pawlak, Z.: Rough Sets: Present State and The Future. Foundations of Computing and Decision Sciences 18(3-4), 157–166 (1993)

Pawlak, Z.: Rough Sets and Intelligent Data Analysis. Information Sciences 147(1-4), 1–12 (2002)

Pedergnana, M., Marpu, P.R., Mura, M.D., Benediktsson, J.A., Bruzzone, L.: A Novel supervised feature selection technique based on Genetic Algorithms. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, Munich, July 22-27, pp. 60–63 (2012), doi:10.1109/IGARSS.2012.6351637

Peng, Y.H., Wu, Z., Jiang, J.: A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics 43(1), 15–23 (2010)

Rentzsch, R., Orengo, C.: Protein function prediction-the power of multiplicity. Trends in Biotechnology 27(4), 210–219 (2009)

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., Ofran, Y.: Automatic prediction of protein function. Cellular and Molecular Life Sciences 60(12), 2637–2650 (2003)

Saeys, Y., Inza, I.N., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nature Review Genetics 11(9), 647–657 (2010)

Seok, L.K., Geem, Z.W.: A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. Computer Methods in Applied Mechanics and Engineering 194(36-38), 3902–3933 (2005)

Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998)

Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.M., Xie, J.: Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33(1), 69–74 (2007)

Siedlecki, W., Sklansky, J.: A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters 10(5), 335–347 (1989)

Velayutham, C., Thangavel, K.: Unsupervised Quick Reduct Algorithm Using Rough Set Theory. Journal of Electronic Science and Technology 9(3), 193–201 (2011)

Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R.: Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters 28(4), 459–471 (2007)

Wei, X.: Computational approaches for biological data analysis. Doctoral Dissertation, Tufts Uiversity Medford, MA, USA (2010) ISBN: 978-1-124-21198-5

Wong, A., Shatkay, H.: Protein Function Prediction using Text-based Features extracted from the Biomedical Literature: The CAFA Challenge. BMC Bioinformatics 14(3), S14 (2013), doi:10.1186/1471-2105-14-S3-S14

Xie, J., Xie, W., Wang, C., Gao, X.: A Novel Hybrid Feature Selection Method Based on IFSFFS and SVM for the Diagnosis of Erythemato - Squamous Diseases. In: Proceedings of JMLR Workshop and Conference Proceedings. Workshop on Applications of Pattern Analysis, vol. 11(1), pp. 142–151. MIT Press, Windsor (2010)

Autonomic Discovery of News Evolvement in Twitter

Mariam Adedoyin-Olowe, Mohamed Medhat Gaber,
Frederic Stahl, and João Bártolo Gomes

Abstract. Recently the continuous increase in data sizes has resulted in many data processing challenges. This increase has compelled data users to find automatic means of looking into databases to bring out vital information. Retrieving information from 'Big data', (as it is often referred to) can be likened to finding 'a needle in the haystack'. It is worthy of note that while big data has several computational challenges, it also serves as gateway to technological preparedness in making the world a global village. Social media sites (of which Twitter is one) are known to be big data collectors as well as an open source for information retrieval. Easy access to social media sites and the advancement of technology tools such as the computer and smart devices have made it convenient for different entities to store enormous data in real time. Twitter is known to be the most powerful and most popular microbloging tool in social media. It offers its users the opportunity of posting and receiving instantaneous information from the network. Traditional news media follow the activities on Twitter network in order to retrieve interesting tweets that can be used to enhance their news reports and news updates. Twitter users include hashtags symbols (#) as prefix to keywords used in tweets to describe its content and to enhance the readability of their tweets. This chapter uses the *Apriori* method

Mariam Adedoyin-Olowe · Mohamed Medhat Gaber
School of Computing Science, Robert Gordon University,
Aberdeen, United Kingdom, AB10 7GJ, UK
e-mail: {m.a.adedoyin-olowe,m.gaber1}@rgu.ac.uk

Frederic Stahl
School of Systems Engineering, University of Reading,
P.O. Box 225, Whiteknights, Reading, RG6 6AY, UK
e-mail: F.T.Stahl@reading.ac.uk

João Bártolo Gomes
Institute for Infocomm Research (I2R), A*STAR, Singapore,
1 Fusionopolis Way Connexis, Singapore 138632
e-mail: bartologjp@i2r.a-star.edu.sg

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_7

for Association Rule Mining *(ARM)* and a novel methodology termed Rule Type Identification-Mapping (RTI-*Mapping*) which is inherited from Transaction-based Rule Change Mining *TRCM* (Adedoyin-Olowe et al., 2013) and Transaction-based Rule Change Mining-Rule Type Identification (TRCM-RTI) (Gomes et al., 2013) to map Association Rules *(ARs)* detected in tweets' hashtags to evolving news reports and news updates of traditional news agents in real life. *TRCM* uses Association Rule Mining (ARM) to analyse tweets on the same topic over consecutive periods t and $t+1$ using Rule Matching (RM) to detected changes in *ARs* such as emerging, unexpected, new and dead rules. This is obtained by setting user-defined Rule Matching Threshold (RMT) to match rules in tweets at time t with those in tweets at $t+1$ in order to ascertain rules that fall into the different patterns. TRCM-RTI is a methodology built from TRCM, it identifies rule types of evolving *ARs* present in tweets' hashtags at different time periods. This chapter adopts RTI-*Mapping* from methodologies in (Adedoyin-Olowe et al., 2013) and (Gomes et al., 2013) to map *ARs* with online evolving news of top traditional news agents in order to detect and track news and news updates of evolving events. This is an initial experiment of *ARs* mapping to evolving news. The mapping is done manually at this stage and the methodology is validated using four events and news topics as case studies. The experiments show substantial result on the selected news topics.

1 Introduction

'Big data' refers to the accumulation of data in different data warehouses ranging from personal data stored on smart devices in private homes to 'big data' generated on the World Wide Web. The word 'big data' was first coined in 2005 and Web 2.0 a year earlier. Digital 'big data' is generated at different levels of data handling at a very remarkable speed and the prospect of its continuity is evident currently in data warehouses globally. Some 'big data experts have anticipated the likelihood of data explosion termed the new 'industrial revolution. It is believed that the revolution will cause a complete overhaul of business-operational and societal opportunities that may result in new innovation business models and news business insights that promotes competitive advantage (Bloem et al., 2012). Currently business retailers sell their products without being confined within building walls, selling and buying products and services on the Internet is just a click away. This gives both sellers and buyers the opportunity of transacting business without leaving their domain. International businesses experience a boom due to Internet business transactions. Parties involved in Internet business dealings are gaining confidence in technologies used in online business transactions all around the world despite security risks involved in using some of these technologies .

'Big data' poses huge challenges to traditional database storage and handling systems. As of 2011 the world is said to create 2.5 quintillion bytes of data daily (Henno et al., 2013). The 10 largest databases in the world stores well over 9.786322 petabytes of data; excluding databases of organisations that cannot be made public. The World Data Centre for Climate (WDCC) alone has 220 terabytes of data

available online with 110 terabytes of climate simulation data and 6 petabytes of data on magnetic tapes. National Energy Research Scientific Computing Centre (NERSCC) is said to have about 2.8 petabytes of data. Google is reported to have over 33 trillion entries in their database while Sprint, America largest wireless telecommunication network processes over 365 million call detail records of its over 55 million users per day. LexisNexis, a company providing computer-assisted legal research services keep a database of over 250 terabytes, while YouTube has a video database of around 45 terabytes. Amazon, an online shopping store is reported to have over 42 terabytes and the Library of Congress has 130 million items in books, photographs and maps and nearly 530 miles of shelves. Social media sites generate data every second, and the generated data requires more up-to-date techniques to refine and store the 'big data' in a way that it can be available in a meaningful state when needed. Every minute Twitter users send over 100,000 tweets while Facebook users share 684,478 pieces of contents with about 34,722 'Likes' for brands and organisations. In addition, YouTube users upload 48 hours of new videos and Google receives over 2 million search queries. Users on Instagram share 3,600 new photos while Apple receives about 47,000 App downloads.

It is evident that the global databases are increasing at an alarming rate and as the data increases, human knowledge about data decreases. It is therefore necessary to employ high quality tools in retrieving constructive, valid and useful contents from 'big data'. Many big organisations are already using innovative technologies that enable information technology to use big data analytics to generate higher levels of insight, resulting in significant competitive business advantages. However, the role of the social media in the growth of global data warehouses cannot be overemphasized. Individuals (such as celebrities), big organisations, government bodies and even presidents of nations visit social media sites like Twitter, Facebook and YouTube, either to post contents or to find out what stakeholders and the public say about them.

This chapter uses *Apriori* method of Association Rule Mining *(ARM)* and a novel methodology termed Rule Type Identification-Mapping (RTI-*Mapping*) which is inherited from Transaction-based Rule Change Mining *TRCM* (Adedoyin-Olowe et al., 2013) and Transaction-based Rule Change Mining-Rule Type Identification (TRCM-RTI) (Gomes et al., 2013) to retrieve information in form of news and news updates from Twitter 'Big Data'. TRCM-RTI is a methodology used to identify evolving Association Rules *(ARs)* present in tweets' hashtags at different periods of time t and $t+1$. In Adedoyin-Olowe et al. (2013) four evolving *ARs* patterns namely; New Rules, Emerging Rules, Unexpected Consequent and unexpected Conditional Rules and Dead Rules were identified. This chapter extends the experiments in Adedoyin-Olowe et al. (2013) and Gomes et al. (2013) by proposing RTI-*Mapping* which is used to map *ARs* in tweets' hashtags with online evolving news of selected traditional news agents.The mapped news is then tracked over a specified time slot in order to detect every updates of the news. The mapping is done manually at this stage and the methodology is validated using four events/news topics as case studies. This novel methodology shows substantial results on the selected case studies. The term *Time Frame Windows (TFWs)* of rules as well as Tweet-originated and

news-originated (’TwO-NwO’) stories are also explored in this chapter. The *TFWs* of *ARs* present in tweets is compared to *TFWs* of news stories in real life and the experimental results show that news in real life have more rapid *TFWs* than hashtags of tweets. The Static Hashtags Evolving News *(SHEN)* state is also discussed.

The remainder of the chapter is organised as follows: Section 2 discusses the related work. Section 3 explores Twitter background while Section 4 explains the overview of Association Rule Mining (ARM) and ARs in tweets. Section 5 discusses the evolution of the TRCM methodologies and Section 6 analyses tweets trend using *TRCM- RTI*. Section 7 evaluates the empirical study in this chapter. The chapter is concluded in Section 8 by stating the plan for future work.

2 Related Work

2.1 Big Data: Challenges and Opportunities

The affordance of the web 2.0 in recent decades is known to be instrumental to the massive net-centric data generation around today (Evans, 2010; Parameswaran and Whinston, 2007; Tang and Liu, 2010). Big data generation can be traced to diverse sectors of humanity, starting from personal data to organisation data and then to World Wide Web data stored in different data storage systems such as mobile devices, sensors, CDs and web-clicks. Organisational data for instance begins with the collection of customers and personnel records stored in organisational databases then to customers Relationships Management (CRM) (Bloem et al., 2012). Organisational data progresses to the web where other data resides and all these data form the digital ‘big data’. In this era of Internet, data grows rapidly from megabytes to petabytes, which concludes that data collection, and generation is never ending. Big data poses many challenges including (but not limited to) management and semantics of ’big data’ (Bizer et al., 2012). Businesses now go online for many reasons, including the need to cut down on running cost and to promote the opportunity of reaching more customers and creating social presence. According to Kaplan and Haenlein (2009), business organisations is said to benefit virtual social worlds in the areas of advertising/communication, virtual product sales, marketing research human resources, and internal process management. Internet business has resulted in massive increase of data generated globally. One of the challenges of ’big data’ generated through Internet businesses is that of illegal use of customers’ data by business organisations (Bollier and Firestone, 2010). An enormous amount of customers’ data is collected by business organisation continually. Some of these data sometimes end up in wrong hands breaching the confidentiality and privacy such data are meant to portray. Big throughput and big analytics are challenges related with storage and deployment of big data using limited resources and transforming ’big data’ to useful knowledge.

The issue of filtering, developing, disseminating and administering ’big data’ and automatically generating the appropriate meta-data of data has also been identified as an important challenge when handling ‘big data’ (Labrinidis and Jagadish, 2012). However, business intelligence and analytics (BI&A) has helped business organisa-

tions to analyse complex business data for decision-making. Even though 'big data' has many challenges, organisations like IBM still have their success stories about effective handling of their ever-increasing databases. Some of the opportunities of analysing 'big data' is to improve performance and to support human decision making with computational concept. Just as natural plants grow when they get the necessary nutrients, so do data grow with all the technology available today serving as nutrients to data growth. 'Big data' is measured using different parameters. Data can be described as 'big data' by measuring its volume and frequency of accumulation in relation to its lifespan while maintaining the relevance. The number of times a data is cited in academia can be used to measure its relevance and such citing increases the chances of the data being reused, thus adding to the ever-growing 'big data' (Lynch, 2008). It can be observed that 'big data' has succeeded in transforming the way knowledge is perceived, analysed, used, stored and transferred. Most data that would have been lost in the past due to lack of space and expertise to preserve it are now accessible for as long as they remain relevant. The term 'big data' is now common all over the globe and the necessity of protecting data (big or small) has become important. Sensitive organisational and government data are guarded with high-tech data protection devices in order to preserve their relevance. The opportunities of 'big data out-weigh its challenges where adequate data analytics is employed in 'big data' handling.

2.2 The Social Media and 'Big Data'

Social Media (SM) is a group of Internet-based applications that has improved on the concept and technology of Web 2.0, enabling the creation and interchange of User Generated Content (Kaplan, 2012). SM can simply be defined as the media used to be social (Safko, 2010). SM sites have no doubt become big data generators, day-by-day Internet users visit different social media sites either to post information or to retrieve information. Oftentimes SM users post contents pertaining to their personal lives (on Facebook and Twitter) some post pictures (on Instagram) while

Fig. 1 Common Social Media

others post videos (on YouTube, Facebook and Twitter). SM users also create personal blogs where they post real issues for discussion with other users or with the public. Different entities read personal blogs of experts in certain fields or blogs of those with large number of followers; bloggers with large number of followers have large audience and whatever they write on the blogs often receives wide readability. SM sites are real-time 'big data' generators and data generated on these sites contributes immensely to the rapid growth of global databases. Different entities including celebrities, business organisations, schools and governments bodies rely on the social media as one of the means of communicating with their audience. The high rate of acquisition and the use of personal computers and other sophisticated smart mobile devices by people around the world has aided SM sites to stream large-scale data. Telecommunications, electronic mail and electronic messengers like Skype, Yahoo messenger, Google Talk and MSN Messenger are also considered as SM (Aggarwal, 2011). Local events on social media hardly remain local nowadays, as users are quick to post interesting events on the Internet turning local events into issues for global discussions. The majority of mobile phone users in the world today use their phones to connect to the Internet more than using it for the primary purpose of making and receiving calls/text messages. Many retail stores now include online stores to their chains and encourage buyers to leave reviews and/or 'Likes' on products/services they have experienced on popular social media sites, thereby adding to the size of data generated online. SM has contributed significantly to the success stories of many popular big businesses (Kaplan, 2012) in so many ways. SM has also endowed customers with unimaginable power of participation in businesses they have stake in (Evans, 2010). More people are relying on information given by strangers on SM to decide on products/services to buy, film to watch at the cinema or school to enrol in (Pang and Lee, 2008), thus eradicating the possibility of many people making the same mistake for lack of information (Qualman, 2012). Reviewing products and services online is also a means of compelling businesses to improve on their products and services since high percentage of patronage is often derived based on reviews of other customers. Big businesses devote time to filter big data generated from SM sites to make valuable decisions. Research on extracting top quality information from social media (Agichtein et al., 2008; Liu et al., 2009) and on presenting Twitter events content has attracted more attention in recent times (Evans, 2010). The work of Becker et al. (2012) used events features to develop query design approaches for retrieving content associated with an event on different social media sites and discovered ways of using event content detected on one social media to retrieve more information on other social media sites. On the other hand, the work of Kaplan (2012) outlined how firms can utilise mobile social media for marketing research, communication, sales promotions/discounts, and relationship development/loyalty programs. They presented four pieces of advice for mobile social media usage, which they term the 'Four I's of mobile social media.

2.2.1 Analysing 'Big Data' on Twitter

Since Twitter generates enormous amount of data on a daily basis, it is pertinent to design computational ways of mining and analysing Twitter data (tweets) for meaningful use. Interestingly, research on Twitter network is becoming very popular, authors and researchers develop different methods of mining Twitter streaming data. Twitter users tweet for different reasons ranging from expression of personal mode to information dissemination (Jansen et al., 2009), real life events reporting, opinion/sentimental expression to posting breaking news in real time (Adedoyin-Olowe et al., 2013; Agarwal et al., 2012; Chakrabarti and Punera, 2011; Kwak et al., 2010; Weng and Lee, 2011). Topics and events detection and tracking is currently attracting high level of attention as a handful of research and experiments are being conducted on how stories and events are detected and how they evolve over time on Twitter (Adedoyin-Olowe et al., 2013; Gomes et al., 2013; Mathioudakis and Koudas, 2010; Okazaki and Matsuo, 2011; Osborne et al., 2012; Petrović et al., 2010; Phuvipadawat and Murata, 2010; Popescu and Pennacchiotti, 2010). The authors of Popescu and Pennacchiotti (2010) presented methods for discovering a definite type of controversial events that triggers public discussion on Twitter. The work in Becker et al. (2011) used topically related message clusters to spot real-world events and non-events messages on Twitter. On the other hand Weng and Lee (2011) used EDCoW (Event Detection with Clustering of Wavelet-based Signals) to cluster words to form events with a modularity-based graph partitioning method. The experiments in Osborne et al. (2012) was conducted to determine whether Wikipedia, when observed as a stream of page views, could be utilised to improve the quality of discovered events in Twitter. The results of the study confirmed the timeliness of Twitter in detecting events compared to Wikipedia. In Chakrabarti and Punera (2011) SUMMHMM was employed to summarise the 'big data' of real time events such as sporting events posted on Twitter network. The authors of Cataldi et al. (2010) used Page Rank algorithm to retrieve tweets of users with authority on the network. The retrieved tweets were used to create a navigable topic graph that connects the emerging topics under user-defined time slot. The experiments in Adedoyin-Olowe et al. (2013) utilized a combination of Association Rule Mining and a novel methodology called *TRCM* (Transaction-based Rule Change Mining) to identify four Association Rules *(ARs)* patterns in tweets' hashtags at two consecutive time periods. The identified *ARs* were linked to real life events and news reports to show the dynamics of the tweets. Furthermore Gomes et al. (2013) built on *TRCM* by using Transaction-based Rule Change Mining-Rule Type Identification *(TRCM-RTI)* to analyse rule trend of *ARs* present tweets. They employ the approach of Time Frame Window *(TFW)* to measure the evolvements of rules and to calculate the life span of rules on Twitter network.

This chapter progresses to use Rule Type Identification-Mapping *(RTI-Mapping)* to map evolving *ARs* in tweets with evolving news reports and news updates of well-known traditional news agents.

3 Background of Twitter Network

...There may not be newsagents around at a scene of event but there will always be tweeters on ground to broadcast the event live on Twitter even before professional newsagents arrive at the scene...

Twitter has become an acclaimed microblog for information dissemination since its launch in 2006 (Pak and Paroubek, 2010). This can be attributed to the increase in the acceptability of Twitter over the years. The network allows the effective collection of 'big data' with a record of about 500,000 tweets per day as of October 2012. The number of its registered users is estimated at 500 million, while the number of regular users stands at 200 million. It is not necessary to tweet to be on Twitter network as it is reported that 40% of its users simply use the network to track tweets that are of interest to them. According to Twitter, the network reached the billionth tweet in 3 years, 2 months and 1 day. Tweets per second (TPS) record reached 456 when Michael Jackson (a popular American pop star) died on June 25, 2009. The current and most recent TPS record was set in January 2013, when the citizens of the Japan tweeted 6,939 tweets per second on New Year's Day. Twitter data can be referred to as news in real time, the network reports useful information from different perspectives for better understanding. Tweets posted online include news, major events and topics that could be of local, national or global interest. Different events/occurrences are tweeted in real time all around the world making the network generate data rapidly. Individuals, organisations, and even government bodies follow activities on the network in order to obtain knowledge on how their audience reacts to tweets that affect them. Twitter users follow other users on the network and are able to read tweets posted by their following users and also make their contribution to the topic tweeted. The enormous volume of Twitter data has given rise to major computational challenges that sometimes result in the loss of useful information embedded in tweets. Apparently, more and more people are relying on Twitter for information, products/services reviews as well as news on diverse topics. Twitter has been tagged a strong medium for opinion expression and information dissemination on diverse issues. It is also a remarkable source for breaking news and celebrities' gossips. The death of Whitney Houston (a popular American pop star) in February 2012 was reported on Twitter about 27 minutes before the press. Other news and events are often made public faster by Twitter users due to the fact that Twitter users do not need more than a mobile device connected to the Internet to report news to hundreds of millions subscribers on the network. Tweets posted on the network are re-tweeted in order to accelerate the speed of spreading the news. More importantly tweets posted by users known as influencers on the networks are given more attention because such tweets are believed to have more credibility (influencers are users on the network that have large numbers of followers due to the quality of tweets they post on Twitter). Even though some messages carried on Twitter are sometimes incorrect, other Twitter subscribers are often quick to correct any wrong information aired on the network. Many organisations, public figures, educational institutions and government officials set up Twitter accounts in order to reach out to their

audience. Presidents of countries of the world also tweet for many reasons, one of which is to communicate with their subjects on an informal platform.

Considering the enormous volume of tweets generated continuously on a daily basis, users have invented a common way of labelling tweets. This is often attained by including a number of hashtags (#) as prefix to keywords in tweets to describe the tweets' contents. The use of hashtags makes it easy to search for and read tweets of interest. The network reports useful information from different perspectives for better understanding (Zhao and Rosson, 2009). 'Big data' generated on Twitter can be analysed using *ARM* to extract and present interesting hashtags, which can be applied to real life news evolvement.

3.1 Twitter as Decision Support Tool

Opinion of other people is always important to most individuals during the process of decision making. Twitter is often used as information and opinion retrieval network. It also serves as a medium of advertisement for products and services as well as for recommending films, schools or even candidates to vote for in an election. During the US election in 2009, Twitter played a very vital role in the campaign for Barack Obama as presidential candidate. Supporters used Twitter to campaign for his candidature, this can be said to be instrumental to Barack Obama's success in the US national election in 2009. Since Twitter can be categorized as a decision support tool, it is very important to filter tweets to present their interestingness. Using *ARM* to mine tweets' hashtags is an effective way of revealing Association Rules *(ARs)* present in tweets and an effective tool for retrieving useful information embedded in Twitter 'big data'.

4 Overview of Association Rules Mining

Association Rule Mining (ARM) finds frequent patterns, associations, connections, or underlying structures among sets of items or objects in transactional databases, relational databases and other information repositories. Rakesh Agrawal introduced the technique for determining association between items in large-scale transaction data recorded by point-of-sale (POS) structure in superstores. A rule $milk, bread => sugar$ indicates that a transaction including milk and bread together will likely result in the purchase of bread. This information can be used for decision support regarding marketing activities such as pricing, products shelf arrangement and product promotions. Association rules tend to reveal every probable association that satisfies definite boundaries using the lowest support and confidence. *Apriori* method of Association Rules (ARs) is a commonly used algorithm in data mining. It reveals frequent itemsets with minimum support that can be used to establish *ARs* that highlight common trends in the database. A subset of a frequent itemset must also be a frequent itemset. For instance, if *milk*, *bread* is a frequent itemset, both *milk* and *bread* should be frequent itemsets as well. *Apriori* iteratively find frequent itemsets with cardinality from *1 to k (k-itemset)*. It uses the frequent itemsets to generate

association rules. *ARM* technique uncovers different patterns in both transactional and relational datasets. Table. 1 presents a matrix of hashtags extraction in tweets.

Table 1 Hashtags extraction in Tweet

Tweet 1	#datamining	#bigdata	#sql	#KDD
Tweet 2	#ecommerce	#ISMB	#datamining	
Tweet 3	#bigdata	#facebook	#data mining	#analytics
Tweet 4	#analytics	#privacy	#datamining	
Tweet 5	#datamining	#KDD	#bigdata	

4.1 Association Rules in Tweets

Association Rule Mining (ARM) is used to analyse tweets on the same topic over consecutive time periods t and $t+1$ [1]. Rules present at the two periods are matched using Rule Matching *(RM)* to detect four rules patterns present in tweets' hashtags. Rules detected are namely; 'emerging rule', 'unexpected consequent' and 'unexpected conditional rule' ,'new rule' and 'dead rule'. These rules were obtained by setting a user-defined Rule Matching Threshold *(RMT)* to match rules in tweets at time t with those in tweets at $t+1$ in order to determine rules that fall into the different patterns. Transaction-based Rule Change Mining *(TRCM)* was proposed to show rules evolvements in tweets at different points in time. All the rules are linked to real life occurrences such as events and news reports. Rule Type Identification (a technique based on TRCM) was proposed to discover the rule trend of tweets' hashtags over a consecutive period. Rule trend demonstrates how rules patterns evolve into different Time Frame Windows *(TFWs)*. TFWs on the other hand is used to calculate the lifespan of specific hashtags on Twitter and how the lifespan of the hashtags on Twitter network is related to evolvements of news and events pertaining to the hashtags in reality.

4.2 Rule Similarities and Differences

Rule similarities and differences were calculated in Liu et al. (2009) and Song and Kim (2001) to discover association rules in relational datasets. Their methods were adopted in calculating similarities and differences in *ARs* present in tweets hashtags. Similarity is defined using the concepts of degree of similarity proposed in Liu et al. (2009) and Song and Kim (2001). The calculations and notations used are described as follows:

Table 2 Notation for Rule Similarities

n	Number of hashtags
i	an association rule in dataset 1 presented in binary vector
j	an association rule in dataset 2 presented in binary vector
lh_i/lh_j	number of hashtags with value 1 in conditional part of rule i/j
rh_i/rh_j	number of hashtags with value 1 in consequent part of rule i/j
lh_{ij}/rh_{ij}	number of same hashtags in conditional/consequent part of rules i and j
p_{ij}/q_{ij}	degree of similarity of features in conditional/consequent part of rules i and j
r_j^t	Rule present at time t
r_j^{t+1}	Rule present at time $t+1$

4.3 Measuring Similarity

$$p_{ij} = \frac{lh_{ij}}{max(lh_i, lh_j)} \quad (1), \qquad q_{ij} = \frac{rh_{ij}}{max(rh_i, rh_j)} \quad (2)$$

The left hand side (*lhs*)/*conditional* and the right hand side (*rhs*)/*consequent* parts of rules in Apriori principle is used to analyse hashtags as conveyed in tweets over a defined period of time. The co-occurrence of frequent hashtags is used to detect Association Rules *(ARs)* present in tweets at different periods of time. The similarities and differences in the *ARs* discovered in tweets at time t and time $t+1$ are measured in order to categorize them under a rule pattern (for example emerging rule). Changes in the rules dynamics patterns is generated using the *Apriori* principle. Emerging rule detection such as breaking news of a disaster like Typhoon in the Philippines can trigger an instantaneous action from disaster emergency response organisations. It can also serve as a decision support tool for such organisations on how to prioritize their services. News of disaster often generates emerging rule in tweets at an early stage, and can be termed a speedy rule emergence. The emergence of this rule can result in wide broadcast by news agencies globally.

5 Evolution of TRCM

TRCM is a framework set up to define rule change patterns in tweets at different periods. The application of *Apriori algorithm* of ARM to hashtags in tweets at t and at $t + 1$ generates two association rulesets. In Adedoyin-Olowe et al. (2013) Transaction based Rule Change Mining *(TRCM)* was used to detect four (temporal) dynamic rules in tweets. The four rules identified are namely 'new' rules, 'unexpected' rules, 'emerging' rules and 'dead' rules. The rules were obtained by matching rules present in tweets at two periods, t and $t+1$. Binary vectors 0 and 1 were used as the Rule Matching Threshold (RMT) with 0 indicating rule dissimilarity and 1 indicating rule similarity. The Degree of similarity and difference measures are developed to detect the degree of change in rules as presented in Fig.2. The changes are categorized accordingly under the four identified rules. TRCM revealed the dynamics of ARs present in tweets and demonstrates the linkage between the

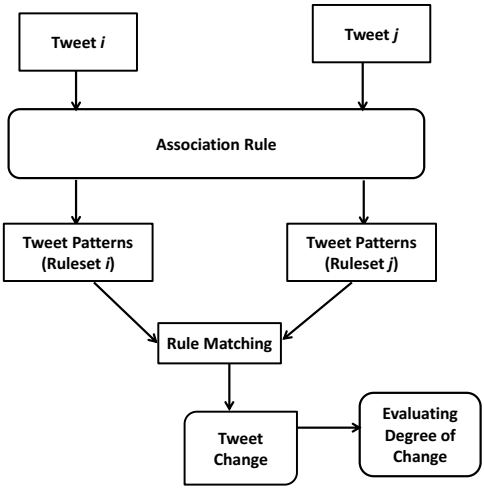

Fig. 2 The Process of Tweet Change Discovery

different types of rule dynamics investigated. Experimental investigations conducted in Adedoyin-Olowe et al. (2013) shows that rules present in tweets hashtags evolve over time giving rise to what is known as **rule trend**.

5.1 Definitions of TRCM Rules

Unexpected Consequent Change in Tweet $(p_{ij} \geq thp_{ij}$ *and* $q_i j < thq_{ij})$. This is when two rules in r_j^t and r_j^{t+1} have similar conditional parts but different consequent parts measurements greater than 0 (for example 0.10).
Unexpected Conditional Change in Tweet ($p_{ij} < thp_{ij}$ and $q_{ij} \geq thq_{ij}$). This change occurs when the consequent parts of rules r_j^t at and r_j^{t+1} are similar, but the conditional parts are different. If the absolute difference measure is less than 0, then the consequent part is similar and the conditional part is different. On the other hand, if the absolute value of the difference measure is greater than 0, then an unexpected conditional change in is said to have occurred.

Emerging Change in Tweet ($p_{ij} \geq thp_{ij}$ and $q_{ij} > thq_{ij}$). 'Emerging' rule is discovered when two hashtags at time t and $t+1$ have similar conditional and consequent part. The similarity measure in this case must be greater than the user-defined threshold.

New Rules. All rules are 'new' until a matching rule is discovered. Every hashtag at time $t+1$ is completely different from all the hashtags in time t ($p_{ij} < thp_{ij}$ and $q_{ij} < thq_{ij}$). However, this changes when a matching is discovered in any parts of t.
'Dead' Rules. 'Dead' rule occurrence is the opposite of new rule detection. A rule in t is labelled 'dead' if its maximum similarity measure with all the rules in $t+1$ is

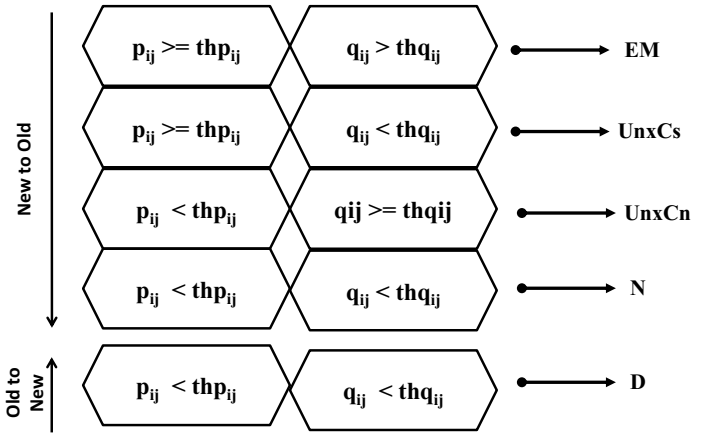

Fig. 3 Transaction-based Rule Change Mining (TRCM)

less than the user-defined threshold ($p_{ij} < thp_{ij}$ and $q_{ij} < thq_{ij}$). The four defined rules detected in tweets hashtags (as presented in Fig.3) are used for analysing trend of *ARs* present in tweets.

6 Analysing Tweets Trend Using *TRCM-RTI*

Trend Analysis *(TA)* of Twitter data is a way of analysing the progression of rules present in tweets hashtags over a temporal period of time. The ultimate goal of *TA* is to be able to trace back the origin of a rule (rule trace). A rule $X => Y$ may have actually started up as $A => B$ and over time the rule has evolved unexpectedly. The time frame between $X => Y$ and $A => B$ may vary depending on factors that may affect the rule status at different point in time. Time Frame Window *(TFW)* describes the different evolvement chains/sequences rule statuses discovered in tweets hashtags is measured in throughout its lifespan on the network. Factors effecting time frame of rules include unexpected discovery relating to an on-going event. Such discovery may elongate or truncate the lifespan of a hashtag. In Gomes et al. (2013) TRCM-RTI was applied to learn the rule trend of tweets' hashtags over a sequential period. Time Frame Windows *(TFWs)* were created to show the different rule evolvement patterns which can be applied to evolvements of news and events in reality. TFWs were employed to calculate the lifespan of specific hashtags on Twitter and to link the hashtags to lifespan of related occurrences in real life. Using the experimental study results, it demonstrated that the lifespan of tweets' hashtags could be linked to evolvements of news and events in reality.

6.1 Rule Trend Analysis

Rule trend analysis demonstrates the chain/sequence a rule is measured in during its period of evolvement on Twitter network. A rule may evolve by taking on a different status in every Time Frame Window *(TFW)*, while another rule may retain a status for more than one TFW in a consecutive evolving period. In some other trend, a rule may evolve back to assume its former status during the course of evolvements. While most rules end up being dead, some may not; such rules will still be present on Twitter network even though they may become static. Evolving rules are synonymous to updates on trending news topics and the pattern of evolving rules could be linked to news updates contents in real life scenario. Different evolvements of rule patterns of *TA* in tweets is demonstrated in the formalisation in Table 3.

Table 3 Evolving Rules Patterns

T	The total time period intervals a rule status is measured in.
C_t	The category of the rule
C_tN	New rule
$C_tU_t^i$	Unexpected conditional rule
C_t^jt	Unexpected consequent rule
C_tE	Emerging rule
C_tD	Dead rule
TFW	Number of frame window

6.2 Time Frame Window of Evolving Rules in Tweet

Time frame plays an important role in Trend Analysis of tweets. While a rule may take a short period to evolve from a new rule to an unexpected rule, another rule may take a longer time to evolve from one rule pattern to another. On the other hand, a rule may become 'dead' straight after becoming a new rule. Such a rule would present a single *TFW* window (new - 'dead'). As time frame window is important to trend analysis of tweets, so it is to news updates in real life situations. In Fig. 4 Sequence A shows how a rule evolved over a time frame period of 47 days and in 4 TFWs before it became a 'dead' rule. The rule has a time frame sequence of C_tN, $C_tU_t^i$, C_t^jt, C_tE, C_tD. In sequence B, the rule did not go into the 'dead' rule status. It went back to assume the C_tE status for the second time and evolved over time frame period of 112 days and having 3 TFW. Lastly, sequence C shows that the rule evolved only once from C_tN to C_tE, then to C_tD after a time frame period of 121 days and in 2 TFWs. All the sequences in Fig.4 explains how events/occurrences in reality affect the dynamics of rules. It also shows the importance of some events in real life when we consider their sequence of evolvements and how long they retain some statuses. Understanding the *TA* of rule evolvements in tweets hashtags enable different entities to understand tweets better and make advantageous use of its contents, either as decision support tool or for information retrieval.

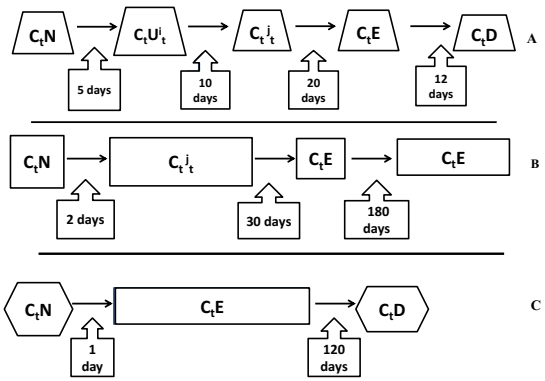

Fig. 4 Chains/Sequences of Evolving Rules

6.3 Which Comes First, the News or the Hashtag? - The "TwO - NwO" State

Twitter users oftentimes tweet about events in real time or disseminate on-the-spot information on the network. Such tweets may trigger the broadcast of the events/information by newsagents as presented in Fig.5. An example of such instance is the news of the death of Whitney Houston mentioned earlier in the chapter.

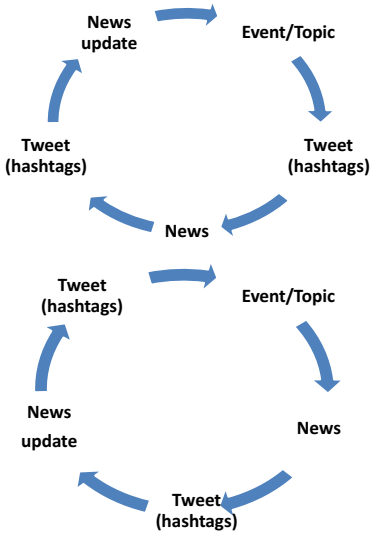

Fig. 5 Which Comes First?

In this case the tweet comes before the news - **tweet-originated**. On the other hand, event or topic broad-casted by newsagents may result in Twitter users hashtagging keywords in the news while expressing their opinion/sentiment on the topic via Twitter network. The topic can be referred to as **news-originated**. The opinion/sentiment expressed on a topic (tweets-originated or news originated) may go on to result in chains of news updates to earlier news reports. This is termed *"TwO - NwO"* state.

7 Empirical Evaluation

Twitter users post tweets of diverse topics ranging from personal daily activities, information on local and national events to breaking news. Twitter streams its data at a fast rate making the network a 'big data' generator. Analysing Twitter data in order to retrieve useful contents has become a concern for many stakeholders including big business organisations and traditional newsagents. Twitter has become an unofficial news media for some individuals and other entities. Different entities sometimes rely on tweets posted online during their process of decision-making. It is however worthy of note that not all tweets carry tangible information, thus the necessity to filter those tweets that embeds the much-needed information for analysis and use. Apart from intangible tweets that constitute noise on the network, inappropriate ways of hashtagging tweets is also responsible for the loss of vital information embedded in tweets.

It is also observed that users replace hashtags used in particular tweets based on the evolvement of the topic being tweeted. Tweet about a local event may trigger a global discussion thus resulting in hashtags replacements. An example of this is the news of 'baby59' found in a sewage pipe in China on May 25, 2013. The news started locally, and later became global news. Hashtags used by Twitter users when the incident unfolded changes as the news relating to the incident continue to evolve. However, this is not always the case as Twitter users are sometimes not quick to update hashtags used in tweets, even when the content of their tweets changes. Also re-tweeting the same tweet repeatedly can result in TRCM-RTI detecting the same set of hashtags more than once (but in the same ARs pattern).

Considering these challenges RTI-*Mapping* goal is to: 1) Extract tweets hashtags that co-occur with the user-specified keyword and also satisfied the minimum support and confidence. 2) Apply *Apriori* of ARM on sets of hashtags at two different time periods to identify ARs patterns present in the tweets using rule matching method. 3) Map evolving *ARs* with relating evolving news in real life. 4) Track updates of evolving news as broad-casted by traditional newsagents. The results of the experiments will offer an autonomic way of retrieving evolving news and news update, allows Twitter users to detect and track evolving news, and updates without having to visit the website of the traditional newsagents targeted in the experiments. Events in real life are often on-going in nature and it requires newsagents to broadcast timely updates of such news. Newsagents update news of events/occurrences in order to keep their audience abreast of how events unfold. RTI-*Mapping* will offer

users the opportunity to be up-to-date with the news updates in a quick, effective and efficient way.

7.1 Experimental Setup

Data: In carrying out the preliminary experiments a small sample dataset of tweets is used. Given a topic #Shutdown, the algorithm first extracts all hashtags co-occurring with #Shutdown from tweets within a specified time t and at $t+1$:
$tweets0 <-$ searchTwitter('#Shutdown', since= '2013-07-15', until='2013-07-16', ...")
$tweets1 <-$ searchTwitter('#Shutdown', since= '2013-07-17', until='2013-07-18', ...")
Next *ARs* are obtained at both period $(r0, r1)$ using *apriori* of ARM and setting user-defined support and confidence parameters to; supp = 0.01, conf = 0.05, target = "rules" for rules present at t and at $t+1$. Support and confident parameters are set so that strong *ARs* are extracted and at the same time interested rules are not lost due to high support and confident setting. The setting increases the performance accuracy of RTI-*Mapping*. Hashtags in both left hand side and the right hand side of rules are matched *(lhsT and rhsT)* in order to detect *TRCM* at time $t+1$ using Rule Matching threshold (RMT). Where rule similarity is detected in the *LHS* (left hand side/conditional part) of $r1$ *(sim_lhs >= lhsT)*, then unexpected consequent rule evolvement is said to have occurred. However if similarity is detected in the *RHS*(right hand side/consequent part) of $r1$ *(sim_rhs >= rhsT)*, then unexpected conditional rule evolvement has occurred. Note that unexpected consequent and unexpected conditional rules evolvements are the same in real life situation and therefore treated as one evolving rule pattern. On the other hand, emerging rules evolvements occur when there is similarity in both lhs and rhs of rules *(sim_lhs >= lhsT & sim_rhs >= rhsT)*.

Annotation: Annotation for mapping the evolving *ARs* in tweets with evolving news and news updates in reality is done manually. Similar to Becker et al. (2011), evolving rules in $t+1$ are clustered and classified into unexpected consequent, unexpected conditional and emerging rules. RTI-*Mapping* is then applied to all the hashtags present in the clusters of the evolving *ARs*. For each evolving rule detected in each case study under review, the combination of all the hashtag keywords in the detected *ARs* are used as search terms in the online news databases of the chosen newsagents. RTI-*Mapping* revealed hashtags within the emerging rules clusters are mostly breaking news, this describes the interestingness of such hashtags. The four case studies in this section are used to validate the experiments.

Emerging rules and unexpected rules are considered in this experiment because they are evolving *ARs* that best describe the dynamics of real life news and news updates. As defined earlier in the chapter all rules in $t+1$ are news until a matching is found in tweets at t. For this reason, new rule are not clustered in the experiments.

7.2 Experimental Case Studies

Case Study 1 - #*Woolwich* - A British soldier murdered in Woolwich, South-East of London on 22nd May, 2013 made a global news while Twitter users include different # in tweets relating to the incident. #*Woolwich* is used as keyword for the experiment. TRCM- RTI experiment conducted within the first 7 days of the incident on #*Woolwich* revealed a number of emerging and unexpected rules. #*EDL* =>#Woolwich and #*Benghazi* =>#Woolwich evolved as emerging and unexpected rules respectively. The rules are used as keywords to search the online version of *'The Telegraph'* newspaper. The ARs detected within 22nd May 2013 and 25th May 2013 are mapped to news headlines and news updates of *'The Telegraph'* newspaper within the 23rd May 2013 and 30th June 2013 as presented in Fig.6.

7.2.1 Analysis of Case Study 1

#*EDL* =>#Woolwich evolved as emerging rule within 22nd May 2013 and 25th May 2013, the rule was mapped to news headlines in *'The Telegraph'* newspaper during and after the period the rule evolved on Twitter network. The first news item mapped to #*EDL* =>#Woolwich was dated 23rd May 2013 at 12:53 am BST with headlines **"Retaliations and Demonstrations follow Woolwich Murder"**. RTI-*Mapping* tracked an update of the news on 3rd June 2013 with caption **"Woolwich and the dark underbelly of British Islam"**. By the 29th June 2013 another update relating to EDL demonstration was tracked with caption **"EDL Leaders Arrested during March to Woolwich"**.

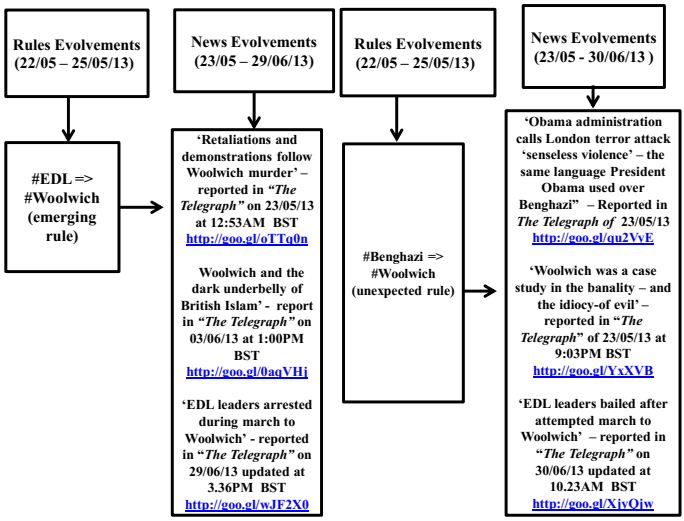

Fig. 6 Rule Mapping 1

On the other hand, *#Benghazi =>*#Woolwich evolved as unexpected rule on Twitter during the period of the experiment. Benghazi and Woolwich were used as search terms and mapped to news headlines of *The Telegraph*. The first news relating to *#Benghazi =>*#Woolwich was mapped on 23rd May 2013 with captioned **"Obama Administration Calls London Terror Attack 'Senseless Violence' the Same Language President Obama used over Benghazi"**. On the same day at 9:03pm BST an update in the *The Telegraph* was mapped to the caption**"Woolwich was a Case Study in the Banality' and the Idiocy 'of Evil"**. The last update was mapped to on 30th June 2013 at 16:23AM BST with caption **"EDL leaders bailed after attempted march to Woolwich"**.

7.2.2 Case Study 2 - *#GayMarriage*

Gay marriage debates became very intense in many countries of the world in 2013. Religious bodies, groups and individuals expressed their views on the passing of the bill legalising gay marriage in some countries in Europe and America. While many African countries rebuff legislation of gay marriage, the governments bill covering England and Wales was passed in the House of Commons in 2013. RTI-*Mapping* experiments conducted on *#Gaymarriage* between 1st June and 4th June 2013 revealed a number of evolving *ARs*. *#gayrights =>* *#gaymarriage* and *#politicalparties =>* *#gaymarriage* evolved as unexpected rules, the rules are mapped to *The BBC Politics online News* from 4th June to 27th June 2013 as shown in Fig.7.

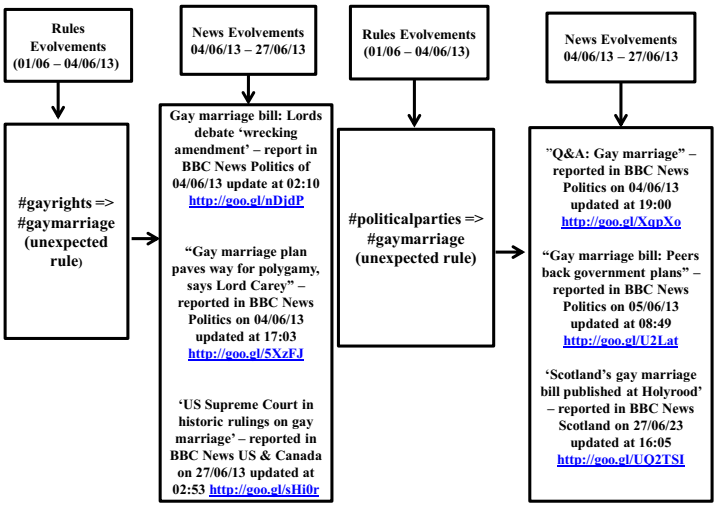

Fig. 7 Rule Mapping 2

7.2.3 Analysis of Case Study 2

On the 4th June 2013, two *BBC News* captioned **"Gay marriage bill: Lord debate wrecking amendment"** and **QA: Gay marriage** were mapped to $\#gayrights => \#gaymarriage$. The former was reported at 2:10am and on the same day at 17:03 another update captioned **"Gay marriage paves was for polygamy, says Lord Carey"** was mapped. Within 24 hours another update captioned **'Gay marriage bill: Peers back government plans'** was mapped.

On 27th June 2013 at 02:53 an update captioned **"US Supreme Court in historic rulings on gay marriage"** reported by *BBC US and Canada news* was also mapped to $\#politicalparties => \#gaymarriage$. Later on the same day *The BBC, Scotland* updated a related news captioned **"Scotland's gay marriage bill published at Holyrood"**.

7.2.4 Case Study 3 - *#Shutdown*

The United States federal government went into a shutdown on 1st October 2013 to 16th October 2013 curtailing most routine operations. The shutdown occurred because of the Congress failing to enact legislation appropriating funds for fiscal year 2014, or a continuing resolution for the interim authorisation of appropriations for fiscal year 2014. RTI-*Mapping* experiments conducted within 1st and 4th October 2013 with *#shutdown* as keyword detected a number of emerging and unexpected evolving rules on Twitter. As expected, all the rules discovered pointed to US

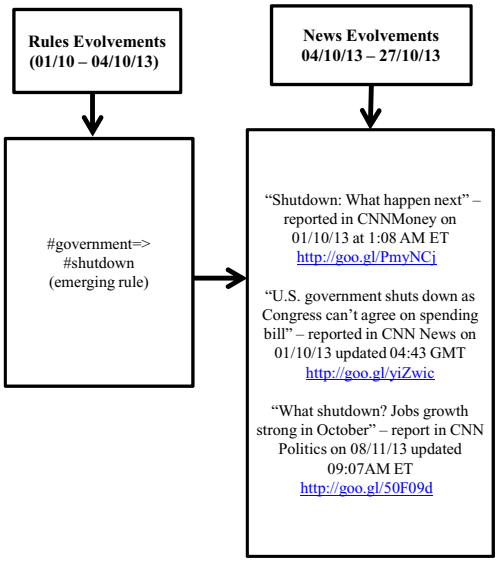

Fig. 8 Rule Mapping 3

government shutdown. *#government => #shutdown* was mapped to a number of *CNN* news reports and updates as presented in Fig.8.

7.2.5 Analysis of Case Study 3

On 1st October 2013, two news items captioned **"Shutdown: What happen next?"** and **"U.S. Government Shuts Down as Congress can't Agree on Spending Bill"** were mapped to *#government => #shutdown*The former was reported on *CNNMoney* less than an hour into the US shutdown while the later was reported on *CNN politics* about 3 hours later. On 8th November 2013 another update on *CNN politics* was mapped to *#government => #shutdown*. *ARs #Shutdown => #PlannedParenthood* evolved unexpectedly and is mapped to *CNN* news captioned **"In Shutdowns, an Attack on Women's Health"** that was reported on 1st October 2013 at 1402 GMT.

7.3 Case Study 4 - #BusinessNews

Business news (unlike the first three case studies) is not based on any event or occurrence. The rules detected are also mapped to business news items on *BBC News*. As shown in Fig.9, some of the news relating to *#BusinessNews* were on going before the experiments commenced. However, *#SprintNextel => #BusinessNews* evolved unexpectedly on Twitter during the experiments, this is attributed to the take-over of Sprint corporation (the biggest intercom corporation in the US) by a Japanese company (Softbank) at the time.

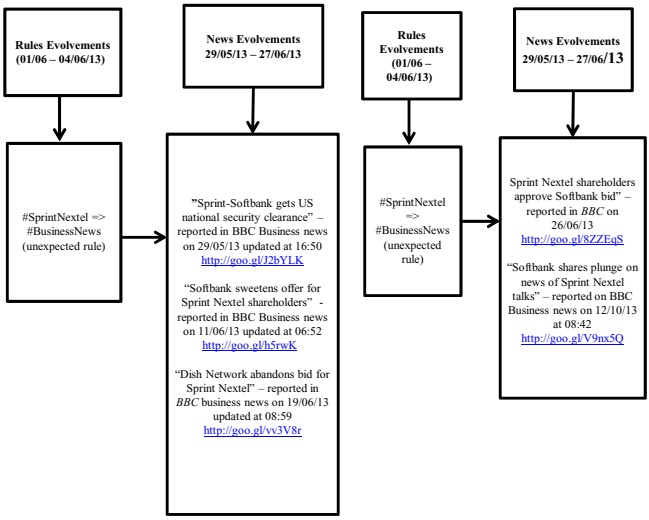

Fig. 9 Rule Mapping 4

7.3.1 Analysis of Case Study 4

On 29th May 2013, *BBC news* reported about *Softbank* getting US national security clearance to buy 70% stake in Sprint Nextel for $20.1bn under the caption, **"Sprint-Softbank gets US national security clearance"**. An update to the story was mapped to $\#SprintNextel => \#BusinessNews$ on 11th June 2013 with caption, **"Softbank Sweetens Offer for Sprint Nextel Shareholders"**. On 19th June 2013 the story evolved with caption **"Dish Network abandons bid for Sprint Nextel"**. Consequently on 26th June 2013 another update was spotted with caption **"Sprint Nextel shareholders approve Softbank Bid"**. The last mapping in the case study was done on 12th October 2013 with caption **"Softbank Shares Plunge on News of Sprint Nextel talks"**. All the news items mapped includes keywords Sprint Nextel and business news detected as *ARs* on Twitter.

8 Conclusion

Hashtag is most commonly used in tweets when compared to its usage on other social media. This chapter uses a novel methodology termed RTI-*Mapping* to map ARs identified in tweets' hashtags at two consecutive time period t and $t+1$ to evolving news reported by three online traditional newsagents. The chapter explained *"TwO - NwO"* stories. *TwO* being news that start from the Twitter network and *NwO* being news that started as reports by newsagents. An example of such news is breaking news tweeted in real time at the scene of occurrence. As explained in Section 6, news triggered by tweets posted online can be tagged tweet-originated topic. However, static rule, evolving news situation can be because of many reasons including hashtags updates by tweeters.

Linking ARs evolvements in tweets hashtags to real life news demonstrates that rule evolvements on Twitter can enhance rapid information retrieval from 'big data' generated on Twitter. It can also enhance timely news updates if hashtags are updated when necessary. Mapping ARs evolvements to news evolvements in reality has been identified as one of the benefits of tweets as regards information retrieval from 'big data'.

8.1 Future Work

The preliminary experiments conducted in this chapter used small sample data sets (tweets) to demonstrate how RTI-*Mapping* technique can map *ARs* present in tweets hashtags to evolving news topics in reality. The technique tracks news updates as they evolve in the news of targeted traditional newsagents. In the experiments, RTI-*Mapping* was able to map all the *ARs* detected in tweets hashtags to news items reported by traditional newsagents in reality. Future experiments will use large datasets (tweets) from *Seeders* and *topsy.com*. Seeders have become an integral part of the Twitter, and will be used for future experiments in order to find newsworthy tweets rapidly. Topsy.com on the other hand has always been a good source for

extracting meaningful tweets from Twitter noise. In 2013, Topsy.com, Topsy Pro, and Topsy APIs began to offer the whole history of public tweets on Twitter, including the first tweet by Jack Dorsey in March 2006. Sourcing tweets from seeders and topsy.com will offer the opportunity of extracting quality news tweets as well as minimize noise characterize by non-event tweets on Twitter network. *TRCM* will be applied to tweets hashtags to detect evolving *ARs* at $t + 1$. Detected evolving rules will be clustered, and keywords in the clusters will then be mapped to evolving news reported by targeted traditional newsagents. The classifier will be trained to detect and track news in an autonomic way. The system will send alert each time there is an update relating to detected corresponding ARs. The alert will state; the headlines topic, the date and time of update, name of newsagents updating news and the number of times the news is updated.

References

Adedoyin-Olowe, M., Gaber, M.M., Stahl, F.: TRCM: A methodology for temporal analysis of evolving concepts in twitter. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2013, Part II. LNCS, vol. 7895, pp. 135–145. Springer, Heidelberg (2013)

Agarwal, P., Vaithiyanathan, R., Sharma, S., Shroff, G.: Catching the long-tail: Extracting local news events from twitter. In: Proceedings of Sixth International AAAI Conference on Weblogs and Social Media, Dublin, June 4-7, pp. 379–382. ICWSM (2012)

Aggarwal, C.C.: An introduction to social network data analytics. Springer (2011)

Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 183–194. ACM (2008)

Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 533–542. ACM (2012)

Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. In: ICWSM, vol. 11, pp. 438–441 (2011)

Bizer, C., Boncz, P., Brodie, M.L., Erling, O.: The meaningful use of big data: four perspectives–four challenges. ACM SIGMOD Record 40(4), 56–60 (2012)

Bloem, J., Van Doorn, M., Duivestein, S., van Ommeren, E.: Creating clarity with big data. Sogeti VINT, Sogeti 3rd edn. (2012)

Bollier, D., Firestone, C.M.: The promise and peril of big data. Aspen Institute, Communications and Society Program, Washington, DC (2010)

Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, p. 4. ACM (2010)

Chakrabarti, D., Punera, K.: Event summarization using tweets. In: Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, July 17-21, pp. 66–73. ICWSM (2011)

Evans, D.: Social media marketing: the next generation of business engagement. John Wiley & Sons (2010)

Gomes, J.B., Adedoyin-Olowe, M., Gaber, M.M., Stahl, F.: Rule type identification using trcm for trend analysis in twitter. In: Research and Development in Intelligent Systems XXX, pp. 273–278. Springer (2013)

Henno, J., Jaakkola, H., Mäkelä, J., Brumen, B.: Will universities and university teachers become extinct in our bright online future? In: 2013 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO), pp. 716–725. IEEE (2013)

Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11), 2169–2188 (2009)

Kaplan, A.M.: If you love something, let it go mobile: Mobile marketing and mobile social media 4x4. Business Horizons 55(2), 129–139 (2012)

Kaplan, A.M., Haenlein, M.: The fairyland of second life: Virtual social worlds and how to use them. Business Horizons 52(6), 563–572 (2009)

Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media?. In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)

Labrinidis, A., Jagadish, H.: Challenges and opportunities with big data. Proceedings of the VLDB Endowment 5(12), 2032–2033 (2012)

Liu, D.-R., Shih, M.-J., Liau, C.-J., Lai, C.-H.: Mining the change of event trends for decision support in environmental scanning. Expert Systems with Applications 36(2), 972–984 (2009)

Lynch, C.: Big data: How do your data grow? Nature 455(7209), 28–29 (2008)

Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp. 1155–1158. ACM (2010)

Okazaki, M., Matsuo, Y.: Semantic twitter: Analyzing tweets for real-time event notification. In: Breslin, J.G., Burg, T.N., Kim, H.-G., Raftery, T., Schmidt, J.-H. (eds.) BlogTalk 2008/2009. LNCS, vol. 6045, pp. 63–74. Springer, Heidelberg (2010)

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First story detection using twitter and wikipedia. In: the Workshop on Time-aware Information Access. TAIA, vol. 12 (2012)

Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of Seventh International Conference on Language Resources and Evaluation, LREC, Malta, May 17-23, pp. 1320–1326. LREC (2010)

Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1–135 (2008)

Parameswaran, M., Whinston, A.B.: Social computing: An overview. Communications of the Association for Information Systems 19(1), 37 (2007)

Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 181–189. Association for Computational Linguistics (2010)

Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in twitter. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 3, pp. 120–123. IEEE (2010)

Popescu, A.-M., Pennacchiotti, M.: Detecting controversial events from twitter. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1873–1876. ACM (2010)

Qualman, E.: Socialnomics: How social media transforms the way we live and do business. John Wiley & Sons (2012)
Safko, L.: The Social media bible: tactics, tools, and strategies for business success. John Wiley & Sons (2010)
Song, H.S., Kim, S.H.: Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications 21(3), 157–168 (2001)
Tang, L., Liu, H.: Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery 2(1), 1–137 (2010)
Weng, J., Lee, B.-S.: Event detection in twitter. In: The Proceedings of the International Conference on Weblogs and Social Media, Barcelona, July 17-21, pp. 401–408. ICWSM (2011)
Zhao, D., Rosson, M.B.: How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, pp. 243–252. ACM (2009)

Hybrid Tolerance Rough Set Based Intelligent Approaches for Social Tagging Systems

H. Hannah Inbarani and S. Selva Kumar

Abstract. The major challenge with Big Data analysis is the generation of huge amounts of data over a short period like Social tagging system. Social Tagging systems such as BibSonomy and del.icio.us have become progressively popular with the widespread use of the internet. The social tagging system is a popular way to annotate web 2.0 resources. Social tagging systems allow users to annotate web resources with free-form tags. Tags are widely used to interpret and classify the web 2.0 resources. Tag clustering is the process of grouping the similar tags into clusters. The tag clustering is very useful for searching and organizing the web2.0 resources and also important for the success of social tagging systems. Clustering the tag data is very tedious since the tag space is very large in several social bookmarking websites. So, instead of clustering the entire tag space of Web 2.0 data, some tags frequent enough in the tag space can be selected for clustering by applying feature selection techniques. The goal of feature selection is to determine a marginal bookmarked URL subset from Web 2.0 data while retaining a suitably high accuracy in representing the original bookmarks. In this chapter, Unsupervised Quick Reduct feature selection algorithm is applied to find a set of most commonly tagged bookmarks and this paper proposes TRS approach hybridized with Meta heuristic clustering algorithms. The proposed approaches are Hybrid TRS and K-Means Clustering (TRS-K-Means), Hybrid TRS and Particle swarm optimization (PSO) K-Means clustering algorithm (TRS-PSO-K-Means), and Hybrid TRS-PSO-K-Means-Genetic Algorithm (TRS-PSO-GA). These intelligent approaches automatically determine the number of clusters. These are in turn compared with K-Means benchmark algorithm for Social Tagging System.

Keywords: Tag Clustering, Bookmark Selection, K-Means, Tolerance Rough Set (TRS), Particle Swarm Optimization (PSO), Genetic Algorithm (GA).

1 Introduction

Web sites have introduced a collection of refined techniques known as Web 2.0. The emergence of Web 2.0 and the consequent success of social network websites such as del.icio.us and Flickr introduce us to a new concept called social tagging

H. Hannah Inbarani · S. Selva Kumar
Department of Computer science, Periyar University, Salem-636011
e-mail: hhinba@gmail.com, info_selva@yahoo.co.in

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_8

system. Since then, different social systems have been built that support tagging of a variety of resources. Given a particular web object or resource, tagging is a process where a user assigns a tag to an object (Gupta et al. 2010). Social tagging is one of the most important forms of user generated content. Tagging provides an easy and intuitive way for users to annotate, organize and refined resources. For this reason, a huge number of systems have added tagging functionality. The Social Tagging System is an application of social media that has succeeded as a means to ease information search and sharing. With the rapid growth of Web 2.0, tagged data is becoming more and more abundant on the social networking websites. The tags are collected from the user epitomize part of this user's favorite or interests in the social bookmarking website (Heymann et al. 2008).

Tagging can be seen as the action of connecting a relevant user-defined keyword to a document, image or video, which helps user to better organize and share their collections of interesting stuff. The tags are collected by the user's favorites or interests in the social bookmarking website. In Social Tagging systems, Tagging can be seen as the act of linking of entities such as users, resources and tags (Caimei et al. 2011). It helps user better way to understand and distribute their collections of attractive objects. When a user applies a tag to a resource in the system, a multilateral relationship between the user, the resource and the tag is formed as shown in Fig.1.

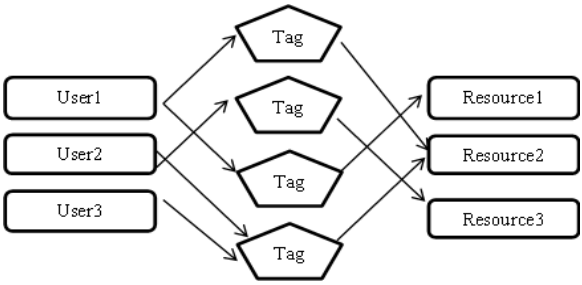

Fig. 1 Multilateral relationship

Massive and complex data are generated every day in many areas. Complex data refer to data sets that are so sizable voluminous that conventional database management and data analysis tools are insufficient to deal with them. Managing and analysis of Social Tagging big data involve many different issues regarding their structure, storage and analysis (Azar and Hassanien 2014). Feature selection also called as Dimensionality reduction is a process of selecting a subset of highly relevant features which is responsible for future analysis (Inbarani et al 2014a). Dimensionality reduction is one popular technique to remove noisy (i.e. irrelevant) and redundant attributes. Dimensionality reduction techniques can be categorized mainly into feature extraction and feature selection. Feature Selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. In feature extraction approach, features are projected into a new space with lower

dimensionality. The feature selection approach aims to select a small subset of features that minimize redundancy and maximize relevance to the target Bookmarks. Both dimensionality reduction approaches are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage. However, feature selection is superior in terms of better readability and interpretability since it maintains the original feature values in the reduced space while feature extraction transforms the data from the original space into a new space with lower dimension, which cannot be linked to the features in the original space. Therefore, further analysis of the new space is problematic since there is no physical meaning of the transformed features obtained from feature extraction technique.

The explosive popularity of social Tagging System produces mountains of high-dimensional data and the nature of social network also determines that its data are often unlabeled, noisy and partial, presenting new challenges to feature selection. The nature of social tagging system also determines that each view is often noisy and incomplete (Jiang et al. 2004). Feature Selection (FS) is an essential part of knowledge discovery. It is a process which attempts to select features which are more informative (Velayutham and Thangavel 2011). FS is separated into the supervised and unsupervised categories. When class labels of the data are available, supervised feature selection is applicable, otherwise the unsupervised feature selection is applicable (Jothi and Inbarani 2012). The feature selection technique is applied for Bookmark Selection (BMS). The goal of Bookmark selection is to find out a marginal bookmarked URL subset from a Web 2.0 data while retaining a suitably high accuracy in representing the original bookmarks (Kumar and Inbarani 2013a). The BMS is a must due to the profusion of noisy, irrelevant or misleading bookmarks added to web 2.0 sites. BMS is also used to increase the clustering accuracy and reduce the computational time of clustering algorithms. Web 2.0 user-generated tagging bookmark data usually contains some irrelevant bookmarks which should be removed before knowledge extraction. In data mining applications, decision class labels are often unknown or incomplete. In this situation the unsupervised feature selection plays a vital role to select features. In this chapter, Unsupervised Quick Reduct (USQR) method is used for selection of tagged bookmarks since there are no class labels for web 2.0 tagged bookmark data.

Clustering is one of the important tasks in data mining. Clustering is an approach to analyze Social Network data at a higher level of abstraction by grouping the data according to its similarity into meaningful clusters. Clustering is one of the general approaches to a descriptive modelling of a large amount of data, allowing the analyst to focus on a higher level representation of the data. Clustering methods analyze and explore a dataset to associate objects into groups, such that the objects in each group have common characteristics. These characteristics may be expressed in different ways: for example, one may describe the objects in a cluster as the population generated by a joint distribution, or as the set of objects that minimize the distances from the centroid of the group (Kisilevich et al. 2010).

Data clustering is a common technique for statistical data analysis. Clustering provides partitioning of a data set into subsets of similar objects or data clusters. Before actually using a clustering technique, the first task one has to do is to transform the problem at hand into a numeric representation that can be used by clustering algorithms (Begelman et al. 2006). Clustering is the process of finding similarities in data and putting similar data into groups (Kumar and Inbarani 2013b). Clustering partitions a dataset into several groups such that the similarity within a group is larger than that among groups (Hammouda 2006). Tag clustering is the process of grouping similar tags into the same cluster and is important for the success of Social Systems. The goal of clustering tags is to find frequently used tags from the tagged bookmarks. On the tag clustering, similar tags are clustered based on tag weights associated with bookmarks (Kumar and Inbarani 2013a). On the tag clustering, the similar tags are clustered based on tag weights associated with bookmarks. In this chapter, Hybrid TRS and K-Means Clustering (TRS-K-Means), Hybrid TRS and Particle swarm optimization (PSO) K-Means clustering algorithm (TRS-PSO-K-Means), and Hybrid TRS-PSO-K-Means-Genetic clustering Algorithm (TRS-PSO-GA) for social systems and the techniques are implemented and tested against a delicious dataset. The performance of these techniques is compared based on 'goodness of clustering' evaluation measures such as Mean Square Quantization Error (MSQE) and Sum of Intra Cluster Distances (SICD).

The proposed work consists of

- **Data Extraction**: fetching data from del.icio.us (www.delicious.com) and the Data sets are converted into matrix representation. "Delicious" (del.icio.us) is a social bookmarking web service for storing, sharing, and determining web bookmarks.
- **Data Formatting**: Data formatting consists of mapping the tags and bookmarks based on tag weights represented in matrix format.
- **Bookmark Selection**: BMS is the progression of selecting more useful tagged bookmarks from a set of bookmarks associated with tags.
- **Tag Clustering**: to cluster relevant tags based on tag weights associated with selected bookmarks

The rest of this Chapter is organized as follows: Section 2 presents some of the related work in web 2.0 tag clustering and feature selection. Section 3 Present Methodology of this research work. In Section 4, the experimental results have been reported. And the conclusion has been addressed in Section 5.

2 Related Work

This section gives a brief review about the Feature Selection and Clustering Techniques. The FS and Clustering Technique plays vital role in the social data analysis. It is widely accepted by the machine learning community that, there are no good mining results without pre-processed data. The performance of the tag

clustering is improved from feature selection (Mangai et al. 2012). FS plays an important role in the data mining process.

FS is needed to deal with the excessive number of features, which can become a computational burden on the learning algorithms. It is also necessary, even when computational resources are not scarce, since it improves the accuracy of the machine learning tasks. Feature selection, as a pre-processing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing predictive accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness (Jothi et al 2013, Azar et al. 2013). The benefits of feature selection are twofold: considerably decrease the running time of the induction algorithm, and increase the accuracy of the resulting model. All feature selection algorithms fall into four categories: (1) the filter approach (2) the wrapper approach (3) Embedded and (4) Hybrid approaches. In the filter approach, feature selection is performed as a pre-processing step to induction. The filter methods are independent of learning algorithms, with good generality. In the wrapper approach, the feature selection is being "wrapped around" an induction algorithm, so that the bias of the operators that define the search and that of the induction algorithm interact. Although the wrapper approach suffers less from feature interaction, running time makes this approach infeasible in practice, especially if there are many features (Hu and Cercone 1999). The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected sub-sets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. Bolón-Canedo (Bolón-Canedo et al. 2012) presents a review of feature selection methods on synthetic data. Hybrid approaches are a technique that finds the feature subset with combination of other approaches. The hybrid approaches give better accuracy compare than other approaches.

The FS Technique is applied for Bookmark Selection (BMS) for Tag Clustering. BMS is a pre-processing step in social tagging systems, and it is very effective in reducing dimensions, reducing the irrelevant data, increasing the learning accuracy and improving comprehensiveness. BMS aims to determine a minimal feature subset from a Social tagging System domain while retaining a suitably high accuracy in representing the original features. Rough set theory has been used as such a tool with much success. RST enables the discovery of data dependencies and the reduction of the number of attributes contained in a dataset using the data alone requiring no additional information (Parthalain and Jensen 2013). Velayutham et al. Have proposed a new unsupervised quick reduct (USQR) algorithm using rough set theory. The USQR algorithm attempts to

calculate a Reduct without exhaustively generating all possible subsets (Velayutham and Thangavel 2011).

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis (Jain et al. 1999). The term "clustering" is used in several research communities to describe methods for grouping of unlabelled data.

Dattolo presents an approach for detecting groups of similar tags and relationships among them. The authors apply clustering processes to find different categories of related tags, presenting three ways of calculating tag weights within a graph: intersection, Jaccard and a more complex approach that considers additional distributional measures of tags in a vector space representation. In the past years, many studies have been carried out on employing clustering in social tagging systems. Dattolo et al had presented an approach for detecting groups of similar tags and relationships among them. The authors find the different categories of related tags by applying a clustering process (Dattolo et al. 2011). In (Xu et al. 2011), Guandong Xu et al. has presented a working example of tag clustering based on similarity and he proposed a clustering technique called Kernel information propagation for tag clustering.

Begelman presented several clustering techniques and provided some results on del.icio.us and Raw-Sugar to prove that clustering can improve the tagging experience (Begelman et al. 2006). Andriy Shepitsen applied hierarchical clustering for tag clustering (Shepitsen et al. 2008). Artificial intelligence based clustering algorithms which compute the number of clusters automatically have been proposed for analyzing web page, click stream patterns by (Inbarani and Thangavel 2009). Marco Luca has used Self-Organizing Map (SOM) to cluster tagged bookmarks (Sbodio and Simpson 2009). Jonathan Gemmell proposes a method to personalize a user's experience within folksonomy using clustering (Gemmell et al. 2008). A personalized view can overcome ambiguity and idiosyncratic tag assignment, presenting users with tags and resources that correspond more closely to their intent. Specifically, we examine unsupervised clustering methods for extracting commonalities between tags, and use the discovered clusters, as intermediaries between a user's profile and resources in order to tailor the results of the search to the user's interests.

Meta heuristics are another diplomatic rational method for problem solving using modern approaches. A development in Meta heuristics research is the investigation of hybridization of Meta heuristic approaches such as GA, PSO and etc. This may also result out in often finding decent solutions with less computational effort than any other methods. Meta heuristic techniques are emerging as viable tools and alternatives to more traditional clustering techniques. Among the many meta-heuristics techniques, clustering with PSO techniques has found success in solving clustering problems. It is suitable for clustering complex and linearly non-separable datasets. Clustering is one of the widely used data mining techniques for Big Data of Social Tagging System (Martens et al. 2011). A large number of meta-heuristic

algorithms have been implemented for clustering the Big Data. Meta heuristic algorithms have some shortcomings such as the slowness of their convergence and their sensitivity to initialize values. The clustering algorithms classify Social Tag data into clusters and the functionally related Tags are grouped together in an efficient manner (Dhanalakshmi and Inbarani 2012).

Ahmadi presented an approach of the Flocking based approach to data clustering (Ahmadi et al. 2010). Ahmed presented a literature survey on PSO algorithm and its variants for clustering high-dimensional data (Esmin et al. 2013). Mehdi Neshat proposed a co-operative clustering algorithm based on PSO and K-means and he also compared that algorithm with PSO, PSO with Contraction Factor (CF-PSO) and K-means algorithms (Neshat et al. 2012). Kuo et al (2011) intended to integrate Particle Swarm Optimization Algorithm (PSOA) with K-means to cluster data and he had shown that PSOA can be employed to find the centroids of a user-specified number of clusters. Modification in PSO and its hybridization with other algorithms give better results in various optimization problems in terms of execution, efficiency, accuracy in comparison with other meta heuristic algorithms such as GA, SA etc. (Rana et al. 2011). Taher Niknam proposed an efficient hybrid approach based on PSO, ACO and K-means for cluster analysis (Taher and Babak 2010). Yau-King Lam proposed PSO-based K-Means clustering with enhanced cluster matching of gene expression data (Yau et al. 2013). R.J. Kuo proposed a hybridization of GA and PSO algorithm for order clustering (Kuo and Lin 2010).

3 Phases of Clustering Social Tagging Data

Clustering Social Tagging System data comprises the following steps.

a) Preprocessing

b) Clustering

c) Pattern Analysis

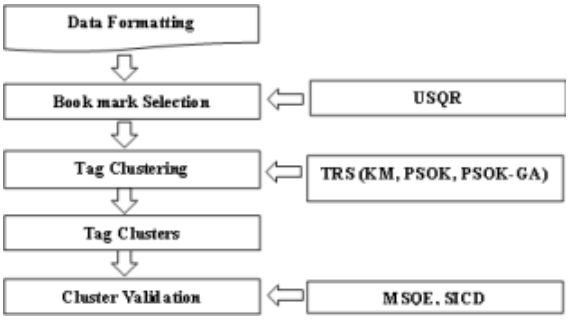

Fig. 2 Methodology

3.1 Data Formatting

The initial step of this work is data extraction and data formatting. The dataset is extracted from delicious which is a free and gentle tool to preserve, organize and discover interesting links on the WWW. Delicious uses a non-hierarchical classification system in which users can tag each of their bookmarks with freely chosen index terms. Delicious is an internet site designed to allow access to any website using social bookmarking.

Social bookmarking is a method of storing, organizing, searching and managing bookmarks of web pages. Instead of saving bookmarks on your computer, you save them on a social bookmarking site like delicious. You can identify your websites by assigning keywords and organize your websites according to your own manner. Tags are used to assign names to the saved bookmarks. Tags are keywords to help you remember what the website is about. You can access your bookmarks from any computer that is linked to the net. The dataset contains tagid, bookmark id and tag weight associated with the bookmarks. Table 1 (a) indicates that the format of extracting dataset from delicious.

Table 1 (a) Format of Tag Dataset from Delicious

S.No	Tag Id	Boomark Id	Tag Weight
1	1	1	214
2	2	4	44
3	2	2	48
4	4	2	85
5	3	1	521

After taking the dataset, convert the tag data set into matrix representation. Table 1 (b) indicates the matrix representation of the tag data set. The Tag data set is normally symbolized by a matrix, with rows corresponding to Tags, and columns corresponding to Social Bookmarking URLs.

In the tag matrix, n- Represents tags

m- Represents bookmarked URLs

Wij - represents tag weight associated with bookmarks

Table 1 (b) Matrix Representation of Tag Dataset

	Bm1	**Bm2**	**Bm3**	**Bm4**
Tag1	214	52	32	14
Tag2	30	48	32	44
Tag3	521	26	47	0
Tag4	14	85	26	33
Tag5	72	54	32	48

3.2 Pre-processing

Bookmark selection is a pre-processing step in data mining, and it is very effective in reducing dimensions, reducing the irrelevant data, increasing the learning accuracy and improving the comprehensiveness (Inbarani et al. 2007). Bookmark selection can be a powerful tool for simplifying or speeding up computations. It can prove the tag clustering efficiency and performance in the ideal case, in which Bookmarks are selected based on class information. Bookmark selection not only reduces the high dimensionality of the Bookmark space, but also provides better data understanding, which improves the tag clustering result (Mitra et al. 2002). The selected bookmark set should contain sufficient or more reliable information about the tag data set. For Tag clustering, this will be formulated into the problem of identifying the most informative frequent tags used by user in bookmarking web pages.

USQR Algorithm

The Rough Set (RS) theory can be used as a tool to reduce the input dimensionality and to deal with vagueness and uncertainty in datasets. Bookmark selection is a pre-processing step in data mining, and it is very effective in reducing dimensions, reducing the irrelevant data, increasing the learning accuracy and improving the comprehensiveness (Inbarani et al. 2014b). Bookmark selection can be a powerful tool for simplifying or speeding up computations. It can prove the tag clustering efficiency and performance in the ideal case, in which Bookmarks are selected based on class information. Bookmark selection not only reduces the high dimensionality of the Bookmark space, but also provides better data understanding, which improves the tag clustering result. The selected bookmark set should contain sufficient or more reliable information about the tag data set. For Tag clustering, this will be formulated into the problem of identifying the most informative frequent tags used by user in bookmarking web pages.

Algorithm 1:USQR Algorithm

Algorithm 1: USQR Algorithm - USQR(C)
C, the set of all conditional features;
(1) R ⟵ { }
(2) Do
(3) T ⟵ R
(4) ∀x ∈ (C – R)
(5) ∀y ∈ C
(6) $\gamma_{RU\{x\}(y)} = \frac{|POS_{RU\{x\}(y)}|}{|U|}$
(7) if $\overline{Y_{RU\{x\}}{}^{(y)}, \forall y| \in c} > \gamma_r{}^{(y)}, \overline{\forall y| \in c}$
(8) T⟵ R U {x}
(9) R ⟵ T
(10) until $\overline{Y_{RU\{x\}}{}^{(y)}, \forall y| \in c} > \gamma_r{}^{(y)}, \overline{\forall y| \in c}$
(11) return R

In many data mining applications, class labels are unknown, and so considering the significance of Unsupervised Feature Selection (UFS). In this work, UFS is applied to book mark selection. Unsupervised Quick Reduct (USQR) (Velayutham and Thangavel 2011) algorithm attempts to compute a bookmark subset without comprehensively generating all possible subsets. According to the algorithm, the mean dependency of each bookmark subset is calculated and the best bookmark subset is selected. The USQR algorithm is presented in Algorithm1.

3.3 Clustering

Clustering Tag data is the process of grouping the similar Tags into the same cluster based on Tag Weight associated with selected bookmarks by using clustering techniques. Tags in the same cluster have been frequently used in the Bookmarks by users. This Chapter reviews three clustering techniques: TRS-K-Means TRS-PSO-K-Means, and TRS-PSO-K-Means-GA. These techniques are implemented and tested against a Delicious dataset.

The input for the tags clustering algorithm consists of:
Tags t_i, i = 1: : : I , Bookmarks b_j , j = 1 : : : J , K – Number of clusters

Objective Function for Clustering

The indeed motto of the clustering algorithm is to moderate the intra-cluster remoteness and also to maximize the inter-cluster distance based on the distance measures, here the objective function is referred as validation measure. To demonstrate it briefly, we have preferred Mean square quantization error (MSQE) and sum of intra cluster distances (SICD) for comparative analysis. MSQE and SICD briefly explained in Experimental Results

TRS Approach

Tolerance Rough Set Model (TRSM) was developed by Ho, T.B, and Nguyen N.B as basis to model tags and Bookmarks in information retrieval, Social Tagging Systems, etc. With its ability to deal with vagueness and fuzziness, TRSM seems to be a promising tool to model relations between tags and bookmarks. The use of Tolerance Space and upper approximation to enrich inter-tag and tag-cluster relation allows the algorithm to discover subtle similarities not detected otherwise (Ho and Nguyen 2002).

Around 1980's, Pawlak presented the basic concepts of rough set and approximation space, i.e. lower/upper approximations (Jianwen and Bagan 2005). The objects in the universe (U) are classified into equivalence classes according to equivalence relation. The objects belonging to the same equivalence class is indiscernible. The set divided by equivalence relation is approximation space.

The set composed of the union of equivalence class contained in any subset X of approximation space is lower approximation while the set composed of the summation of equivalence classes which intersects with X not empty is called upper approximation (De and Krishna 2004). The difference between upper approximation and lower approximation is called a boundary region, which contains the objects that cannot be judged whether it belongs to given class or not. The tolerance rough set of an object x is defined by a set of all objects that has the tolerance relation to the object x with respect to all attributes as $R(t) = \{s \in T, sRt\}$ such a relation is called tolerance or similarity.

The pair (U, T) is called a tolerance space. We call a relation T ⊂ X*U a tolerance relation on U if (i) is reflexive: xTx for any x ∈ U (ii) is symmetric: xTy implies yTx for any pair x; y of elements of U. Sets of the form T(x) = {y∈U: xTy} are called tolerance sets. A binary tolerance relation R is defined on T. The lower approximation of P, denoted by $\underline{R}(P)$ and the upper approximation of P, denoted by $\overline{R}(P)$ are respectively defined as follows:

$$\underline{R}(P) = \{t \in P, R(t) \subseteq P\} \quad 3.1$$

And

$$\overline{R}(P) = U_{t \in P} R(t) \quad 3.2$$

Definition 1: The similarity class of t, denoted by R(t), is the set of Tagged Bookmarks which are similar to t. It is given by

$$R(t) = \{s \in T, sRt\} \quad 3.3$$

Definition 2: A binary tolerance relation R is defined on T. The lower approximation of P, denoted by $\underline{R}(P)$ and the upper approximation of P, denoted by $\overline{R}(P)$ are respectively defined as follows:

$$\underline{R}(P) = \{t \in P, R(t) \subseteq P\} \quad 3.4$$

And

$$\overline{R}(P) = U_{t \in P} R(t) \quad 3.5$$

Let $t_i \in T$ be a Social tag Vector. The upper approximation $\overline{R}(t_i)$ is a set of Tagged Bookmarks similar to t_i, i.e. a Tagged Bookmark in t_i, is more or less similar to other tags in $\overline{R}(t_i)$. This can be called as Similarity Upper Approximation and denoted by S_i.

Algorithm 2: TRS Approach (Intelligent approach)

Input: *Set of N Tagged bookmarks, threshold* $\boldsymbol{\delta}$.
Output*: Number of K- clusters*

Step 1*: Construct similarity matrix between Tagged Bookmarks using eq.3.6*

$$cosine\ similarity = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} {x_i}^2 + \sum_{i=1}^{t} {y_i}^2}} \qquad 3.6$$

Step 2*: Find similarity upper approximation using eq.3.7 for each Tagged Bookmark based on threshold* $\boldsymbol{\delta}$.

$$\bar{R}X = \{x \in U : R(x) \cap X \neq 0\} \qquad 3.7$$

Step 3*: Find centroid for each set of tags' upper approximation*

Step 4*:* Merge similar centroids into one cluster (set) and set the value of K as distinct number of sets

Step 5: *initialize the number of clusters to K and centroids as cluster representatives.*

We provide an example of Social tag data for constructing similarity matrix and finding upper approximation**.**

Initialization:
Let t= {t1, t2, t3, t4} be the set of tags and b= {bm1, bm2, bm3, bm4, bm5} be the set of distinct bookmarks. Let t1 ={bm1,bm2} , t2={bm2,bm3,bm4},t3={bm1,bm3,bm5},t4={bm2,bm3,bm5}.

Table 2 Example Data

	BM1	**BM2**	**BM3**	**BM4**	**BM5**
T1	1	1	0	0	0
T2	0	1	1	1	0
T3	1	0	1	0	1
T4	0	1	1	0	1

Then the tag can be represented as vectors.
t1={1,1,0,0,0},t2={0,1,1,1,0},t3={1,0,1,0,1},t4={0,1,1,0,1}

Step1: Construct Similarity Matrix
Construct cosine Similarity Matrix between all tags, for example the similarity between tag1 and tag2 are given below Table 2 shows the similarity matrix between the tags

$$= cosine\ similarity = \frac{1*0+1*1+0*1+0*1+0*0}{\sqrt{(1^2+1^2+0^2+0^2+0^2)}*\sqrt{(0^2+1^2+1^2+1^2+0^2)}} = \frac{1}{\sqrt{2}*\sqrt{3}} = \frac{1}{1.14*1.73} = \frac{1}{1.97} = 0.50$$

Table 3 Similarity Matrix

	T1	**T2**	**T3**	**T4**
T1	1	0.50	0.50	0.50
T2	0.50	1	0.33	0.66
T3	0.50	0.33	1	0.66
T4	0.50	0.66	0.66	1

Step 2: ***Find upper approximation***
Based on threshold =0.6, the upper approximation for the tags are

u (t1) = {t1}, u(t2) = {t2,t4}, u(t3) = {t3,t4}, u(t4) = {t2,t3,t4}.

Step 3: ***Find centroid for each set of tags' upper approximation***

Table 4 Find Mean Value for the set of all tags

	BM1	**BM2**	**BM3**	**BM4**	**BM5**
T1	**1**	**1**	**0**	**0**	**0**
T2	**0**	**1**	**1**	**1**	**1**
T3	**1**	**1**	**1**	**0**	**1**
T4	**0**	**1**	**1**	**1**	**1**

Step 4: Merge similar centroids into one cluster (set) and set the value of K as distinct number of sets

The next iteration, we get the upper approximation for the tags are

U(t1) =(t1) , U(t2)=(t2,t3,t4) , U(t3)=(t2,t3,t4), U(t4)=(t2,t3,t4)

Then merge similar centroids into one cluster (set) and set the value of K as distinct number of sets

U(t1)=(t1) ,U(t2)=U(t3)=U(t4) = (t2,t3,t4)

Step 5*:* initialize the number of clusters to K and centroids as cluster representatives

For the above example finally we get two sets. u(t1)={t1},u(t2)={t2,t3,t4}, then K is assigned 2 and the mean value is assigned to cluster centroid.

K-Means Clustering

The important feature of the clustering algorithm is to measure the similarity which is used for determining how close two patterns are to one another. The K-Means algorithm is one of the most widely used clustering technique (Moftah et.al

2013). The K-Means algorithm is very simple and can be easily implemented in solving many problems. The K-Means algorithm begins with K clusters; each cluster contains randomly selected cluster centroid. Place the tag in the cluster identified with this nearest centroid. After assigning each tag to one of the clusters, compute the centroid of the altered cluster (Grbovic et al. 2013).

Algorithm 3: TRS-K-Means Clustering Algorithm

Input: *Set of N Tagged bookmarks, K -number of clusters,*
Output*: K overlapping Tag clusters*

Step 1: *Initialize Number of clusters and its centroids using TRS*
Step 2: *Assign each tagged bookmarks X_i to the nearest cluster center Z_j, Where I = 1, 2,…, m and j =1,2,…,K, If and only if*

$$||x_i - z_j|| < ||x_j - z_p||,\ p=1, 2,\ldots K \text{ and } j \neq p.$$

These are resolved arbitrarily, and cluster center for each point x_i is computed as follows,

$$Z_i = (1/n_i) \sum X_i,\ i = 1, 2\ldots, K \text{ and } X_j \varepsilon Z_i.$$

Where n_i is the number of elements belonging to cluster Z_i.
Step 3*: Repeat this step 2 until there are no changes in centroid values.*

PSO-K-Means Clustering

Particle Swarm Optimization (PSO) algorithm is one of the swarm intelligence methods and evolutionary optimization techniques are applied for clustering. PSO is a population-based, globalized search algorithm that uses the principles of the social behavior of the swarm and it is an efficient, simple, and an effective global optimization algorithm that can solve discontinuous, multimodal, and non-convex problems (Kuo et al. 2011). It is computationally effective and easier to implement when compared with other mathematical algorithms and evolutionary algorithms.

In PSO, N particles are moving around in the D dimensional search spaces. Each particle move towards nearest neighborhood. Each particle communicates with some other particle and is exaggerated by the best centroid point found by any member of its current centroid value p_i. The vector p_i for that best neighbour is denoted by p_g. Initialize the particle's location best known position to its initial position: $p_i \leftarrow x_i$. Then correspondingly update the particles or the location of the centroid value position and their velocity (3.8) to know the best global position (3.9) or of the best centroid value to group the data. These steps are repeated until a termination criterion is met. Finally, after finding the global best position, best value of cluster centroid is obtained.

$$v_{id} = w * v_{id} + C_1 * rand1 * (P_{id} - x_{id}) + C_2 * rand2 * (P_{gd} - x_{id}) \quad 3.8$$

$$x_{id} = x_{id} + v_{id} \quad 3.9$$

Where, v_{id} : velocity of particle , x_{id} : current position of particle

w : weighting function, $w = w_{max} - \frac{w_{max}-w_{min}}{iter_{max}} * iter$

w_{min}, w_{max} : initial and final weight

iter : current iteration , itermax : maximum iteration

c1&c2: determine the relative influence of the social and cognitive components

p_{id} : pbest of particle i, p_{gd} : gbest of the group.

The personal best position of particle is calculated as follows

$$p_{id}(t+1) = \begin{cases} p_{id}(t) & \text{if } f(X_{id}(t+1)) \geq f(p_{id}(t)) \\ X_{id}(t) & \text{if } f(X_{id}(t+1)) < f(p_{id}(t)) \end{cases}$$

The particle to be drawn toward the best particle in the swarm is the global best position of each particle. At the start, an initial position of the particle is considered as the personal best and the global best can be identified with minimum fitness function value.

Algorithm 4: TRS-PSO-K-Means Clustering Algorithm

Input: ***D*** *set of N Tags,* ***K*** *-number of clusters,* ***δ*** *– upper approximation threshold.*

Output: ***K*** *overlapping clusters of tags from* ***D*** *with associated membership value*

Step 1: *Initialize Number of clusters and its centroids using TRS approach*

Step 2: *Assign each vector in the data set to the closest centroid vector*

Step 3: *Calculate the fitness value for each Tag vector and update the velocity and Particle position, using equations (3.8) and (3.9) and generate the next solutions*

Step 4: *Repeat steps (2) and (3) until one of the following termination conditions is Satisfied.*

(a) The maximum number of iterations is exceeded

or

(b) *The average change in centroid vectors between iterations is less than a predefined value.*

PSO-K-Means-GA Clustering Algorithms

In this Clustering Algorithm, we are going to combine two global optimization algorithms, i.e., GA and PSO. Since PSO-KM and GA both work with a population of solutions, combining the searching abilities of both methods seems to be a wiser approach (Martens et al. 2011). On the contrary sense, GA works based on evolution from generation to generation, so the changes of individuals in a single generation are not considered. Based on the compensatory property of GA and PSO, we insist a new algorithm that combines the evolutionary ideas of both. In the reproduction and the crossover operation of GAs, individuals are reproduced or selected as parents directly to the next generation without any

enhancement (Kuo and Lin 2010). However, in nature, individuals will grow up and become more suitable to the environment before producing offspring.

Algorithm 5: TRS-PSO-K-Means-GA

Input: *Set of N Tagged bookmarks, K -number of clusters,*
Output*: K overlapping Tag clusters*

Step 1: *Initialize Number of clusters and its centroids using TRS*
Step 2: *Assign each vector in the Data set to the closest centroid vector.*
Step 3: *Calculate the fitness value and update the velocity and particle position, using equations (3.8) and (3.9) to generate the next solutions.*
Step 4*: If the particle position is stagnated, then reassign the clusters using the genetic operators, crossover and mutation*
Step 5: *Repeat step (2-4) until one of the following termination conditions is satisfied.*
(a)The maximum number of iterations is exceeded
or
(b)The average change in centroids vectors between iterations is less than a predefined value.

Validity Measures for Clustering Algorithms

In this section, the validity measures such as MSQE and SICD are explained. The Validity Measure is an Objective Function of the Clustering algorithms. Clustering algorithms which give the minimum MSQE and SICD value is the algorithm which provides better performance than others.

Mean Square Quantization Error (MSQE)

$$f(X,C) = \sum_{i=1}^{N} Min\{\{||X_i - C_l||^2| \quad where \quad l = 1, \dots, K\}\} \qquad 3.10$$

The K clusters are the total within-cluster variance or the total mean-square quantization error (MSE), where $||X_i - C_l||^2$ is a Euclidean distance measure between the i^{th} data point of the tag x_i and the j^{th} cluster center c_l, $1 < i < n$, $1 < j < K$ and is an indicator of the distance of the n tags from their respective cluster centroids (Taher and Babak 2010).

Sum of Intra Cluster Distance (SICD)

$$J(C_1, C_2, \dots., C_k) = \sum_{i=1}^{k} \left(\sum_{X_j \in C_i} \|Z_i - X_j\| \right) \qquad 3.11$$

The Euclidean distance between each data vector in a cluster and the centroid of that cluster is calculated and summed up. Here K is the number of clusters, Z_i represents cluster centroids, X_j is the data vector (Neshat et al. 2012)

4 Experimental Results

The experimental data set is collected from del.icio.us, which is a popular Web 2.0 web site that helps users to share links to their favorite information items. A short experimental evaluation for 3 benchmark datasets is presented. The information about the data sets contains names of dataset, the number of tags and number of Bookmarks, which are given in Table 5. In clustering tag data, the bookmarks were treated as attributes and the tags are treated as objects.

Table 5 Dataset Description

S.No	Dataset	Tags	Bookmarks	Url
1	Social	53388	12616	http://nlp.uned.es/socialtagging/socialodp2k9/
2	DAI Labor	67104	14454	www.markusstrohmaier.info/datasets/
3	tags2con	2832	1474	disi.unitn.it/~knowdive/dataset/delicious/

4.1 Bookmark Selections by USQR

The USQR FS method attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value of the dataset. The features are reduced by the unsupervised QR algorithm the number of selected bookmarks is shown in Table 6.

Table 6 Selected bookmarks Using USQR

S.No	Datasets	Tags	Bookmark	Selected bookmarks
1	Social	53388	12616	**4214**
2	DAI Labor	67104	14454	**3482**
3	tags2con	2832	1474	**328**

Feature Selection is a process of finding any one subset from the dataset that can be used to replace the original dataset. This research work also provides the purpose of feature selection in unsupervised approach and the empirical results presented in this following section. Fig 3 shows the actual features and selected features for Social, DAI Labor and tags2con datasets. The features are selected by the unsupervised QR algorithm.

Fig. 3 Before and After Bookmark Selection

4.2 Performance Analysis of Clustering Algorithms

In this section the performance of Benchmark algorithm K-Means are compared with the proposed algorithms TRS-K-Means, TRS-PSO-K-Means and TRS-PSO-K-Means-GA based on MSQE and SICD validity measures before and after bookmark selection.

K-Means: Benchmark Algorithm
The performance of K-Means Bench mark algorithm is analyzed based on MSQE and SICD validity measures on before Bookmark selection and the results are shown in Table 7.

Table 7 Performance Analysis of K-Means Algorithm

K-Means Clustering	Datasets					
	Social		DAI		tags2con	
	MSQE	**SICD**	**MSQE**	**SICD**	**MSQE**	**SICD**
K= 2-clusters	19.248	108.3	24.785	119.8	11.241	45.8
K= 4-clusters	18.007	97.4	21.219	112.4	10.114	41.2
K= 6-clusters	15.331	85.1	19.998	103.7	08.992	37.5
K= 8-clusters	13.992	78.9	16.332	94.9	05.927	33.2
K= 10-clusters	11.719	70.6	13.328	84.6	04.107	30.9

Fig 4 depicts the performance of K-Means Benchmark Clustering algorithm for various Social Tagging System data sets discussed here based on MSQE validity measure**.**

Fig 5 depicts the performance of K-Means Bench mark Clustering algorithm for various Social Tagging System data sets discussed here based on SICD validity measure**.**

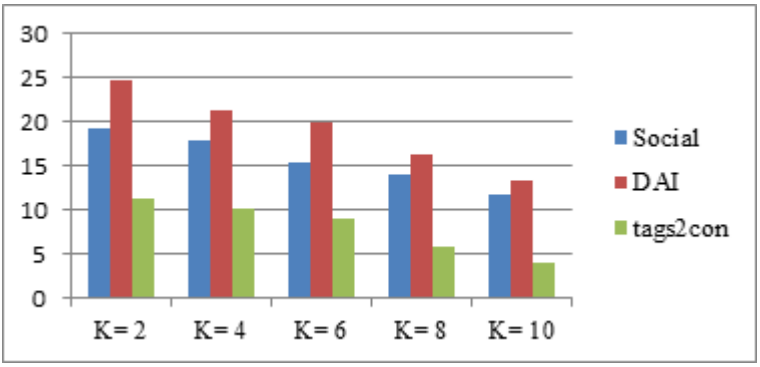

Fig. 4 Performance of K-Means based on MSQE index

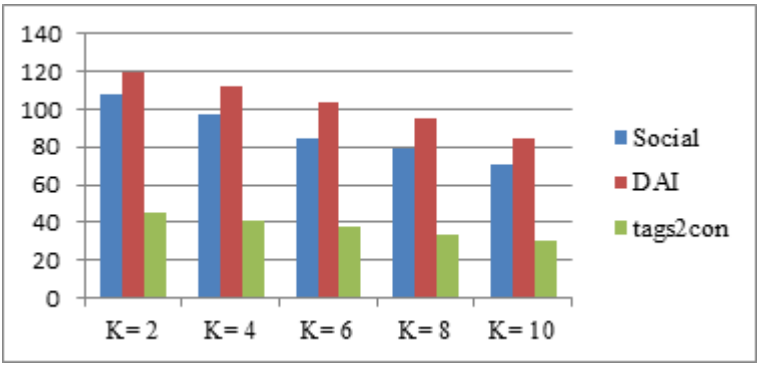

Fig. 5 Performance of K-Means based on SICD index

The performance of K-Means Bench mark algorithm is analyzed based on MSQE and SICD validity measures after bookmark selection and the results are shown in Table 8.

Table 8 Performance Analysis of K-Means Algorithm

K-Means Clustering	Datasets					
	Social		DAI		tags2con	
	MSQE	SICD	MSQE	SICD	MSQE	SICD
K= 2-clusters	12.208	87.6	17.364	102.4	08.334	32.8
K= 4-clusters	11.203	81.7	15.992	96.3	07.009	27.4
K= 6-clusters	09.278	74.3	13.725	91.7	06.168	23.9
K= 8-clusters	08.102	68.3	12.001	84.1	04.998	17.1
K= 10-clusters	06.552	62.5	10.782	78.9	02.004	13.7

Fig 6 depicts the performance of K-Means Benchmark Clustering algorithm for various Social Tagging System data sets discussed here based on MSQE validity measure**.**

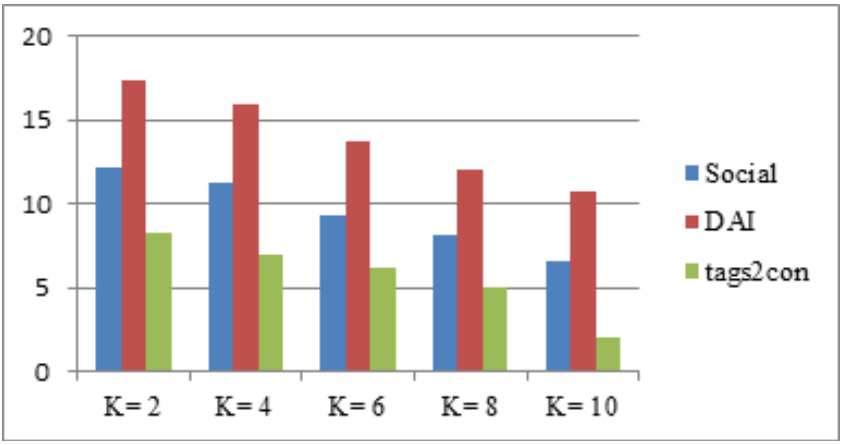

Fig. 6 Performance of K-Means based on MSQE index

Fig 7 depicts the performance of K-Means for various Social Tagging System data sets discussed here based on SICD validity measure**.**

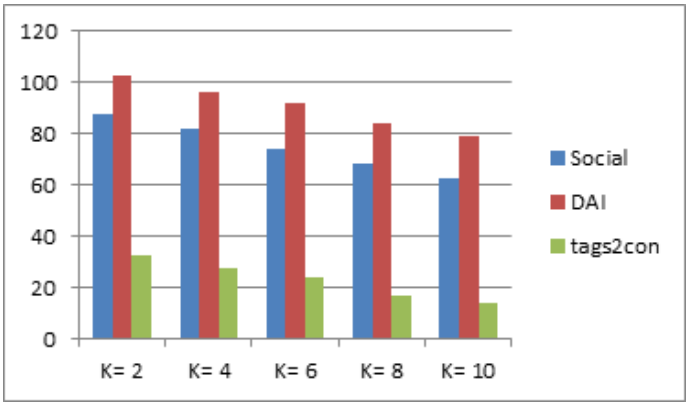

Fig. 7 Performance of K-Means based on SICD index

TRS-K-Means Clustering Algorithm

The performance of TRS-K-Means Clustering algorithm is analyzed based on MSQE and SICD validity measures before and after Bookmark selection and the results are shown in Table 9.

Table 9 Performance Analysis of TRS-K-Means

Datasets	Before Bookmark Selection			After Bookmark Selection		
	No.of. Clusters	MSQE	SICD	No.of. Clusters	MSQE	SICD
Social	16	14.108	81.9	6	09.448	72.8
DAI	22	14.668	89.1	8	11.852	80.1
tags2con	12	09.112	35.7	4	05.874	26.2

Fig 8 depicts the performance of TRS-K-Means for various Social Tagging System data sets discussed here based on a MSQE validity measure before and after bookmark selection**.**

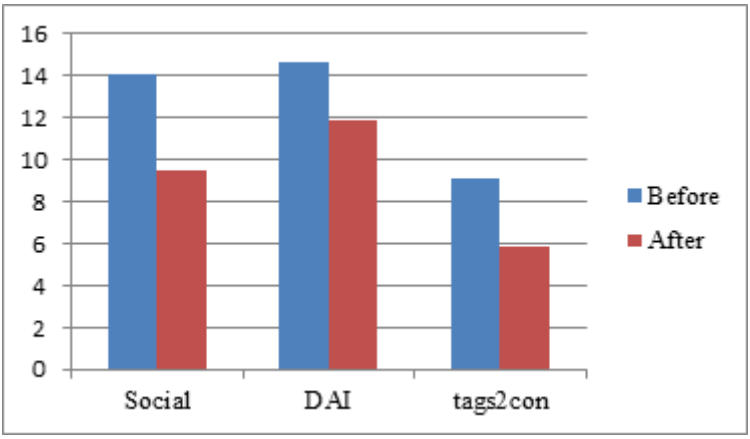

Fig. 8 Performance of TRS-K-Means based on MSQE index

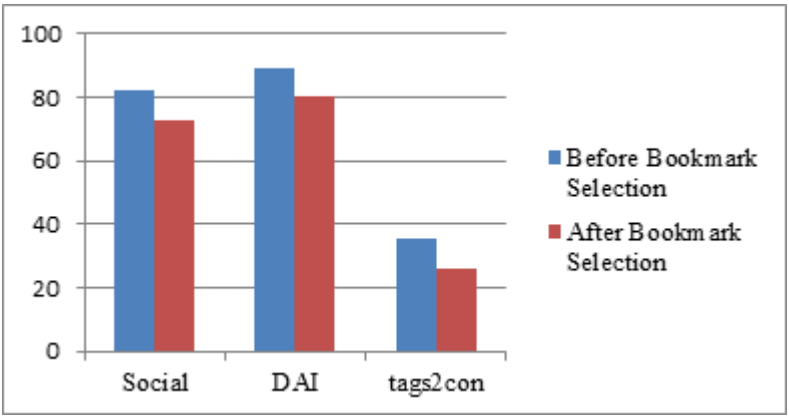

Fig. 9 Performance of TRS-K-Means based on SICD index

Fig 9 depicts the performance of TRS-K-Means for various Social Tagging System data sets discussed here based on SICD validity measure before and after bookmark selection**.**

TRS-PSO-K-Means Clustering algorithm

The performance of TRS-PSO-K-Means Clustering algorithm is analyzed based on MSQE and SICD validity measures before and after Bookmark selection and the results are shown in Table 10.

Table 10 Performance Analysis of TRS-PSO-K-Means

Datasets	Before Bookmark Selection			After Bookmark Selection		
	No.of. Clusters	MSQE	SICD	No.of. Clusters	MSQE	SICD
Social	16	12.552	76.3	6	08.985	69.7
DAI	22	12.901	85.7	8	11.002	77.1
tags2con	12	07.669	29.8	4	05.108	24.1

Fig 10 depicts the performance of TRS-PSO-K-Means for various Social Tagging System data sets discussed here based on a MSQE validity measure before and after bookmark selection**.**

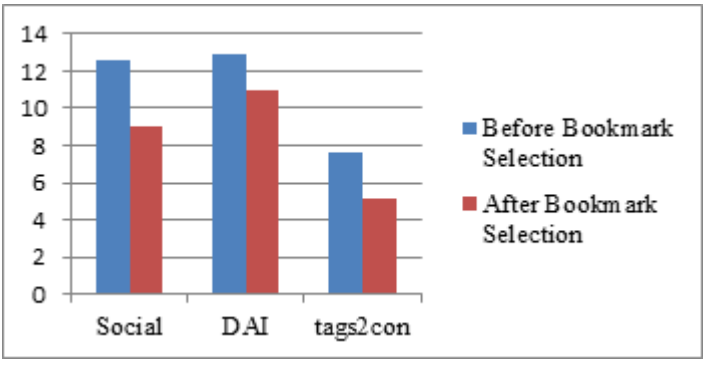

Fig. 10 Performance of TRS-PSO-K-Means based on MSQE index

Fig 11 depicts the performance of TRS-PSO-K-Means for various Social Tagging System data sets discussed here based on SICD validity measure before and after bookmark selection**.**

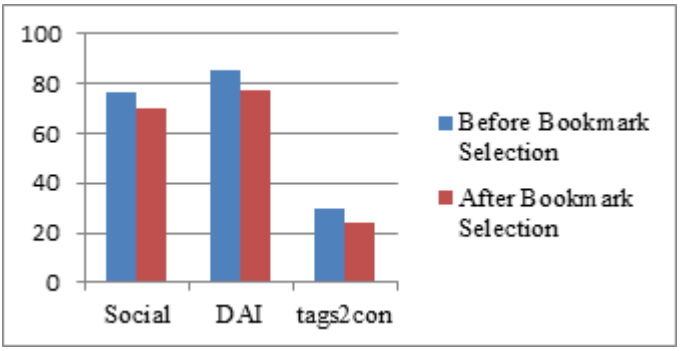

Fig. 11 Performance of TRS-PSO-K-Means based on SICD index

TRS-PSO-K-Means-GA Clustering Algorithm

The performance of TRS-PSO-K-Means-GA Clustering algorithm is analyzed based on MSQE and SICD validity measures before and after Bookmark selection and the results are shown in Table 11.

Table 11 Performance Analysis of TRS-PSO-K-Means-GA

Datasets	Before Bookmark Selection			After Bookmark Selection		
	No.of. Clusters	MSQE	SICD	No.of. Clusters	MSQE	SICD
Social	16	11.235	73.1	6	07.661	66.4
DAI	22	11.621	82.3	8	09.654	73.8
tags2con	12	06.597	26.9	4	04.231	22.7

Fig 12 depicts the performance of TRS-PSO-K-Means-GA for various Social Tagging System data sets discussed here based on a MSQE validity measure before and after bookmark selection**.**

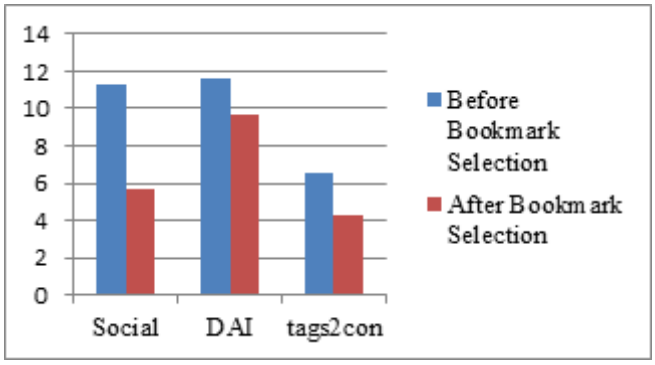

Fig. 12 Performance of TRS-PSO-K-Means-GA based on MSQE index

Fig 13 depicts the performance of TRS-PSO-K-Means-GA for various Social Tagging System data sets discussed here based on SICD validity measure before and after bookmark selection.

Fig. 13 Performance of TRS-PSO-K-Means-GA based on SICD index

4.3 Comparative Analysis

In this section, the comparative analysis of K-Means Clustering, TRS-K-Means clustering, TRS-PSO-K-Means, and TRS-PSO-K-Means-GA clustering for Social Tagged Data using distance measures such as MSQE and SICD measures are provided. This section attempts to carry out comparative analysis of proposed algorithms TRS-K-Means clustering, TRS-PSO-K-Means, and TRS-PSO-K-Means-GA clustering, in terms of the quality of the clusters.

Comparative Analysis Based on MSQE

The Table 12 shows the comparative analysis of clustering algorithms based on MSQE validity measures for all the Social Tagged datasets before Bookmark selection. The experimental results showed that PSO-K-Means-GA clustering algorithm had shown better results than other clustering algorithms such as K-Means, TRS-K-Means and TRS-PSO-K-Means.

Table 12 Comparative analysis based on MSQE for clustering algorithms

Datasets	No. of Clusters Set by TRS	TRS Based Clustering Algorithms		
		K-Means	PSO-K-Means	PSO-K-Means-GA
Social	16	14.108	12.552	11.235
DAI	22	14.668	12.901	11.621
tags2con	12	09.112	07.669	06.597

Figure 14 shows the comparative analysis of various clustering approaches for Social Tagging Systems. It can be observed from the figure that TRS-PSO-K-

Means-GA outperforms other approaches such as K-Means, TRS-K-Means and TRS-PSO-K-Means clustering algorithms based on a MSQE measure before bookmark selection.

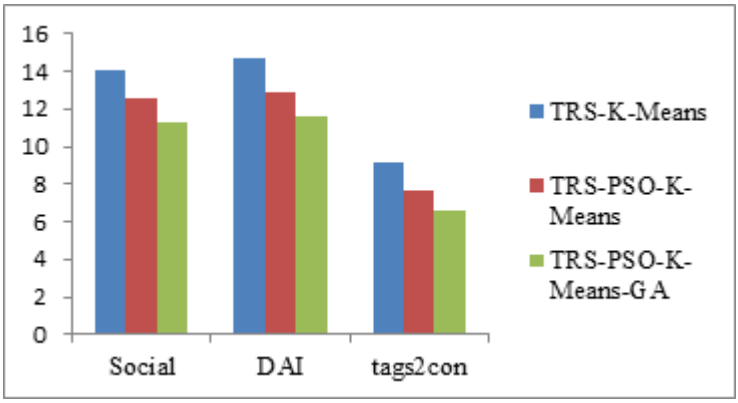

Fig. 14 Comparative Analysis of clustering algorithms based on MSQE

Table 13 shows the comparative analysis of clustering algorithms based on a MSQE validity measure for all the Social Tagged datasets after Bookmark selection. The experimental results showed that PSO-K-Means-GA clustering algorithm had shown better results than other clustering algorithms such as K-Means, TRS-K-Means and TRS-PSO-K-Means.

Table 13 Comparative analysis based on MSQE for clustering algorithms

Datasets	No. of Clusters Set by TRS	K-Means	TRS Based Clustering Algorithms		
			K-Means	PSO-K-Means	PSO-K-Means-GA
Social	6	09.278	09.448	08.985	07.661
DAI	8	12.001	11.852	11.002	09.654
tags2con	4	07.009	05.874	05.108	04.231

Figure 15 and Table 13 shows the comparative analysis of various clustering approaches for Social Tagging Systems. It can be observed from the figure that TRS-PSO-K-Means-GA outperforms other approaches K-Means, TRS-K-Means and TRS-PSO-K-Means clustering algorithms based on MSQE measures before bookmark selection.

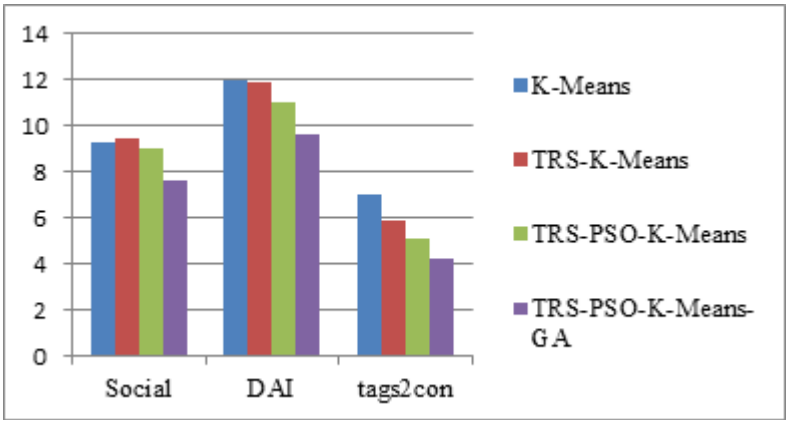

Fig. 15 Comparative Analysis of Clustering algorithms based on MSQE

Comparative Analysis Based on SICD

The Table 14 shows the comparative analysis of clustering algorithms based on SICD validity measure for all the Social Tagged datasets on before Bookmark selection. The experimental results showed that PSO-K-Means-GA clustering algorithm had shown better results than other clustering algorithms such as K-Means, TRS-K-Means and TRS-PSO-K-Means.

Table 14 Comparative analysis based on SICD for clustering algorithms

Datasets	No. of Clusters Set by TRS	TRS Based Clustering Algorithms		
		K-Means	PSO-K-Means	PSO-K-Means-GA
Social	16	81.9	76.3	73.1
DAI	22	89.1	85.7	82.3
tags2con	12	35.7	29.8	26.9

Figure 16 shows the comparative analysis of various clustering approaches for Social Tagging Systems. It can be observed from the figure that TRS-PSO-K-Means-GA outperforms other approaches K-Means, TRS-K-Means and TRS-PSO-K-Means clustering algorithms based on SICD measures on before bookmark selection.

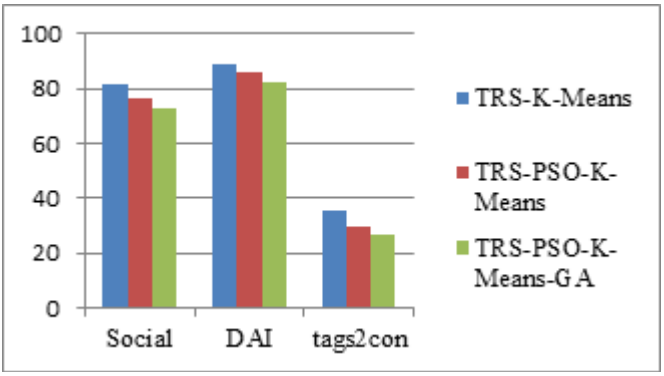

Fig. 16 Comparative Analysis of Clustering algorithms based on SICD

The Table 15 shows the comparative analysis of clustering algorithms based on SICD validity measure for all the Social Tagged datasets after Bookmark selection. The experimental results showed that PSO-K-Means-GA clustering algorithm had shown better results than other clustering algorithms such as K-Means, TRS-K-Means and TRS-PSO-K-Means.

Table 15 Comparative analysis based on SICD for clustering algorithms

Datasets	No. of Clusters Set by TRS	K-Means	TRS Based Clustering Algorithms		
			K-Means	PSO-K-Means	PSO-K-Means-GA
Social	6	74.3	72.8	69.7	66.4
DAI	8	84.1	80.1	77.1	73.8
tags2con	4	27.4	26.2	24.1	22.7

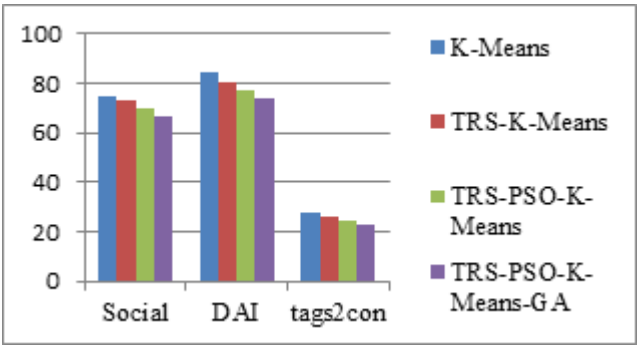

Fig. 17 Comparative Analysis of Clustering algorithms based on SICD

Figure 17 shows the comparative analysis of various clustering approaches for Social Tagging Systems. It can be observed from the figure that TRS-PSO-K-Means-GA outperforms other approaches K-Means, TRS-K-Means and TRS-PSO-K-Means based on SICD measures after bookmark selection.

5 Interpretations of Result

Data mining techniques are widely used for knowledge extraction in all areas. Social tagging system generated a huge amount of data in a short period. The data has more irrelevant bookmarks. So we can apply data mining techniques for knowledge extraction. Here feature selection techniques are applied to bookmark selection and clustering technique is applied for tag clustering. Bookmark selection is used to select bookmarks that are more related to the tags. Tag clusters are used for future web information retrieval systems. Table 16 shows sample bookmarks selected by applying USQR approach. Table 17 shows a sample tag cluster along with related bookmarks. The tag clusters are very useful to classify the web 2.0 websites.

Table 16 Sample of selected BM from tag2con dataset

S.no	Tag ID	Bookmark	Book mark
1	174	59	www.classroom20.com
2	52	48	www.writingfix.com
3	36	35	www.edublogs.org
4	242	14	www.countryreports.org
5	88	37	www.secondlife.com

Table 17 Sample of Tag cluster with related BM from tag2con dataset

S.no	Tag ID	Tag	Bookmark
1	2	library	www.ifla.org
2	88	games	www.secondlife.com
3	36	education	www.edublogs.org
4	68	technology	www.classroom20.com
5	343	blog	www.edublogs.org

6 Conclusion

The bookmark selection and tag clustering problem is a very important problem and has attracted much attention of many researchers. In this Chapter, USQR algorithm is applied to Bookmark selection and also this chapter has proposed a new hybrid algorithm for solving the clustering problem which is based on the combination of TRS and Meta Heuristic Clustering Algorithms. The proposed Clustering algorithms are Hybrid TRS-K-Means, Hybrid TRS-PSO-K-Means, and Hybrid TRS-PSO-K-Means-GA. The proposed clustering algorithms are compared with K-Means benchmark algorithm for Social Tagging System Dataset. The goodness of the clusters is obtained using the two well-known measures such as MSQE and SICD. The comparative analysis shows that TRS-PSO-K-Means-GA had given the best performance over the other two approaches for Social Tagging System Dataset. This research work also shows the importance of bookmark selection method for the performance of clustering approaches after bookmark selection is better than the performance before feature selection

Acknowledgement. The authors would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-41-650/2012 (SR).

References

Ahmadi, A., Karray, F., Kamel, M.S.: Flocking based approach for data clustering. Nat. Comput. 9(3), 767–791 (2010)

Azar, A.T., Banu, P.K.N., Inbarani, H.H.: PSORR - An Unsupervised Feature Selection Technique for Fetal Heart Rate. In: 5th International Conference on Modelling, Identification and Control (ICMIC 2013), Egypt, August 31-September 1-2 (2013)

Azar, A.T., Hassanien, A.E.: Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier. Soft computing (2014), doi:10.1007/s00500-014-1327-4

Begelman, G., Keller, P., Smadja, F.: Automated Tag Clustering: Improving search and exploration in the tag space. In: 15th WWW Conference on Collaborative Web Tagging Workshop, Edinburgh (2006)

Bolón-Canedo, V., Snchez-Maroón, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowledge and Information Systems 34(3), 483–519 (2012)

Lu, C., Hu, X., Park, J.-R.: Exploiting the social tagging network for web clustering. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 41(5), 840–852 (2011)

Dattolo, A., Eynard, D., Mazzola, L.: An Integrating Approach To Discover Tag Semantics. In: Proceedings of the 2011 ACM Symposium on Applied Computing, TaiChung, Taiwan, March 21-24 (2011)

De, S.K., Krishna, P.R.: Clustering web transactions using rough approximation. Fuzzy Set. Syst. 148(1), 131–138 (2004)

Dhanalakshmi, K., Inbarani, H.H.: Fuzzy Soft Rough K-Means Clustering Approach For Gene Expression Data. Int. J. of Scientific Engineering and Research 3(10), 1–7 (2012)

Esmin, A.A., Coelho, R.A., Matwin, S.: A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 1–23 (2013)

Gemmell, J., Shepitsen, A., Mobasher, B., Burke, R.: Personalization in Folksonomies Based on Tag Clustering. In: Intelligent Techniques for Web Personalization & Recommender Systems, pp. 259–266. ACM, Chicago (2008)

Grbovic, M., Djuric, N., Guo, S., Vucetic, S.: Supervised clustering of label ranking data using label preference information. Machine Learning 93(2-3), 191–225 (2013)

Gupta, M., Li, R., Yin, Z., Han, J.: Survey on social tagging techniques. ACM SIGKDD Explor. Newsl. 12(1), 58–72 (2010)

Hammouda, K.: A Comparative Study of Data Clustering Techniques. Technical Report, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada (2006)

Heymann, P., Koutrika, G., Garcia-Molina, H.: Can social bookmarking improve web search. In: Proceedings of the International Conference on Web Search and Web Data mining, pp. 195–206. ACM, New York (2008)

Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. Int. J. of Intelligent Systems 17(2), 199–212 (2002)

Hu, X., Cercone, N.: Data mining via discretization, generalization and rough set feature selection. Knowledge and Information System 1(1), 33–60 (1999)

Inbarani, H.H., Thangavel, K., Pethalakshmi, A.: Rough Set Based Feature Selection for Web Usage Mining. In: International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), Sivakasi, December 13-15, pp. 33–38. IEEE (2007)

Inbarani, H.H., Thangavel, K.: Mining and analysis of clickstream patterns. In: Abraham, A., Hassanien, A.E., De Carvalho, A.P., Snasel, V. (eds.) Foundations of Comput. Intel. Vol. 6. SCI, vol. 206, pp. 3–27. Springer, Heidelberg (2009)

Inbarani, H.H., Banu, P.K.N., Azar, A.T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Computing and Applications (2014a), doi:10.1007/s00521-014- 1552-x

Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Computer Methods and Programs in Biomedicine 113(1), 175–185 (2014b)

Jain, A.K., Murty, M.N., Flyn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)

Jiang, D., Tang, C., Zhang, A.: Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 16(11), 1370–1386 (2004)

Jianwen, M.A., Bagan, H.: Remote sensing data classification using tolerant rough set and neural networks. Science in China Ser. D Earth Sciences 48(12), 2251–2259 (2005)

Jothi, G., Inbarani, H.H.: Soft Set Based Feature Selection Approach for Lung Cancer Images. Int. J. of Scientific Engineering and Research 3(10), 1–7 (2012)

Jothi, G., Inbarani, H.H., Azar, A.T.: Hybrid Tolerance-PSO Based Supervised Feature Selection For Digital Mammogram Images. International Journal of Fuzzy System Applications (IJFSA) 3(4), 15–30 (2013)

Kisilevich, S., Mansmann, F., Nanni, M.: Rinzivillo S Spatio-Temporal Clustering. In: Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 855–874. Springer Press, New York (2010)

Kumar, S.S., Inbarani, H.H.: Web 2.0 social bookmark selection for tag clustering. In: Pattern Recognition, Informatics and Medical Engineering (PRIME), Periyar University, Salem, February 22-23, pp. 510–516. IEEE (2013a)

Kumar, S.S., Inbarani, H.H.: Analysis of mixed C-means clustering approach for brain tumour gene expression data. Int. J. of Data Analysis Techniques and Strategies 5(2), 214–228 (2013b)

Kuo, R.J., Wang, M.J., Huang, T.W.: An application of particle swarm optimization algorithm to clustering analysis. Soft Computing 15(3), 533–542 (2011)

Kuo, R.J., Lin, L.M.: Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering. Decis. Support. Syst. 49(4), 451–462 (2010)

Mangai, J.A., Kumar, V.S., Appavu, S.: A Novel Feature Selection Framework for Automatic Web Page Classification. Int. J. of Automation and Computing 9(4), 442–448 (2012)

Martens, D., Baesens, B., Fawcett, T.: Editorial survey: swarm intelligence for data mining. Machine Learning 82(1), 1–42 (2011)

Mitra, P., Murthy, C., Pal, S.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intel. 24(4), 301–312 (2002)

Moftah, H.M., et al.: Adaptive k-means clustering algorithm for MR breast image segmentation. Neural Computing and Applications (2013)

Neshat, M., Yazdi, S.F., Yazdani, D., Sargolzaei, M.: A New Cooperative Algorithm Based on PSO and K-Means for Data Clustering. J. of Computer Science 8(2), 188–194 (2012)

Parthalain, N.M., Jensen, R.: Unsupervised fuzzy-rough set-based dimensionality reduction. Inf. Sci. 229, 106–121 (2013)

Rana, S., Jasola, S., Kumar, R.: A review on particle swarm optimization algorithm and their application to data clustering. Artificial Intelligence Review 35(3), 211–222 (2011)

Sbodio, M.L., Simpson, E.: Tag Clustering with Self Organizing Maps. Hewlett-Packard Development Company (2009)

Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM Conference on Recommender Systems, New York, USA, pp. 259–266 (2008)

Taher, N., Babak, A.: An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Appl. Soft Comput. 10(1), 183–197 (2010)

Velayutham, C., Thangavel, K.: Unsupervised Quick Reduct Algorithm Using Rough Set Theory. J. of Electronic Science and Technology 9(3), 193–201 (2011)

Xu, G., Zong, Y., Pan, R., Dolog, P., Jin, P.: On kernel information propagation for tag clustering in social annotation systems. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 505–514. Springer, Heidelberg (2011)

Yau, K.L., Tsang, P.W.M., Leung, C.S.: PSO-based K-means clustering with enhanced cluster matching for gene expression data. Neural Computing and Application 22(7-8), 1349–1355 (2013)

Exploitation of Healthcare Databases in Anesthesiology and Surgical Care for Comparing Comorbidity Indexes in Cholecystectomized Patients

Luís Béjar-Prado*, Enrique Gili-Ortiz, and Julio López-Méndez

Abstract. Objective: Charlson comorbidity index (CCI) and Elixhauser comorbidity index (ECI) have been used as prognostic tools in surgical and medical research. We compared their ability to predict in-hospital mortality among cholecystectomized patients using a Spanish large database of 87 hospitals during the period 2008-2010.

Methods: The electronic healthcare database Minimal Basic Data Set (MBDS) contain information of diseases, conditions, procedures and demographic data of patients attended in hospital setting. We used available information to calculate CCI and ECI, and analyzed their relation to in-hospital mortality.

Results: The models including age, gender, tobacco use disorders, hospital size and CCI or ECI were predictive of in-hospital mortality among cholecystectomized patients, when measured in terms of adjusted Odds Ratios and 95% confidence limits. There was a dose-effect relationship between score of prognostic indexes and risk of death. Area under the curve for ROC predictive models for

Luís Béjar-Prado · Julio López-Méndez
Department of Preventive Medicine and Public Health, University of Seville, Spain
e-mail: lmbprado@us.es

Enrique Gili-Ortiz
Unit of Anesthesiology and Pain Therapy,
Hospital Universitario Virgen Macarena, Seville, Spain

Julio López-Méndez
Unit of Preventive Medicine, Surveillance and Health Promotion,
Hospital Universitario Virgen Macarena, Seville, Spain
* Corresponding author.

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_9

in-hospital mortality were 0.8717 for CCI and 0.8771 for ECI, but differences were not statistically significant ($p>10^{-6}$).

Conclusion: Both CCI and ECI were predictive of in-hospital mortality among cholecystectomized patients in a large sample of Spanish patients after controlling for age, gender, group of hospital and tobacco use disorders. The availability of more hospitals databases through Big Data can strengthen the external validity of these results if we control several threats of internal validity such as biases and missing values.

1 Introduction

1.1 Clinical Databases

Clinical databases can be very powerful and productive research tools. The long-term and prospective collection of accurate clinical information on specific disease processes can provide valuable knowledge to researchers and clinicians. They can provide useful information for the perioperative process, the field of research and evaluation of surgeons and anesthesiologists.

The easiest part to a database is collecting the data. What is far more difficult is to then process those data into information. A further level of analysis is then required to turn all this information into knowledge that a surgeon and/or an anesthesiologist may use and comprehend; and finally, detailed analysis is required to answer difficult or challenging clinical questions (Platell 2008)

While the delivery of anaesthesia care is largely a safe process and adverse events are infrequent, they can have devastating consequences for patients and providers when they occur. The goal of perioperative effectiveness research is to improve efficacy, effectiveness and efficiency in surgical and anesthesia care. Generally, delivery of anaesthesia care is a safe process and adverse events are rare, but they can have severe consequences for patients when they occur. Perioperative effectiveness research could afford new approaches leading to improved surgical and anaesthesia care using large databases.

Big data is the statistical and analytical use of many thousands of data points in a retrospective manner to determine subtle patterns and correlations that would otherwise not be detected in a prospective study with a smaller sample. Big Data offers the ability to collate and process large amounts of data, thanks to the wonder of modern computers and databases, and to derive useful information from these data. Probably, the next step in epidemiologic research and evaluation of surgical and anesthesia procedures is to use the whole set of databases for analyzing results, Big Data analysis.

An important advantage of having a large sample size from multiple data sources is the ability to study rare exposures, outcomes, or subgroups. Data from multiple sources also often represent demographically and geographically diverse study populations, which allow for study of geographic or practice variations or treatment heterogeneity. There is a growing body of real-world experience that

demonstrates how distributed networks of electronic healthcare databases support large-scale epidemiologic studies or surveillance assessments (Toh and Platt 2013).

Nevertheless, obsession with study power and precision may have blurred fundamental validity issues not solved by increasing sample size, for example, measurement error, selection bias, or residual confounding (Chiolero 2013).

So, our approach in this chapter is first, to review perioperative comparative effectiveness research and types of epidemiologic studies used, their strengths and their disadvantages and limitations. Second, we'll analyze the use of large databases, including administrative databases in perioperative research. Third, we'll evaluate the importance of comorbidities in risk stratification, using the comparative analysis of two comorbidity indexes in the risk of mortality of cholecystectomized patients in a sample of 87 Spanish hospitals. Fourth, we'll comment the results of some additional applications of large database analysis in perioperative research. Finally, in a Discussion section we'll analyze some limitations and cautions in the analysis and interpretation of results, including validity of results, such as data quality and depth, control of bias, missing data, type I (alpha) error and the cautions and checklists for evaluating the quality of published studies that use large databases and Big data.

1.2 Perioperative Comparative Effectiveness Research

Comparative Effectiveness Research (CER) has been defined by the Federal Coordination Council for Comparative Effectiveness Research (FCCCCER) as the following (FCCCCER 2009):

"Comparative effectiveness research is the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in "real world" settings. The purpose of this research is to improve health outcomes by developing and disseminating evidence-based information to patients, clinicians, and other decision makers, responding to their expressed needs, about which interventions are most effective for which patients under specific circumstances. To provide this information, comparative effectiveness research must assess a comprehensive array of health-related outcomes for diverse patient populations and subgroups. Defined interventions compared may include medications, procedures, medical and assistive devices and technologies, diagnostic testing, behavioral change, and delivery system strategies. This research necessitates the development, expansion, and use of a variety of data sources and methods to assess comparative effectiveness and actively disseminate the results."

The ultimate goal identified is that CER should produce useful evidence suited to help physicians to make treatment choices that would be most beneficial in a specific setting. One way to control cost is to ensure that medical care is based on scientific evidence. Research suggests that up to 50 percent of medical care provided in the United States is not based on sufficient scientific support and that

approximately 30 percent of medical interventions are of "uncertain or questionable value"(Manchikanti et al. 2010). The overarching goal of CER is therefore to improve the efficacy, effectiveness and efficiency of health care. The goals of an ideal CER study have been summarized and it has been suggested that it should address the questions of 1) Is the intervention working, that is, efficacious?, 2) Is it applicable to a large portion of the population, that is, effective? 3) Is the cost of the intervention worth it, that is, efficient? (Kheterpal 2009).

Randomized controlled trials (RCTs) are considered the best study designs for the evaluation of outcomes because randomization prevents the confounding effects of measured and unmeasured variables. If efficacy is defined as the potential benefit or risk of a procedure under ideal circumstances in a specific patient population, a well conducted and analyzed RCTs provide high levels of internal validity of results (Myles et al. 2011).

The RCT design may, because of its high internal validity, be well suited to study the efficacy of a new drug such as a pain medication. While potentially very effective, as determined in an RCT, everybody wants to know about the potential risks of harm and side effects. This information, however, would require external validation in a larger and more diverse population. An alternative has been the design and implementation of Practical Clinical Trials (PCTs),that compare clinically relevant alternative interventions, select diverse populations for study, use diverse practice settings and evaluate a wide range of outcomes. Therefore, PCTs are better equipped to address the "real world" questions of interest to researchers (Tunis et al. 2003; Evley et al. 2010).

Given the wide range of perioperative clinical practice throughout the country, PCTs may therefore be a preferred design; however, they would require the collaboration of multiple study locations, similar to a multi-centre design, but not limited to high-volume, academic settings.

However, despite the advantages, traditional RCTs may not be appropriate to answer every CER question in perioperative care. For example, many reports suggest that high-volume centers have better outcomes for specialized procedures than do low-volume centers (Weiss et al. 2009).

Randomly selecting diverse candidates for assignment to high-volume centers would not be logical or ethical, and even if voluntary participation could be ensured, processes required to identify, stratify and assign candidates to hospitals could complicate the external validity of trial results. Therefore, not every perioperative CER question can be answered by RCTs. RCTs can be very expensive when infrequent outcomes require long study periods and large numbers of participants. On the other hand, RCT results, even when well designed, can lack external validity (the results cannot be generalized to other populations). Finally, it is ethically not feasible to randomly assign patients to a potentially harmful intervention (Memtsoudis and Besculides 2011).

In the perioperative setting, observational studies could be particularly useful in assessing the potential impact in a larger segment of the population. Observational studies can be very useful when studying rare outcomes and they frequently are based on results obtained from actual clinical settings. Internal validity of

observational studies is lower than that of RCTs. Among observational studies, cohort studies (prospective or retrospective) analyze the effects of exposures in an initially healthy population after a follow-up period. They have the advantage of analyzing multiple outcomes simultaneously, and this feature is important because perioperative interventions may impact in a number of outcomes. In cohort studies time-effect relationships between intervention and outcome are well defined, because outcomes were not present initially when patients were exposed to procedure (Powell et al. 2011; Fuller et al. 2011).

In case-control studies a sample of patients with outcome (cases) is compared with a sample of patients without outcome (controls) analyzing the prevalence of exposure (procedure) in both groups. These studies are cheaper and require relatively little time for completion (Turan et al. 2011). Nevertheless, the risk of biased estimations is greater and time-effect relationships between exposure and outcome can be very difficult to prove in case-control studies.

A cohort study or a case–control study could provide useful data, especially when trying to assess the propensity for potential harm due to an intervention that may affect specific subpopulations of perioperative patients. It is not unlikely that many research questions may have to be addressed by a combination of RCTs and observational studies. For example, while the efficacy of a new drug may be evaluated with an RCT, its effectiveness and safety require evaluation in a broader range of patients that is not confined by the strict entry criteria of an RCT.

The extraction of valid findings of causal therapeutic benefits using non randomized studies, when RCTs are impossible or unethical to perform, has been extensively debated and multiple analytic designs have been developed. Many administrative databases in USA and other countries are being made available by the governments for research purposes. Informing the results of the clinical trials from the content of the secondary large databases is useful, but the methods for attaining the validity of results have become an important issue. A consensus-based report recommends the use of stratification analysis before multivariable modeling, propensity scoring, and other methods, where appropriate for secondary data. Sensitivity analyses and calibration of the extent of residual confounding are also recommended (Berger et al 2009; Cox et a. 2009; Johnson et al. 2009).

Many questions can be answered by a systematic review (i.e. meta-analysis) of existing, published data of RCTs and observational studies; thus, developing new CER studies may not always be necessary (White et al. 2009).

The most appropriate study design is determined by the question to be answered. If the intervention to be studied is potentially associated with a high risk of harm or cost, a high degree of internal validity is required and a well-designed and conducted controlled randomized clinical trial is the best option. If the intervention is to be implemented in a diverse population or environment, or would be difficult to replicate, the studies will require high external validity, such as RCTs performed in different populations or high-quality observational studies. Many real-world questions will require tradeoffs to balance requirements for internal validity, external validity, feasibility and timeliness. In these cases, the researcher must assign weights of importance to these components, and if, for example,

requires a moderate internal validity but a high external validity, the best option might be a well-designed observational study.

1.3 Large Databases and Perioperative Research

Over the last decade the systematic exploitation of large databases for perioperative and anaesthesia-related research has increased notably. Advances in information technology, epidemiologic research methods, and the realization that large databases information, in spite of being designed primarily for administrative purposes, could provide valuable clinical and epidemiologic information, have been supporting this trend (Freundlich and Kheterpal 2011; Memtsoudis et al. 2011).

A variety of large databases useful for clinical and epidemiologic research are available. Many are accessible to the public free of charge (e.g. the National Hospital Discharge Database in the USA or the Minimal Basic Data Set in Spain). Most require formal applications before use and in order to protect against misuse, many contain encrypted data for identifying participants, and all of them require the completion of a data use agreement. Because of the lack of identifiable information, work with these databases is often excused from review by research ethics committees.

The choice of a specific database for research will depend to a large extent on the hypothesis to be verified and the availability of relevant information within the data collection set. In most instances, the number of variables collected is usually limited and varies in completeness and accuracy. Further, a clear understanding of the primary purpose of the database (administrative or clinical), the number and the type of contributing sites, and also the database design, and the recompilation of data and verification process are some of many factors to be considered when deciding to work with a particular dataset. For example, researchers need to consider if they need nationally representative data (usually provided by specially designed weighting procedures) collected from a large number of sites or if institutional data from select hospitals are sufficient. In other cases they might need the whole database, as in those studies analyzing very rare events or outcomes. An important consideration before deciding to proceed with this type of research is the availability of an adequate infrastructure that includes programmers, statisticians, epidemiologists, and also hardware and software equipment suitable to handle these large datasets

Large databases allow the performance of many types of studies with the same dataset, including descriptive, epidemiologic and trend analytical investigations, various types of outcomes, and comparative effectiveness research, and also risk factor evaluations.

The availability of information from large numbers of patient allows for the study of rare events of otherwise difficult to study outcomes in subpopulations of patients. As data are not subject to the strict inclusion and exclusion criteria of randomized control trials, they represent an overview of actual practice, thus

affording results a high level of external validity. These characteristics have therefore led to publications on topics that have not been previously studied using randomized controlled trials. The comparison of outcomes of various types of anaesthetics on perioperative adverse effects seemed complicated because of the rarity of specific adverse events that require extremely large sample sizes (Stundner et al. 2012; Memtsoudis et al.2012; Neuman et al.2012).

The value of large databases has been demonstrated by published studies of the epidemiology of malignant hyperthermia, the perioperative outcomes after joint arthroplasty and the study of risk factors for cerebrovascular accidents (Rosero et al.2009; Memtsoudis et al.2009; Sharipfour et al. 2013).

Some administrative data are electronic data records that are typically generated at the time of hospital discharge or provision of other services. They usually contain primary and secondary diagnoses and at least some information about procedures performed and medications received, and they either contain or can be linked to files containing demographic information. Administrative data generally do not contain laboratory results, such as pathology or radiology reports, or clinical measures, such as blood pressure or height and weight (Virnig and McBean 2001).

Use of such data is becoming increasingly common. It is encouraged in the USA by the National Institutes of Health, the Agency for Health Care Research and Quality, the Health Care Financing Administration (HCFA), and the Department of Veterans Affairs (VA). Most states collect hospital discharge summaries, as do Canadian provinces. Electronic hospital discharge summaries form a major component of the data collected for the National Center for Health Statistics' National Hospital Discharge Survey. Many states collect hospital discharge summaries that are used for planning and may be available for research purposes. These databases contain information on admission and discharge dates, diagnoses and procedures, demographic information, outcomes and payer information. They are useful for analysis of admissions and discharges, and in some countries every patient has a code number, which makes them useful for studies of readmission and transfer patterns (Roos et al.1996; Ministerio de Sanidad 2011)

1.4 The Importance of Comorbidities in Risk Stratification

Age, gender, social class and disease-specific factors are crucial predictive factors, but besides them, perhaps the most important source of potential confounding in anaesthesia and perioperative studies is preoperative risk due to comorbidities.

Accurate prediction of perioperative risk is an important goal to enable informed consent for patients undergoing surgery and to guide clinical decision making in the perioperative period. In addition, by adjusting for risk, an accurate risk stratification tool enables meaningful comparison of surgical outcomes between providers for service evaluation or clinical audit. Some risk stratification tools have been incorporated into clinical practice, and indeed, have been recommended for these purposes (Nashef et al.1999; Copeland, Jones and Walters 1991).

Risk stratification tools may be subdivided into risk scores and risk prediction models. Both are usually developed using multivariable analysis of risk factors for a specific outcome. Risk scores assign a weighting to factors identified as independent predictors of an outcome; with the weighting for each factor often determined by the value of the regression coefficient in the multivariable analysis. The sum of the weightings in the risk score then reflects increasing risk. Risk scores have the advantage that they are simple to use in the clinical setting. However, although they may score a patient on a scale on which other patients may be compared, they do not provide an individualized risk prediction of an adverse outcome. Examples of risk scores are the American Society of Anesthesiologists' Physical Status score (ASA-PS) and the Lee Revised Cardiac Risk Index (Saklad 1941; Lee et al.1999).

Several prognostic tools and indexes have been evaluated in the perioperative setting, but a majority of them use primary data (clinical records, laboratory data and other results) (Moonesinghe et al. 2013). Some of them have been validated using administrative databases. The most commonly used means of comorbidity adjustment is the Charlson Comorbidity Index (Charlson et al. 1987). The Charlson index assigns weights to 17 disease-specific groups, most of which are given a weight of one. Diabetes mellitus with organ involvement, hemiplegia, severe and moderate renal disease, and any cancerous tumor, leukemia or lymphoma are assigned a weight of two. Moderate or severe hepatic disorders are assigned a weight of three, and metastatic solid tumors and AIDS a weight of six. The index is calculated by adding the weights for each condition in each patient. Although the Charlson Comorbidity Index was originally developed to predict 1-year mortality from medical record data (not administrative data) in a medical (not surgical) population, adaptations of it for use with ICD-9 codes have been validated and widely used in studies of mortality after surgery (Deyo, Cherkin and Ciol 1992; Romano, Roos and Jolli 1993; Cleves, Sanchez and Draheim 1997). Although comorbidities tend to be systematically undercoded in administrative data, the Charlson index as an aggregate measure performs similarly well whether derived from medical records or from administrative data (Quan, Parsons and Ghali 2002; Malenka et al. 1994).

The Elixhauser Comorbidity Index assigns a weight of 1 to each of 30 groups of specific diseases. Elixhauser Comorbidity Index was developed using administrative data (not medical records) for a mixed medical-surgical population and performs similarly well for the prediction of in-hospital mortality (Elixhauser et al.1998; Southern, Quan and Ghali 2004; Gili et al.2011).

Regardless of the method used, attention to several factors can optimize risk adjustment. The chosen risk adjustment method should have demonstrated validity for prediction of the outcome of interest (e.g., in-hospital mortality, long-term mortality, length of hospital stay, hospital costs, etc.). Risk adjustment using comorbidity indexes may be improved by empirically deriving site-specific or study-specific weights used to calculate the comorbidity score in the data set being analyzed, rather than simply using the weights (Klabunde et al.2007; Ghali et al. 1996; Martins and Blais 2006).

Finally, risk adjustment using administrative data can be significantly improved if it incorporates codes ("flag") that indicate whether a diagnosis was present on admission (a comorbidity) or developed subsequently (a complication). Some studies have validated comorbidities indexes comparing comorbidities diagnosed with the flag present on admission (POA) and without the flag (Pine et al.2007; Gili et al.2011).

Unfortunately, many administrative data do not contain this information. For example, in Spain only one region (Andalusia) has incorporated the flag POA in MBDS (Servicio Andaluz de Salud 2012).

The codes and methods used to assess the presence and influence of comorbid conditions should be tailored to the research question (Klabunde, Warren and Legler 2002).

In addition, the methods and rationale used to adjust for comorbidities should be clearly communicated. The degree to which these standards are followed, and the transparency with which they are described, are important to inform the reader of the adequacy of risk adjustment in a given analysis.

2 Aims and Structure of the Chapter

Research based on big databases has increased exponentially during the latest years because healthcare policy and clinical decision making needs information of large and diverse populations.

In clinical trials and cohort studies with rare outcomes and case-control studies of rare exposures, Big Data exploitation is the best approach to obtain more precise estimates. Nevertheless, at present researchers need to take into account several limitations of many databases.

The main objective of this Chapter is to compare the ability of two prognostic indicators (Elixhauser comorbidity indicator and Charlson comorbidity indicator) to predict in-hospital mortality using a Spanish large database of 87 hospitals.

Besides this, we want to analyze the future impact of Big Data on this kind of studies, and to review the cautions and limitations of available information in these big databases.

The chapter includes four parts:

3. METHODS
    - 3.1. Participants
    - 3.2. Variables
    - 3.3. Data analysis
4. RESULTS
    - 4.1. Patient characteristics
    - 4.2. Crude mortality rates

3 Methods

3.1 Participants

Data for hospital admissions were captured in the administrative minimal basic data set (MBDS) of 87 Spanish Hospitals during the period of 2008-2010.

From written or digitalized information that was provided by the hospital physician who signs the clinical record, each patient's diagnosis, external causes and procedures were codified according to International Classification of Diseases, 9th review (ICD-9-CM) codes. Codification and data entry in the electronic database are performed by dedicated administrative personnel who have completed in-depth training on medical data registration. This administrative database has demographic data, admission and discharge dates, type of admission and type of discharge, diagnostic codes for principal cause and secondary diagnoses, external causes and procedures using the ICD-9-CM. Also included in these data bases are Diagnosis-Related Groups (DRG) and a classification of each hospital into 5 categories based on size and complexity of services, varying from those with a smaller size and lesser complexity to larger hospitals that offer a broader range of services. Analysis was restricted to patients who at the moment of hospital discharge were 18 and older.

3.2 Variables

Cases of cholecystectomy were defined as those which fit the codes specified in the ICD-9-CM as a procedure with codes 51.21 and 51.22 for laparoscopic cholecystectomy and with codes 51.23 and 51.24 for non laparoscopic cholecystectomy. The ICD-9-CM code for disorders associated with tobacco dependence was used (305.1).

Age was measured in years. Also included in these data bases are Diagnosis-Related Groups (DRG) and a classification of each hospital into 5 categories based on size and complexity of services, varying from those with a smaller size and

lesser complexity to larger hospitals that offer a broader range of services (Ministerio de Sanidad 2011). Analysis was restricted to patients who at the moment of hospital discharge were 18 and older. The comorbidities included in Charlson Comorbidity Index and Elixhauser Comorbidity Index were calculated using ICD-9 codes. We used the ICD-9-CM codes proposed for these comorbidities by Quan et al (Quan et al. 2005).

3.3 Data Analysis

The primary outcome of interest was to determine the mortality in patients with cholecystectomy and its relationship with comorbidity indexes.

Univariate analysis was used to examine the association between mortality among cholecystectomized patients and tobacco related disorders, age, gender, and comorbidity indexes. In order to compare continuous variables a Student's t-test or a parametric equivalent was used. For qualitative variables a chi-squared test was used. Dose-effect relationships between comorbidity indexes were analyzed using Mantel-Haenszel method (Mantel and Haenszel 1959).

Multivariate logistic regression was performed to determine the effect of each comorbidity index on in-hospital mortality in cholecystectomized patients. Data were adjusted for age, gender, tobacco related disorders and group of hospitals.

The methods recommended by Hanley and McNeil were used to compare the area under the curve (AUC) of the receiving operative characteristic curves (ROC) for each multivariate predictive model (Hanley and McNeil 1982; Hanley and McNeil 1983). Alpha error was 10^{-6}. Algorithms and statistical analysis were performed using STATA version MP 13.1.

4 Results

4.1 Patient Characteristics

A total of 5,475,315 patients were identified and 83,231 cholecystectomies were performed, 27,038 laparoscopic cholecystectomies and 56,193 non laparoscopic cholecystectomies. Mean Charlson Comorbidity Index among cholecystectomized patients was 0.79 (95% CL 0.77-0.80). Mean Elixhauser Comorbidity Index among cholecystectomized patients was 1.07 (95% CL 1.06-1.08).

4.2 Crude Mortality Rates

Crude mortality rates by age and gender among cholecystectomized patients are exposed in Figure 1. Mortality increased in both genders with age. Male had higher mortality rates than females in all age group excepting in the group aged 85 years or more.

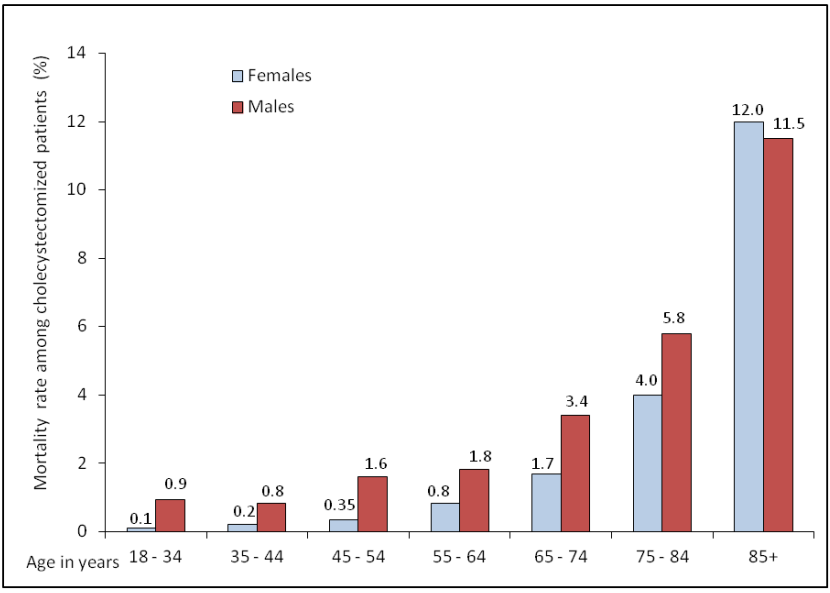

Fig. 1 Mortality rates among cholecystectomized patients in a sample of 87 Spanish hospitals during 2008-2010, by age and gender.

Dose-effect relationships between Charlson Comorbidity Index and mortality among cholecystectomized are exposed in Table 1. There is a direct dose-effect relationship between the value of Charlson Comorbidity Index and the adjusted Odds Ratio of death. Risk of mortality is higher among non laparoscopic cholecystecomized patients, but dose-effect relationships are directly related both in laparoscopic and non laparoscopic cholecystectomies.

ROC curves were developed for each multivariate model and the AUC were measured and compared between models. The results are shown in Figure 2. These show that the model including Elixhauser Comorbidity Index best explained mortality among cholecystectomized patients (AUC=0.8771), followed by the model that incorporated the Charlson comorbidity index (AUC=0.8717), but the differences were statistically non significant ($p= 0.0002$, $>10^{-6}$).

Table 1 Dose-effect relationships between Charlson Comorbidity Index (CCI) and mortality among patients cholecystectomized with laparoscopic cholecystectomy, non-laparoscopic cholecystectomy and any type of cholecystectomy.

Charlson Comorbidity Index	Adjusted Odds Ratio	95% CL	p	p for trend
Laparoscopic cholecystectomy[1]				
0	1	-	-	$<10^{-6}$
1	1.6	1.4 – 1.9	$<10^{-6}$	
2	3.1	2.6 – 3.6	$<10^{-6}$	
3	4.5	3.7 – 5.5	$<10^{-6}$	
4	9.3	7.1 – 12.2	$<10^{-6}$	
5+	14.5	9.2 – 23.0	$<10^{-6}$	
Non laparoscopic cholecystectomy[1]				
0	1	-	-	$<10^{-6}$
1	2.1	1.4 – 3.0	$<10^{-6}$	
2	12.9	8.6 – 19.1	$<10^{-6}$	
3	15.2	8.9 – 26.0	$<10^{-6}$	
4	43.3	16.8 – 111.5	$<10^{-6}$	
5+	59.8	13.1 – 272.0	$<10^{-6}$	
Any type of cholecystectomy[2]				
0	1	-	-	$<10^{-6}$
1	1.6	1.4 – 1.9	$<10^{-6}$	
2	3.6	3.1 – 4.2	$<10^{-6}$	
3	4.9	4.1 – 5.9	$<10^{-6}$	
4	9.8	7.5 – 12.7	$<10^{-6}$	
5+	15.5	10.0 – 24.1	$<10^{-6}$	

[1]Adjusted for age, gender and group of hospital

[2]Adjusted for age, gender, group of hospital and type of cholecystectomy

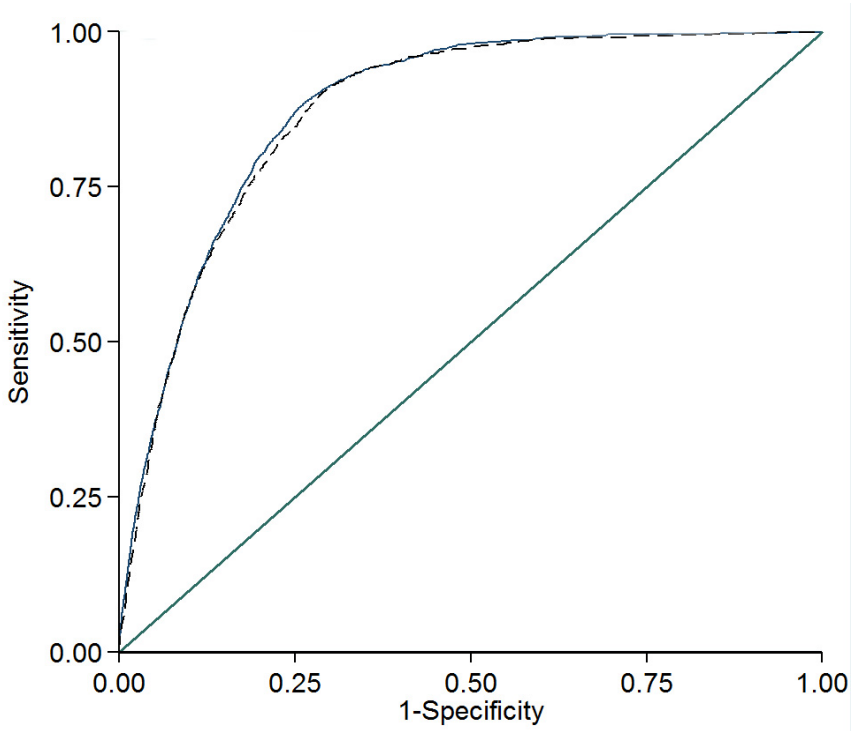

Fig. 2 ROC curves for multivariate models predicting in-hospital mortality among cholecystectomized patients. Both models included age, gender, tobacco related disorders and group of hospitals, but one of them incorporated ECI (_____) and the other CCI (- - - -). AUC for the model with ECI was 0.8771 and AUC for the model with CCI was 0.817 ($p>10^{-6}$).

5 Discussion

5.1 Limitations and Strengths of the Study

Our study has several limitations. The data used were only those found in the MBDS and were not complemented with additional data from patients. We used clinical definitions of celiac disease, alcohol and tobacco related disorders and comorbidities as assigned by physicians and entered by codifiers. Another limitation is potential underreporting if all of the information required to complete such codes is not available in the clinical record, or due to variations in the codifiers interpretation of data. The ICD-9-CM does not include codes specific to the motive for hospitalizing patients, and in some cases this would make an interesting variable to study.

Databases such as the MBDS also have clear advantages. The data collected are usually completed in most hospital admissions and, as they include virtually all cases, provide reasonably accurate estimates of the incidence, prevalence, comorbidity, and mortality of diseases treated in the hospital setting. The data can be analyzed retrospectively, unlike other designs that require prospective data collection, and data from long periods or from large numbers of patients can be gathered quickly and easily. Since the data are collected systematically, costs are considerably reduced. In studies using these databases, there may be less selection bias due to the refusal of patients or their legal representatives to sign consent forms which would allow the patient to participate in the study. Alpha error = 10^{-6} increases the validity of results.

Because of our large sample size and diversity of hospitals, our data are generalizable and not limited to patients admitted to a small number of centers. The availability of costs of hospitalization for each DRG, stratified by hospital group and specific to each year, facilitates the calculation of excess costs due to prolonged stays attributed to different risk exposures (tobacco, alcohol, and the like).

5.2 Other Applications

Chen et al, within a large, national registry, identified hospitals with at least 50 adult in-hospital cardiac arrest cases between January 1, 2000, and November 30, 2009. They used multivariable hierarchical regression to evaluate the correlation between a hospital's cardiac arrest incidence rate and its case-survival rate after adjusting for patient and hospital characteristics. Of 102,153 cases at 358 hospitals, the median hospital cardiac arrest incidence rate was 4.02 per 1000 admissions and the median hospital case-survival rate was 18.8%. In crude analyses, hospitals with higher case-survival rates also had lower cardiac arrest incidence (r, −0.16; P = 0.003). This relationship persisted after adjusting for patient characteristics (r, −0.15; P = 0.004). After adjusting for potential mediators of this relationship (ie, hospital characteristics), the relationship between incidence and case survival was attenuated (r, −0.07; P = 0.18). The one modifiable hospital factor that most attenuated this relationship was a hospital's nurse-to-bed ratio (r, −0.12; P = 0.03). Hospitals with exceptional rates of survival for in-hospital cardiac arrest are also better at preventing cardiac arrests, even after adjusting for patient case mix (Chen et al. 2013).

Ebell et al, developed a simple pre-arrest point score that can identify patients unlikely to survive in-hospital cardiac arrest (IHCA), neurologically intact or with minimal deficits. The study included 51,240 in patients experiencing an index episode of IHCA between January 1, 2007, and December 31, 2009, in 366 hospitals participating in the Get With the Guidelines–Resuscitation registry. Dividing data into training (44.4%), test (22.2%), and validation (33.4%) data sets, they used multivariate methods to select the best independent predictors of good neurologic outcome, created a series of candidate decision models, and used the test data set to select the model that best classified patients as having a very low (<1%), low

(1%-3%), average (>3%-15%), or higher than average (>15%) likelihood of survival after in-hospital cardiopulmonary resuscitation for IHCA with good neurologic status. The final model was evaluated using the validation data set. The best performing model was a simple point score based on 13 pre-arrest variables. The C statistic was 0.78 when applied to the validation set. It identified the likelihood of a good outcome as very low in 9.4%of patients (good outcome in 0.9%), low in 18.9% (good outcome in 1.7%), average in 54.0% (good outcome in 9.4%), and above average in 17.7%(good outcome in 27.5%). Overall, the score identified more than one-quarter of patients as having a low or very low likelihood of survival to discharge, neurologically intact or with minimal deficits after IHCA (good outcome in 1.4%) (Ebell et al.2013).

Sharifpour et al, studied the incidence, predictors, and outcomes of perioperative stroke after noncarotid major vascular surgery using the American College of Surgeons National Quality Improvement Program database. 47,750 patients undergoing noncarotid vascular surgery from 2005 to 2009 at non- Veterans Administration hospitals were identified from the American College of Surgeons National Quality Improvement Program database. An analysis of patients undergoing elective lower extremity amputation, lower extremity revascularization, or open aortic procedures was performed to determine the incidence, independent predictors, and 30-day mortality of perioperative stroke. The overall incidence of perioperative stroke within 30 days of surgery was 0.6%. Multivariate analysis revealed that each 1-year increase in age [odds ratio 1.02, 95% confidence interval (CI) (1.01 to 1.04)], cardiac history [1.42, (1.07 to 1.87)], female sex [1.47, (1.12 to 1.93)], history of cerebrovascular disease [1.72, (1.29 to 2.29)], and acute renal failure or dialysis dependence [2.03, (1.39 to 2.97)] were independent predictors of stroke. Only 15% (95% CI, 11%–20%) of strokes occurred on postoperative day 0 or 1. Perioperative stroke was associated with a 3-fold increase in 30-day all- cause mortality [3.36, (1.77 to 6.36)] and an increased median surgical length of stay from 6 (95% CI, 2 to 28) to 13 (95% CI, 3 to 43) days ($P < 0.001$, in a matched-cohort assessment. Perioperative stroke was an important source of morbidity and mortality, as reflected by significant increases in median surgical length of stay and all-cause 30-day mortality. The independent predictors of stroke that they identified in this population are not readily modifiable and the majority of strokes occurred after the first postoperative day (Sharifpour et al. 2013).

Neuman et al., examined a retrospective cohort of patients undergoing surgery for hip fracture in 126 hospitals in New York in 2007 and 2008. They tested the association of a record indicating receipt of regional versus general anesthesia with a primary outcome of inpatient mortality and with secondary outcomes of pulmonary and cardiovascular complications using hospital fixed-effects logistic regressions. Subgroup analyses tested the association of anesthesia type and outcomes according to fracture anatomy. Of 18,158 patients, 5,254 (29%) received regional anesthesia. In-hospital mortality occurred in 435 (2.4%). Unadjusted rates of mortality and cardiovascular complications did not differ by anesthesia type. Patients receiving regional anesthesia experienced fewer pulmonary complications (359 [6.8%] vs. 1,040 [8.1%], $P <0.005$). Regional anesthesia was associated with

a lower adjusted odds of mortality (odds ratio: 0.710, 95% CI0.541, 0.932, $P<$ 0.014) and pulmonary complications (odds ratio: 0.752, 95% CI 0.637, 0.887, P <0.0001) relative to general anesthesia. In subgroup analyses, regional anesthesia was associated with improved survival and fewer pulmonary complications among patients with intertrochanteric fractures but not among patients with femoral neck fractures. Regional anesthesia was associated with a lower odds of inpatient mortality and pulmonary complications among all hip fracture patients compared with general anesthesia (Neuman et al.2012).

Mathis et al, performed a retrospective database review of all pediatric patients who received a laryngeal mask anesthetic at their institution from 2006 to 2010. Device brands were restricted to LMA Unique™ and LMA Classic™, and primary outcome was laryngeal mask failure, defined as any airway event requiring device removal and tracheal intubation. Potential risk factors were analyzed with both univariate and multivariate techniques and included medical history, physical examination, surgical, and anesthetic characteristics. Of the 11,910 anesthesia cases performed in the study, 102 cases (0.86%) experienced laryngeal mask failure. Common presenting features of laryngeal mask failures included leak (25%), obstruction (48%), and patient intolerance such as intractable coughing/bucking (11%). Failures occurred before incision in 57% of cases and after incision in 43%. Independent clinical associations included ear/nose/ throat surgical procedure, non outpatient admission status, prolonged surgical duration, congenital/acquired airway abnormality, and patient transport. The findings of the study supported the use of the LMA Unique™ and LMA Classic™ as reliable pediatric supraglottic airway devices, demonstrating relatively low failure rates. Predictors of laryngeal mask airway failure in the pediatric surgical population did not overlap with those in the adult population and should therefore be independently considered (Mathis et al. 2013).

Gili et al, assessed the impact of alcohol use disorders (AUD) on the risk of healthcare associated infections (HCAI), length of hospital stay, hospital costs and mortality among surgical patients, using a large administrative database (MBDS). analyzing these adverse events in patients who underwent elective surgery in a sample of 87 Spanish hospitals during the period 2008-2010. HCAI were defined as health care-associated severe sepsis, non-severe sepsis, pneumonia, surgical site infections (SSI), vascular catheter associated infections and urinary tract infections. Primary outcomes were risk of HCAI in patients with AUD. Secondary outcomes were mortality, excess length of stay and over-expenditure. There were 1,511,899 inpatient admissions analyzed; 43,484 (2.9%) cases of HCAI were identified, and SSI was the most frequent HCAI. AUD was diagnosed in 39,226 patients (2.6%) and in 2,587 HCAI cases. Multivariable analysis demonstrated that AUD was an independent predictor of developing every type of HCAI (Odds Ratio: 1.5; 95% CL 1.4-1.6; p<0.0001). Patients with AUD who developed HCAI had a mean 2.1 day longer hospital length of stay. The excess cost for stay among patients with AUD who developed HCAI was 1,871.1 Euros. A significant association between AUD and death from SSI infections was found with multivariable analysis (OR 1.4, 95% CL 1.1-1.8, p: 0.004) but this association was not found

with the rest of HCAIs. Among surgical patients, AUD increase the risk of HCAI, heighten the risk of in-hospital death in SSI, and cause longer hospital stays and over-expenditures (Gili et al. 2013).

Hlatky et al assessed whether clinical characteristics modify the comparative effectiveness of coronary artery bypass graft (CABG) versus percutaneous coronary intervention (PCI) in an unselected, general patient population. The study design was an observational treatment comparison using propensity score matching and Cox proportional hazards models. The setting was United States, during the period 1992 to 2008. Patients were Medicare beneficiaries aged 66 years or older. They measured CABG–PCI hazard ratio (HR) for all-cause mortality, with prespecified treatment-by-covariate interaction tests, and the absolute difference in life-years of survival in clinical subgroups after CABG or PCI, both over 5 years of follow-up. Among 105,156 propensity score–matched patients, CABG was associated with lower mortality than PCI (HR, 0.92[95% CI, 0.90 to 0.95]; $P <$ 0.001). Association of CABG with lower mortality was significantly greater (interaction P =0.002 for each) among patients with diabetes (HR, 0.88), a history of tobacco use (HR, 0.82), heart failure (HR, 0.84), and peripheral arterial disease (HR, 0.85). The overall predicted difference in survival between CABG and PCI treatment over 5 years was 0.053 life-years (range, -0.017 to 0.579 life-years). Patients with diabetes, heart failure, peripheral arterial disease, or tobacco use had the largest predicted differences in survival after CABG, whereas those with none of these factors had slightly better survival after PCI.

One limitation was that treatments were chosen by patients and physicians rather than being randomly assigned. Multivessel CABG was associated with lower long-term mortality than multivessel PCI in the community setting. This association was substantially modified by patient characteristics, with improvement in survival concentrated among patients with diabetes, tobacco use, heart failure, or peripheral arterial disease (Hlatky et al.2013).

de Wit et al, analyzed the risk of healthcare associated infections (HAI) and surgical site infections (SSI) in surgical patients undergoing elective inpatient joint replacement, coronary artery bypass grafting, laparoscopic cholecystectomy, colectomy, and hernia repair. They obtained data from the Nationwide Inpatient Sample for the years 2007 and 2008. HAI were defined as health care-associated pneumonia, sepsis, SSI, and urinary tract infection. Primary outcomes were risk of HAI and SSI in patients with alcohol use disorders (AUD). Secondary outcomes were mortality and hospital length of stay in patients with HAI and SSI, alpha= 10^{-6}. There were 1,275,034 inpatient admissions analyzed; 38,335 (3.0%) cases of HAI were documented, and 5,756 (0.5%) cases of SSI were identified. AUD was diagnosed in 11,640 (0.9%) of cases. Multivariable analysis demonstrated that AUD was an independent predictor of developing HAI: odds ratio (OR) 1.70, p $<10^{-6}$, and this risk was independent of type of surgery. By multivariable analysis, the risk of SSI in patients with AUD was also higher: OR 2.73, p $<10^{-6}$. Hospital mortality in patients with HAI or SSI was not affected by AUD. However, hospital length of stay was longer in patients with HAI who had AUD (multivariable analysis 2.4 days longer, p $<10^{-6}$). Among patients with SSI, those with AUD did

not have longer hospital length of stay. The conclusion was that patients with AUD who undergo a variety of elective operations have an increased risk of infectious postoperative morbidity (de Wit et al. 2012).

Zhan and Miller analyzed the impact of medical injuries in the health care system assessing the excess length of hospital stay, charges, and deaths attributable to medical injuries during hospitalization. They used the Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSIs) to identify medical injuries in 7.45 million hospital discharge abstracts from 994 acute-care hospitals across 28 states of the USA in 2,000 in the AHRQ Healthcare Costs and Utilization Project Nationwide Inpatient Sample database. The main outcome measures were length of hospital stay, charges, and in-hospital mortality that were recorded in hospital discharge abstract and were attributable to medical injuries according to 18 PSIs. Excess length of stay attributable to medical injuries ranged from 0 days for injury to a neonate to 10.89 days for postoperative sepsis, excess charges ranged from 0 US$ for obstetric trauma (without vaginal instrumentation) to US$ 57,727 for postoperative sepsis, and excess mortality ranged 0% for obstetric trauma to 21.96% for postoperative sepsis ($P<0.001$). Following postoperative sepsis, the second most serious event was postoperative wound dehiscence, with 9.42 extra days in the hospital, US$ 40,323 in excess charges, and 9.63% attributtable mortality. Infection due to medical care was associated with 9.58 extra days, US$ 38,656 in excess charges, and 4.31% attributtable mortality (Zhan and Miller 2003).

5.3 Strengths of Large Databases

Of the strengths of administrative data, population coverage is perhaps the most important. Without such population base, inference that can be drawn from the data is limited. Medicare provides health care for approximately 38 million persons, 30 million of whom are age 65 years and older. It has been estimated that over 96% of the elderly in the United States receive health care through the Medicare program. A related strength of administrative data is the slightly less dramatic observation that the files tend to include a large number of observations even if the high degree of population coverage is lacking.

A major drawback to primary data collection is the cost of such efforts. Although administrative data may be limited with respect to the number of data elements they contain, they are much more efficient to obtain. In general, the delay between patient contact and data availability is shorter than when using manually abstracted charts. Furthermore, because personnel costs are lower, obtaining the data is a much less expensive prospect. The timeliness and lower costs are the basis for some researchers to suggest that administrative data are a preferable data source for long-term follow-up of cohorts.

The widespread use of such standard identifiers as social security number, codes numbers as well as more obvious identifying information, such as name and date of birth, make it possible to link administrative data with other sources. In the

nineties some authors described methods for linking Medicare data with SEER cancer registries and VA utilization data (Potosky et al. 1993; Fleming and Fisher 1992).

Such combinations allow for the administrative data to be augmented by data from other sources and increase the usefulness of both sources for research and planning. The flexibility and efficiency associated with the ability to combine administrative data and data from other sources serves to extend the applications of both sources of data

Beyond linking administrative data with individual-level data, administrative data can also be combined with group-level data. Some researchers have used household income estimated from the census to overcome the lack of income information available from Medicare sources (Krieger 1992; Gornick et al.1996).

Despite the clear strengths of administrative data outlined above, there are also some major weaknesses associated with the data source. For some researchers, the fact that they are a "secondary" source of data implies that data were not generated for the specific purpose for which they are used and their internal validity is under suspicion.

5.4 Validity of Large Databases

Several factors influence the validity of large databases as a source of information for perioperative effectiveness research, among them the data quality and depth of information available, control of biases, control of missing data and alpha error (type I error).

Studies on the accuracy of administrative large databases compared with other data sources, such as medical records, generally indicate a high level of validity (Virnig 2001).

The population included by such databases must be evaluated to ensure generalizability to the clinical population in question. Many administrative data are not collected with research objectives in mind, and therefore their appropriateness for specific research questions needs to be judged on a case-by-case basis.

In general, reimbursed procedures are likely to be accurately coded, while the coding of comorbidities and complications (other than death) may be less reliable. For example, the concordance between Medicare claims data and the SEER registry data is generally higher for extensive (i.e., highly reimbursed) surgical procedures than for biopsies and nonsurgical therapy (Cooper et al. 2002). Likewise, the sensitivity of administrative data for identifying complications such as wound infections or hemorrhage following elective lumbar diskectomy was found to be only 35%, but reoperation (which would carry a high reimbursement) was coded with 100% sensitivity (Romano et al. 2002). In a study of knee replacement surgery, administrative data were found to be 95.5% sensitive for identifying procedures, but only 27% for comorbidities and 66% for complications (Hawker et al.1997).

Services reimbursed at a low rate may be inconsistently recorded in administrative databases, even if they were in fact provided. Laboratory values are not usually available from administrative data. For example, the Nationwide Inpatient Sample in the USA or MBDS is Spain which consists of hospital-level discharge data collected on a national basis, includes information on inpatient stays but no information on outpatient procedures.

Whereas external validity (generalizability) of a study using population-based data can be threatened if the study population does not accurately represent the clinical population of interest, internal validity can be threatened by biases. Internal validity specifically refers to the degree to which a study's measure of effect accurately represents the ''truth.'' Bias is any systematic trend in the collection, analysis, interpretation, publication, or review of data that can lead to invalid conclusions.

The three main types of bias are selection bias, information bias and confounding bias (Delgado and Llorca 2004). Selection bias refers to the representativeness of participants in every group and results when inclusion in a study or a particular study group is affected by factors that also affect the outcome. Of particular relevance to surgical studies is treatment selection bias. Information bias results when problems in the quality or type of data measurement lead to subsequent misclassification (cases are classified as non-cases, exposed are classified as non-exposed, or the reverse) (Copeland et al. 1997).

Confounding bias results when a factor that affects the outcome of interest is unevenly distributed among study groups. A variable may introduce confounding bias only if it exhibits three characteristics. First, it must be a risk factor for the outcome of interest. Second, it must be associated with the exposure of primary interest. Finally, it must not be an intermediate factor between the exposure and the outcome of interest in the causal link. Methods for controlling bias during the analysis of results should be employed when bias cannot be eliminated at the study design stage.

Of special concern when evaluating efficacy of therapies using administrative and registry data is nonrandom treatment assignment, which may introduce selection bias. One example of selection bias would be systematically offering surgical therapy only to relatively healthy patients or those with localized disease. Such clinically reasonable decisions can present problems when inferences are to be drawn from observational data, as differences in outcome may be due to selection criteria, the treatment itself, or both. To deal with this, multivariate regression methods that include treatment type as a dummy variable and adjust for other covariates are commonly used.

Multivariate regression methods are appropriate when the distributions of the covariates included in the model overlap sufficiently between study groups. When treatment groups differ substantially with respect to the distribution of covariates reliance on standard regression models may be dangerous.

Largely due to these shortcomings, the use of propensity score analysis to account for nonrandom treatment assignment has become increasingly popular

(Rubin 1997; Rosenbaum and Rubin 1983; Rosenbaum and Rubin 1984; Rosenbaum and Rubin 1985; Braitman 2002).

Propensity score methodology determines the hypothetical probability of an individual patient having received a certain treatment (e.g., surgery) as a function of all confounding covariates, collapsed into a single numerical score. The propensity score is generated using a potentially exhaustive set of available covariates, thereby allowing more complete control for treatment selection bias than conventional multivariate regression methods alone. Propensity score methodology can be especially useful when a treatment is common but the outcome of interest is rare. The propensity score can then be used to stratify or match subjects for further analysis. Stratifying or matching patients based on their individual propensity scores greatly facilitates the construction and comparison of treated and control groups with comparable distributions of measured covariates. Propensity score techniques can only reliably account for measured determinants of treatment selection.

Nevertheless, some studies have found that propensity scores developed using administrative data did not necessarily balance patient clinical characteristics, and the resulting analyses tended to overestimate treatment effects (Austin et al.2005).

Observational studies carry a higher risk of drawing invalid inferences due to incomplete control of confounding. Although multivariate regression methods attempt to adjust for measured confounders, the risk of bias due to unmeasured confounders remains. The potential for confounding may be even more problematic when employing administrative databases because these data often lack clinical depth and the values of potentially important confounders may be unknown.

Most administrative databases and registry data sets have some degree of missing data, and this is a serious problem in epidemiologic research involving human populations (Allison 2001).

In a longitudinal study, subjects may drop out, be unable, or refuse to participate in subsequent waves of data collection, generating non response bias, a frequent cause of selection bias in medical research (Kleinbaum, Morgenstern and Kupper 1981).

The usual approach is to restrict the analysis to subjects with no missing values in the specific set of variables. This so-called available-case analysis can yield biased estimates. Sometimes, where multiple analyses are involved, this approach excludes subjects with any missing values in any of the variables used in at least one analysis (the so-called complete-case or list wise deletion analysis). Other approaches such as treating the missing data as a separate category also result in biased estimates (Vach 1994).

Listwise deletion or complete-case analysis is the default method in most statistical software packages. However, the use of these methods can result in biased estimates unless missingness is completely random (Little and Rubin 2002).

Listwise deletion also reduces the available sample size and consequently results in loss of power. Relatively small numbers of missing values can result in the deletion of a large number of cases.

Another frequently used approach is to create a separate category for missing data (e.g., race unknown). This method can also result in biased estimates. If the amount of missing data is small, one reasonable approach is to perform list wise deletion followed by a sensitivity analysis to explore the potential impact of missing data on inferences.

There are three approaches for correctly analyzing incomplete data (Raghunathan 2004; Horton and Kleinman 2007). The first approach involves attaching weights to each subject included in the analysis to represent subjects who were excluded. In surveys, the weight for an individual is the inverse of his/her selection probability. Thus, the weighted average (where the weights are the inverse of their selection rate) is an unbiased estimate of the population mean whereas the simple mean is not. Weighting to compensate for non response is an extension of the same idea. That is, excluding the subjects because of missing values is a distortion of the representation in the original sample, and weights are attached to subjects included in the analysis to restore the representation. Weighting is a simple device to correct for bias but it still discards partial information from subjects with missing values.

A more appropriate and generally applicable method for handling missing data is multiple imputation (Rubin 2004; Schafer 1999).

Multiple imputation uses the available data to predict plausible values for missing data through the use of regression models. Missing data are then replaced with predicted, or imputed, values. By using multiple imputed data sets, the subsequent analyses appropriately consider both the uncertainty of the observed values and the uncertainty of the imputed values, thereby resulting in more valid inferences. Perhaps the most practical approach is based on multiple imputation. This approach involves an upfront investment in multiply imputing the missing values in the database. Once multiply imputed, any complete data software can be used to repeatedly analyze the completed data sets, extract the point estimates and their standard errors, and combine them (Reiter and Raghunathan 2007).

The third approach is based on the likelihood constructed from the observed incomplete data. In the complete data statistical methodology, maximum likelihood for a given model is a dominant inferential procedure, for example, the linear, logistic, Poisson, log-linear, and random effects models.

A type I error refers to the conclusion that a treatment is effective or a difference exists between two groups when in reality the treatment is not effective or no difference exists. Statistical software packages allow analyses to be performed with relative ease and this may lead to the temptation to engage in data mining, a non hypothesis driven search process for a statistically significant result. This practice carries a higher risk of type I error and spurious conclusions can result. All analyses of administrative data are essentially secondary data analyses and therefore are generally more susceptible to such statistical error.

Researchers may consider imposing a stricter standard for type I error than the conventional alpha = 0.05 as a means of greater insurance against false positive findings, as we did in this study, where alpha was stated as 10^{-6}. The Bonferroni correction divides the type I error by the number of tests and uses this value as the

threshold for statistical significance. For example, if 10 tests are conducted with alpha = 0.05, only P values less than .005 would be considered statistically significant.

The large sample sizes available in administrative data sets have the potential to demonstrate statistical significance even when very small absolute differences exist. Although the conventional threshold for statistical significance of $P < 0.05$ is widely used, one should keep in mind that this threshold is arbitrary. The P value has no meaning with regard to the magnitude and clinical significance of the observed effect, but rather reflects the precision of effect estimation. Excessive focus on a P value of less than 0.05 can exaggerate the importance of statistically significant, but clinically meaningless results. Likewise, this approach can discard potentially meaningful information gleaned from an analysis simply because the P value exceeds an arbitrary threshold.

Researchers should focus on estimates of effect (point estimates and confidence intervals). Confidence intervals quantify imprecision more clearly than P values do, allowing clinicians to assess the clinical significance of the range of plausible values for effect estimates.

Finally, associations derived from observational data alone not always imply causality, regardless of the magnitude of the observed effect or its statistical significance. The strength of an association is only one of several factors that should be evaluated to establish causality following classical Hill's rules (Hill 1965).

5.5 Cautions and Checklists

Research based on retrospective longitudinal databases is increasingly important to health care policy and clinical decision-making, which needs information on "real-world" practice, experience of large and diverse populations, and efficiency in information acquisition.

In an effort to assist decision makers in evaluating the quality of published studies that use health related retrospective databases, a checklist was developed by the International Society For Pharmacoeconomics and Outcomes Research (ISPOR) Task Force that focuses on issues that are unique to database studies or are particularly problematic in database research. This checklist was developed primarily for the commonly used medical databases but could potentially be used to assess retrospective studies that employ other types of databases (Motherol et al. 2003).

In particular, the Checklist calls attention to three of the most critical issues in evaluating the quality of the research: 1) Do the particular large administrative database used for this study adequately represent the intended clinical experience?; 2) Have linkages across data files been managed to assure accuracy of the linkages themselves and compatibility of the information across files? and 3) Have nuances and complexities of longitudinal data been appreciated, including the many changes that occur in a given health plan and in real patient experiences over

time? Some researchers have added additional recommendations to the meticulous ISPOR Task Force checklist (Andrews and Eaton 2003).

6 Conclusions

In a considerable number of situations, such as clinical trials and cohorts studying rare outcomes or case-control studies of rare exposures, a Big Data approach is mandatory to obtain precise estimates.

Despite the advantages associated with database research, researchers need to consider a number of limitations. Some of the largest databases were created for administrative purposes, thus, detailed clinical information, including documentation of perioperative events, is often missing. In this context, comorbidities and complications are usually based on the ICD-9-coding system with considerable risk for coding bias and the inability to discern the severity of diseases and events, and also information on whether the condition was present on admission. The latter has recently been addressed by a number of database administrators through inclusion of a "present on admission" variable as commented previously.

At present, many databases do not allow for the longitudinal analysis of patient outcomes as one cannot determine separate admissions for the same individual because no patient identifiers are available, and additional hospitalizations in a facility outside the hospital universe may not be captured. In this context, efforts to create databases with longitudinal data analysis capability are underway. Although large databases have been used for many decades for medical research, their increasing size and complexity demand continued advances in computer software and statistical methodologies to allow for appropriate manipulation and interpretation of data. The fields of biostatistics and epidemiology have expanded dramatically over the last few decades in order to provide means to satisfy the needs for methodological tools. Collaboration with biostatisticians and epidemiologists is an essential component of perioperative database analysis in order to achieve a rigorous scientific process in the fields of comparative effectiveness research, epidemiologic evaluations, risk factors analysis, and outcomes research.

Despite promising possibilities, researchers in the field of anaesthesiology and surgery need to be aware of the pitfalls of large database research and the need for ongoing methodology development in order to address the important questions posed by perioperative research in a rigorous and robust scientific way.

Acknowledgments. This Research was funded by the Spanish Delegación del Gobierno para el Plan Nacional Sobre Drogas (DGPNSD) (grant number 2009I017, project G41825811). DGPNSD had no further role in study design; in the collection; analysis and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

References

1. Allison, P.D.: Missing Data. Sage, Thousand Oaks (2001)
2. Andrews, E.B., Eaton, S.: Additional considerations in longitudinal database research. Value Health 6(2), 85–87 (2003)
3. Austin, P.C., Mamdani, M.M., Stukel, T.A., et al.: The use of the propensity score for estimating treatment effects: administrative versus clinical data. Stat. Med. 24(10), 1563–1578 (2005)
4. Berger, M.L., Mamdani, M., Atkins, D., et al.: Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part I. Value Health 12(8), 1044–1052 (2009)
5. Braitman, L.E., Rosenbaum, P.R.: Rare outcomes, common treatments: analytic strategies using propensity scores. Ann. Intern. Med. 137(8), 693–695 (2002)
6. Charlson, M.E., Pompei, P., Ales, K.L., et al.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic. Dis. 40(5), 373–383 (1987)
7. Chen, L.M., Nallamothu, B.K., Spertus, J.A., et al.: Association Between a Hospital's Rate of Cardiac Arrest Incidence and Cardiac Arrest Survival. JAMA Intern. Med. 173(13), 1186–1194 (2013)
8. Chiolero, A.: Big Data in Epidemiology. Too Big to Fail? Epidemiology 24(6), 938–939 (2013)
9. Cleves, M.A., Sanchez, N., Draheim, M.: Evaluation of two competing methods for calculating Charlson's comorbidity index when analyzing short-term mortality using administrative data. J. Clin. Epidemiol. 50(8), 903–908 (1997)
10. Cooper, G.S., Virnig, B., Klabunde, C.N., et al.: Use of SEER-Medicare data for measuring cancer surgery. Med. Care 40(8, suppl. IV), 43–48 (2002)
11. Copeland, K.T., Checkoway, H., McMichael, A.J., et al.: Bias due to misclassification in the estimation of relative risk. Am. J. Epidemiol. 105(5), 488–495 (1997)
12. Copeland, G.P., Jones, D., Walters, M.: POSSUM: A scoring system for surgical audit. Br. J. Surg. 78(3), 355–360 (1991)
13. Cox, E., Martin, B.C., Van Staa, T., et al.: Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report–Part II. Value Health 12(8), 1053–1061 (2009)
14. de Wit, M., Goldberg, S., Hussein, E., et al.: Health care-associated infections in surgical patients undergoing elective surgery: are alcohol use disorders a risk factor? J. Am. Coll. Surg. 215(2), 229–236 (2012)
15. Delgado-Rodríguez, M., Llorca, J.: Bias. J. Epidemiol. Commun. Health 58(3), 635–641 (2004)
16. Deyo, R.A., Cherkin, D.C., Ciol, M.A.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol. 45(6), 613–619 (1992)

17. Ebell, M.H., Jang, W., Shen, Y., et al.: Development and Validation of the Good Outcome Following Attempted Resuscitation (GO-FAR) Score to Predict Neurologically Intact Survival After In-Hospital Cardiopulmonary Resuscitation. JAMA Intern. Med. 173(20), 1872–1878 (2013)
18. Elixhauser, A., Steiner, C., Harris, D.R., et al.: Comorbidity measures for use with administrative data. Med. Care 36(1), 8–27 (1998)
19. Evley, R., Russell, J., Mathew, D., et al.: Confirming the drugs administered during anaesthesia: a feasibility study in the pilot National Health Service sites, UK. Br. J. Anaesth. 105(3), 289–296 (2010)
20. Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and Congress. U.S. Department of Health and Human Services, Washington, DC (2009), http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/ (accesed on October 10, 2013)
21. Flemming, C., Fisher, E.S., Chang, C.H., et al.: Studying outcomes and hospital utilization in the elderly: the advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Med. Care 30(5), 377–391 (1992)
22. Freundlich, R.E., Kheterpal, S.: Perioperative effectiveness research using large databases. Best Pract. Res. Clin. Anaesthesiol. 25(4), 489–498 (2011)
23. Fuller, G., Bouamra, O., Woodford, M., et al.: The Effect of Specialist Neurosciences Care on Outcome in Adult Severe Head Injury: A Cohort Study. J. Neurosurg. Anesthesiol. 23(3), 198–205 (2011)
24. Ghali, W.A., Hall, R.E., Rosen, A.K., et al.: Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J. Clin. Epidemiol. 49(3), 273–278 (1996)
25. Gili, M., Sala, J., López, J., et al.: Impact of Comorbidities on In-Hospital Mortality From Acute Myocardial Infarction, 2003-2009. Rev. Esp. Cardiol. 64(12), 1130–1137 (2011)
26. Gili, M., Ramírez, G., López, J., et al.: Alcohol use disorders, healthcare associated infections, hospital stay, over-expenditures and mortality among surgical inpatients of a sample of 87 Spanish Hospitals. Gac San 27(suppl. 2), 163–164 (2013)
27. Gornick, M.E., Eggers, P.W., Reilly, T.W., et al.: Effects of race and income on mortality and use of services among Medicare beneficiaries. N. Engl. J. Med. 335(11), 791–799 (1996)
28. Hanley, J., McNeil, B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1), 29–36 (1982)
29. Hanley, J., McNeil, B.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148(3), 839–843 (1983)
30. Hawker, G.A., Coyte, P.C., Wright, J.G., Paul, J.E., Bombardier, C.: Accuracy of administrative data for assessing outcomes after knee replacement surgery. J. Clin. Epidemiol. 50(3), 265–273 (1997)
31. Hill, A.: The environment and disease: association or causation? Proc. R. Soc. Med. 58, 295–300 (1965)
32. Hlatky, M., Boothroyd, M.D., Baker, D.B.,, L., et al.: Comparative Effectiveness of Multivessel Coronary Bypass Surgery and Multivessel Percutaneous Coronary Intervention. A Cohort Study. Ann. Intern. Med. 158(10), 727–734 (2013)
33. Horton, N.J., Kleinman, K.: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Amer. Stat. 61(1), 79–90 (2007)

34. Johnson, M.L., Crown, W., Martin, B.C., et al.: Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part III. Value Health 12(8), 1062–1073 (2009)
35. Klabunde, C.N., Legler, J.M., Warren, J.L., et al.: A Refined comorbidity measurement algorithm for claims based studies of breast, prostate, colorectal, and lung cancer patients. Ann. Epidemiol. 17(8), 584–590 (2007)
36. Kheterpal, S.: Perioperative comparative effectiveness research: an opportunity calling. Anesthesiology 111(6), 1180–1182 (2009)
37. Klabunde, C.N., Warren, J.L., Legler, J.: Assessing comorbidity using claims data: an overview. Med. Care 40(8, suppl. IV), 26–35 (2002)
38. Kleinbaum, D.G., Morgenstern, H., Kupper, L.: Selection bias in epidemiological studies. Am. J. Epidem. 113(4), 452–463 (1981)
39. Krieger, N.: Overcoming the absence of socioeconomic data in medical records: validation and application of a census based methodology. Am. J. Public Health 82(5), 703–710 (1992)
40. Lee, T.H., Marcantonio, E.R., Mangione, C.M., et al.: Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation 100(10), 1043–1049 (1999)
41. Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data, 2nd edn. Wiley, Hoboken (2002)
42. Malenka, D.J., McLerran, D., Roos, N., et al.: Using administrative data to describe casemix: a comparison with the medical record. J. Clin. Epidemiol. 47(9), 1027–1032 (1994)
43. Manchikanti, L., Falco, F.J., Boswell, M.V., et al.: Facts, fallacies, and politics of comparative effectiveness research: Part I. Basic considerations. Pain Physician 13(1), E23–E54 (2010)
44. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22(4), 719–748 (1959)
45. Martins, M., Blais, R.: Evaluation of comorbidity indices for inpatient mortality prediction models. J. Clin. Epidemiol. 59(7), 665–669 (2006)
46. Mathis, M.R., Haydar, B., Taylor, E.L., et al.: Failure of the Laryngeal Mask Airway Unique™ and Classic™ in the Pediatric Surgical Patient. A Study of Clinical Predictors and Outcomes. Anesthesiology 119(6), 1284–1295 (2013)
47. Memtsoudis, S.G., Ma, Y., Gonzalez Della Valle, A., et al.: Perioperative outcomes after unilateral and bilateral total knee arthroplasty. Anesthesiology 111(6), 1206–1216 (2009)
48. Memtsoudis, S.G., Besculides, M.C.: Perioperative comparative effectiveness research. Best Pract. Res. Clin. Anaesthesiology 25(4), 489–498 (2011)
49. Memtsoudis, S.G., Ma, Y., Swamidoss, C.P., et al.: Factors influencing unexpected disposition after orthopedic ambulatory surgery. J. Clin. Anesth. 24(2), 89–95 (2012)
50. Ministerio de Sanidad, Servicios Sociales e Igualdad. Registro de Altas de los Hospitales Generales del Sistema Nacional de Salud. CMBD. Norma Estatal (2011), `http://www.msc.es/estadEstudios/estadisticas/cmbd.htm` (accessed on November 14, 2013)
51. Moonesinghe, S.R., Mythen, M.G., Das, P., et al.: Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: Qualitative systematic review. Anesthesiology 119(4), 959–981 (2013)

52. Motherol, B., Brooks, J., Clark, M.A., et al.: A checklist for retrospective database studies-Report of the ISPOR Task Force on retrospective databases. Value Health 6(2), 90–97 (2003)
53. Myles, P.S., Peyton, P., Silbert, B., et al.: Perioperative epidural analgesia for major abdominal surgery for cancer and recurrence-free survival: randomised trial. BMJ 342, d1491 (2011)
54. Nashef, S.A., Roques, F., Michel, P., et al.: European system for cardiac operative risk evaluation (EuroSCORE). Eur. J. Cardio-thorac. Surg. 16(1), 9–13 (1999)
55. Neuman, M.D., Silber, J.H., Elkassabany, N.M., et al.: Comparative effectiveness of regional versus general anesthesia for hip fracture surgery in adults. Anesthesiology 117(1), 72–92 (2012)
56. Pine, M., Jordan, H.S., Elixhauser, A., et al.: Enhancement of claims data to improve risk adjustment of hospital mortality. JAMA 297(1), 71–76 (2007)
57. Platell, C.: Secrets to a successful database. ANZ J. Surg. 78(9), 729–730 (2008)
58. Potosky, A.L., Riley, G.F., Lubitz, J.D., et al.: Potential for cancer related health services research using a linked Medicare tumor registry database. Med. Care 31(8), 732–748 (1993)
59. Powell, E.S., Cook, D., Pearce, A.C., et al.: A prospective, multicentre, observational cohort study of analgesia and outcome after pneumonectomy. Br. J. Anaesth. 106(3), 364–370 (2011)
60. Quan, H., Parsons, G.A., Ghali, W.: Validity of information on comorbidity derived from ICD-9-CCM administrative data. Med. Care 40(8), 675–685
61. Quan, H., Sundararajan, V., Halfon, P., et al.: Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med. Care 43(11), 1130–1139 (2005)
62. Raghunathan, T.E.: What do we do with missing data? Some options for analysis of incomplete data. Annu. Rev. Public Health 22, 99–117 (2004)
63. Reiter, J.P., Raghunathan, T.E.: The multiple adaptations of multiple imputation. J. Amer. Stat. Assoc. 102, 1462–1471 (2007)
64. Romano, P.S., Roos, L.L., Jollis, J.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J. Clin. Epidemiol. 46(10), 1075–1079 (1993)
65. Romano, P.S., Chan, B.K., Schembri, M.E., et al.: Can administrative data be used to compare postoperative complication rates across hospitals? Med. Care 40(10), 856–867 (2002)
66. Roos, N.P., Black, C., Froehlich, N., et al.: Population health and health care use: an information system for policy makers. Mil-bank Q. 74(1), 3–31 (1996)
67. Rosenbaum, P.R., Rubin, D.: The central role of the propensity score in observational studies for causal effects. Biometrika 70(1), 41–55 (1983)
68. Rosenbaum, P.R., Rubin, D.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. 79, 516–524 (1984)
69. Rosenbaum, P.R., Rubin, D.: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat. 39, 33–38 (1985)
70. Rosero, E.B., Adesanya, A.O., Timaran, C.H., et al.: Trends and outcomes of malignant hyperthermia in the United States, 2000 to 2005. Anesthesiology. Am. Soc. Anesthesiol. 110(1), 89–94 (2009)
71. Rubin, D.: Estimating causal effects from large data sets using propensity scores. Ann. Intern. Med. 127(8, Pt. 2), 757–763 (1997)

72. Rubin, D.B.: Multiple imputation for non response in surveys. Wiley-Interscience, Hoboken (2004)
73. Saklad, M.: Grading of patients for surgical procedures. Anesthesiology 2(3), 281–284 (1941)
74. Schafer, J.L.: Multiple imputation: a primer. Stat. Methods Med. Res. 8(1), 3–15 (1999)
75. Servicio Andaluz de Salud.: Manual de instrucciones del conjunto mínimo básico de datos de Andalucıa. Consejerıa de Salud de la Junta de Andalucıa, Sevilla (2012), http://www.juntadeandalucia.es/servicioandaluzdesalud (accesed on October 12, 2013)
76. Sharifpour, M., Moore, L.E., Shanks, A.M., et al.: Incidence, predictors, and outcomes of perioperative stroke in noncarotid major vascular surgery. Anesth. Analg. 116(2), 424–434 (2013)
77. Southern, D.A., Quan, H., Ghali, W.: Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. Med. Care 42(4), 355–360 (2004)
78. Stundner, O., Chiu, Y.L., Sun, X., et al.: Comparative perioperative outcomes associated with neuraxial versus general anesthesia for simultaneous bilateral total knee arthroplasty. Reg. Anesth. Pain Med. 37(6), 638–644 (2012)
79. Toh, S., Platt, R.: Is size the next big thing in epidemiology? Epidemiology 24(3), 349–351 (2013)
80. Tunis, S.R., Stryer, D.B., Clancy, C.M.: Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 290(12), 1624–1632 (2003)
81. Turan, A., Mascha, E.J., Roberman, D., et al.: Smoking and perioperative outcomes. Anesthesiology 114(4), 837–846 (2011)
82. Vach, W.: Logistic Regression with Missing Values in Covariates. Springer, New York (1994)
83. Virnig, B.A., McBean, M.: Administrative data for Public Health Surveillance and Planning. Annu. Rev. Public Health 22, 213–230 (2001)
84. Weiss, E.S., Allen, J.G., Meguid, R.A., et al.: The impact of center volume on survival in lung transplantation: an analysis of more than 10,000 cases. Ann. Thorac. Surg. 88(4), 1062–1070 (2009)
85. White, C.M., Ip, S., McPheeters, M., et al.: Using Existing Systematic Reviews To Replace De Novo Processes in Conducting Comparative Effectiveness Reviews. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality, US (2009)
86. Zhan, C., Miller, M.R.: Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 290(14), 1868–1874 (2003)

Sickness Absence and Record Linkage Using Primary Healthcare, Hospital and Occupational Databases

Miguel Gili-Miner*, Juan Luís Cabanillas-Moruno, and Gloria Ramírez-Ramírez

Abstract. Objective: Charlson comorbidity index (CCI) has been adapted to primary care (PC) patients to determine chronic illness costs. We retrospectively evaluated its ability to predict sickness absence, hospital admissions and in-hospital mortality among 1,826,190 workers followed during the period 2007-2009.

Methods: The electronic administrative databases DIRAYA© and MBDS contain information of diseases and conditions of patients attended in primary care and hospital settings, respectively. We retrospectively used available information in the DIRAYA medical record database to calculate CCI adapted to PC (CCIPC), and analyzed its relation to sickness absence, hospital admissions and in-hospital mortality.

Results: The models including age, gender, province of residence, hospital size and CCIPC calculated in PC setting were predictive of every outcome: sick leave (their number and duration), hospital admissions (number and length of hospital stays) and in-hospital mortality, when measured in terms of adjusted Odds Ratios and 95% confidence limits. Area under the curve for ROC predictive models was maximal for in-hospital mortality (0.9254).

Conclusion: The adapted CCIPC was predictive of all outcomes related to sick leave, hospital admissions, and in-hospital mortality among a large sample of

Miguel Gili-Miner · Gloria Ramírez-Ramírez
Unit of Preventive Medicine, Surveillance and Health Promotion. Seville. Spain
e-mail: mgili@us.es

Miguel Gili-Miner · Gloria Ramírez-Ramírez · Juan Luís Cabanillas-Moruno
Department of Preventive Medicine and Public Health. University of Seville. Spain

Juan Luís Cabanillas-Moruno
Council of Health. Autonomous Government of Andalusia. Spain
* Corresponding author.

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_10

Spanish workers. If the goal is to compare outcomes across centers and regions for specific diseases and causes of sickness absence, CCIPC is a promising option worthy of prospective testing. The future availability of information through Big Data can increase the external validity of these results if at the same time biases that threaten the internal validity of the results are avoided.

1 Introduction

1.1 Electronic Health Databases

Health services and epidemiologic research are best conducted with population-level data. This helps ensure the appropriate estimation of incidence and prevalence rates, the minimization of referral bias, and the overall generalizability of the study conclusions to the population of interest.

Because prospective clinical registries and retrospective chart review comprising a representative sample or all residents of a jurisdiction are impractical, electronic health administrative data are an alternative for population-based chronic disease surveillance, outcomes research, and health services research. Health administrative data is information passively collected, often by government and health care providers, for the purpose of managing the health care of patients (Spasoff 1999).

Electronic health databases can report timely data that could facilitate surveillance of infectious diseases, disease outbreaks, and chronic illnesses. Software can extract data from records, analyze them, and electronically submit them to public health authorities, which will likely soon receive unprecedented amounts of information (Chretien et al. 2009)

These databases will also greatly facilitate public health research. Large databases can enable researchers to conduct comprehensive observational studies that include millions of records from patients with diverse demographics who are treated in real clinical settings over many years. Researchers can use these diverse collections of data to study disease progress, comorbidities, health disparities, clinical outcomes, treatment effectiveness, and the efficacy of public health interventions, and their findings may influence many public health decisions. To this end, the Patient Protection and Affordable Care Act of 2010 embraces the concept of comparative effectiveness research and supports the use of observational studies to evaluate and compare health outcomes (Cousens et al. 2011)

Health databases may be particularly valuable during public health emergencies. They may enable to obtain critical medical information about disaster victims in the absence of access to their physician's offices or to local computers (Brown et al. 2007). Electronic heath databases can also be situated at disaster scenes or in field hospitals to facilitate data sharing, decision-making, and efficient administrative operations (Levy et al. 2010).

Electronic databases can serve as a continuous communication channel between healthcare professionals and public health authorities. Public health services could

provide electronic updates and recommendations to clinicians both during emergencies and in ordinary times (Garrett et al. 2011).

In some instances, electronic health databases are incomplete, lacking essential information such as treatment outcomes. Patients who receive medication from their doctors often do not report whether the therapy was effective. The absence of return visits may mean that the patients were cured, but it could also indicate that they failed to improve or deteriorated and decided to visit different doctors or specialists (Newgard et al. 2012; Hoffman and Podgurski 2013).

Even if the electronic health data themselves are flawless, analysts must overcome a variety of analytical challenges. These may be particularly pronounced in the case of studies seeking to answer etiologic questions, such as whether certain public health interventions have had a positive impact. Electronic health data is generally observational, not experimental, and hence treatments and exposures are not assigned randomly. This makes it much more difficult to ensure that causal inferences are not distorted by systematic biases. Analysts and users of research data must be familiar with the risks of selection bias, confounding bias, and measurement bias.

Selection bias can occur when analysts unknowingly employ a study group that is not representative of the population of interest. The group studied might have atypical clinical, demographic, or genetic attributes, and therefore, it would be inappropriate to generalize study conclusions to the population at large (Delgado and Llorca 2004).

Confounding bias is a systematic error that occurs because there is a common cause of the treatment/ exposure variable and the outcome variable. For example, socioeconomic factors may be confounders because low income may cause individuals to choose sub-optimal, inexpensive treatments and may also separately lead to deteriorated health because of stress or poor nutrition. A failure to account for socioeconomic status may thus skew study results (Greenland 2003).

Information (misclassification) biases are generated by errors in measurement and data collection resulting from faulty equipment or software or from human error (Copeland et al. 1997). In addition, patients may provide clinicians with incorrect information regarding their medical histories, symptoms, or treatment compliance because they are confused, have impaired memories, or are embarrassed to tell the truth. Systematic errors, which can arise because of problems such as upcoding, are especially challenging.

Accuracy of the diagnostic codes used to identify patients within these data depends on multiple factors including database quality, the specific condition being identified, and the validity of the codes in the patient group. A large gradient in data quality exists, with some databases being of higher quality than others (De Coster et al. 2006).

Isolated diagnostic codes associated with physician billing records have been shown to be accurate to identify patients with some chronic diseases (Chen et al. 2009) but not others (Hux et al. 2002; To et al. 2006; Benchimol et al. 2009; Guttmann et al. 2010).

Since chronic diseases usually require multiple contacts with the health system to diagnose, a single-visit diagnostic code is often insufficient to accurately identify patients with the disease. The validity of codes is also dependent on the patient group being studied. For instance, the accuracy of diagnostic codes or combinations of codes (algorithms) varies across age groups because of variable use of the health system (Ahmed et al.2005; Benchimol et al. 2009; Guttmann et al. 2010).

Validation of algorithms used to identify patients with different health states (including acute conditions, chronic diseases, and other health outcomes) is essential to avoid misclassification (information) bias (Manuel et al. 2007), which may threaten the internal validity and interpretation of study conclusions. For example, assessment of health services utilization in a cohort of patients with a chronic disease contaminated by large number of healthy residents falsely labeled as having a chronic disease would underestimate the burden of the disease on the health system or the quality and performance of the health system. Similarly, assessment of incidence of the disease in the cohort would overestimate risk to the population. Although the validation of administrative data coding has been identified as a priority in the health services research by an international consortium (De Coster et al. 2006), the complete and accurate reporting of algorithm validation research is equally important to appropriate application.

The translation of research from the literature to medical practice or health policy requires the research to be appropriately designed, reported, and interpreted. As such, consortia have created criteria for the reporting of clinical trials (Begg et al. 1996), observational studies (von Elm et al. 2007), and studies of diagnostic accuracy (Bossuyt et al. 2003).

These criteria are guidelines for researchers involved in study design and for consumers of the literature to assess the quality of the research. Unfortunately, no such criteria exist for the creation or reporting of studies using health administrative data. An international symposium assessed priorities of methodological research using administrative data associated with ICD-9 and ICD- 10 (De Coster et al. 2006). Five of 13 potential areas of research identified were related to reliability and validity of these data. These included assessment of internal consistency of identification algorithms, identification of reliable reference standards against which to validate data, the creation of training standards for coders, development of chart-database comparison studies, and international cross-validation of ICD-10.

Validation of health administrative data for identifying patients with different health states (diseases and conditions) is a research priority, but no guidelines exist for ensuring quality. Benchimol et al. (Benchimol et al. 2011) created reporting guidelines for studies validating administrative data identification algorithms and used them to assess the quality of reporting of validation studies in the literature. Using Standards for Reporting of Diagnostic accuracy (STARD) criteria as a guide, they created a 40-item checklist of items with which identification accuracy studies should be reported. A systematic review identified studies that validated identification algorithms using administrative data. They used the checklist to assess the quality of reporting. In 271 included articles, goals and data sources were

well reported but few reported four or more statistical estimates of accuracy (36.9%). In 65.9% of studies reporting positive predictive value (PPV)/negative predictive value (NPV), the prevalence of disease in the validation cohort was higher than in the administrative data, potentially falsely elevating predictive values. Subgroup accuracy (53.1%) and 95% confidence intervals for accuracy measures (35.8%) were also underreported. The quality of studies validating health states in the administrative data varies, with significant deficits in reporting of markers of diagnostic accuracy, including the appropriate estimation of PPV and NPV. They concluded that these omissions could lead to misclassification bias and incorrect estimation of incidence and health services utilization rates. Use of a reporting checklist, such as the one created for this study by modifying the STARD criteria, could improve the quality of reporting of validation studies, allowing for accurate application of algorithms, and interpretation of research using health administrative data.

These cases show the growing importance of electronic health records in different areas of medical care, control of quality of health care and patient safety, health surveillance and evaluation of new procedures. But the next point is the linkage of these electronic health records and its trends.

1.2 Record Linkage

Record linkage involves bringing together records derived from different sources, but relating to the same individual. Data linkage is a technique for connecting pieces of information that are thought to relate to the same person, family, place or event. Administrative information is created each time a person comes into contact with a particular service, such as the registration of a birth or death, a hospital stay or an emergency department visit. If this information can be connected for the whole population in a way that does not breach an individual's privacy, it can all be analyzed.

The three basic steps are: blocking of records that have a potential relationship, matching to determine if records within a block are likely to be related and linking matched records so they can be analyzed as information for the one individual (Gill et al. 1993).

A major challenge in data matching is the lack of common entity identifiers in the databases to be matched. As a result of this, the matching needs to be conducted using attributes that contain partially identifying information, such as names, addresses, or dates of birth. However, such identifying information is often of low quality. Personal details especially suffer from frequently occurring typographical variations and errors, such information can change over time, or it is only partially available in the databases to be matched.

In record linkage, with data matching commonly relying on personal information such as names, addresses and dates of birth of individuals, privacy and confidentiality need to be carefully considered. This is especially the case when databases are matched between organizations, or when the outcomes (the matched

data set) are to be used by an external organization or by individuals such as academic researchers. The analysis of matched data has the potential to uncover aspects of individuals or groups of entities that are not obvious when a single database is analyzed separately (Christen 2012).

For example, the outcomes of analyzing matched health and population databases can potentially lead to discrimination against certain groups of individuals, if it is discovered that these people have a higher risk of getting a certain serious disease. The discrimination could be in the form of higher life insurance premiums, or even that these individuals would find it much harder to gain employment due to their potentially increased risk of long-term illness. In recent years, research into the development of techniques that facilitate privacy preserving record linkage has received attention from areas such as health informatics. The aim is to facilitate the matching of data across organizations without compromising the privacy and confidentiality of the data to be matched.

Linkage of health records has been used to support public health surveillance (Gill et al. 1993; Lynge and Thygesen 1988), etiologic and primary prevention research (Guend, Engholm and Lynge 1990; Van der Brandt et al. 1990), natural history and prognostic research (Overpeck, Hoffman and Prager 1992), and studies of the utilization, adverse effects and outcomes of health care services (Tyndall, Clarke and Shimmins 1987; Thomas and Holloway 1991).

The large-scale systematic application of record linkage for research purposes in the health field is uncommon due to significant requirements for long-term planning and inter-agency cooperation. The few examples of comprehensive record linkage systems include the Oxford Record Linkage Study (Acheson 1967; Goldacre, Shiwach and Yeates 1994), that applying these techniques on birth, death and hospital data of around 350,000 individuals allowed the study of associations between certain diseases, and using longitudinal matched data enabled the analysis of occupational mortality, migration and related socio-economic factors, the Scottish Record Linkage System (Kendrick and Clarke 1993; Ryan 1994), the Rochester Epidemiology Project (Melton 1996) and the Manitoba Population Health Information System (Roos et al. 1995; Black, Burchill and Roos 1995).

The Canadian Mortality Data Base and National Cancer Incidence Reporting System, while they do not include comprehensive data on health care, have shown the value of record linkage in epidemiological follow-up studies of possible carcinogens and other harmful agents (Smith and Newcombe 1982).

Linkage of census and death records in the Office of Population Census and Surveys Longitudinal Study in the United Kingdom has also provided an example of large-scale record linkage, used to study social class mortality differentials (Fox, Goldblatt and Jones 1985).

Some Scandinavian countries have linked mortality and cancer registry data to their central population registers, with the intention of improving their occupational health surveillance (Lynge and Thygesen 1988). In these countries record linkage has been used extensively in health outcomes research.

Recently, Fazel et al. (Fazel et al. 2013) assessed the prevalence and risks of premature mortality from external causes such as suicide, accidents and assaults in

people with epilepsy with and without psychiatric comorbidity. They linked several longitudinal, nationwide population registers in Sweden: the Patient Register (held at the National Board of Health and Welfare), the Censuses from 1970 and 1990 (Statistics Sweden), the Multi- Generation Register (Statistics Sweden), and the Cause-of-Death Register (National Board of Health and Welfare). The Multi-Generation Register connects every person born in Sweden from 1933 onwards and ever registered as living in Sweden after 1960 to their parents. For immigrants, similar information exists for those who became citizens of Sweden before aged 18 years, together with one or both parents. In Sweden, all residents including immigrants have a unique ten-digit personal identification number that is used in all national registers, thus making the linking of data possible. They selected a cohort of people born between 1954 and 2009, and followed them up for up to 41 years, from 1969 to the end of follow-up in 2009 (n=7,238,800). The Patient Registers started in 1969; hence they began their follow-up at that point, which meant that children diagnosed with epilepsy who died between 1954 and 1968 were not included in the cohort. A sensitivity analysis addressed whether this affected the main findings. Using the Multi-Generation Register, they also identified patients with epilepsy who had full siblings without epilepsy.

They studied all individuals born in Sweden between 1954 and 2009 with inpatient and outpatient diagnoses of epilepsy (n=69,995) for risks and causes of premature mortality. Patients were compared with age-matched and sex-matched general population controls (n=660,869) and unaffected siblings (n=81,396). Sensitivity analyses were done to investigate whether these odds differed by sex, age, seizure types, comorbid psychiatric diagnosis, and different time periods after epilepsy diagnosis.6,155 (8.8%) people with epilepsy died during follow-up, at a median age of 34.5 years with substantially elevated odds of premature mortality (adjusted odds ratio of 11.1 (95% CI 10.6–11.6) compared with general population controls, and 11.4 (95% CI 10.4-12.5) compared with unaffected siblings). Of those deaths, 15.8% (n=972) were from external causes, with high odds for non-vehicle accidents (adjusted OR 5.5, 95 % CI 4.7–6.5) and suicide (adjusted OR 3.7, 95% CI 3.3-4.2). Of those who died from external causes, 75.2% had comorbid psychiatric disorders, with strong associations in individuals with co-occurring depression (13.0, 10.3-16.6) and substance misuse (22.4, 18.3-27.3), compared with patients with no epilepsy and no psychiatric comorbidity. They conclude that reducing premature mortality from external causes of death should be a priority in epilepsy management and that psychiatric comorbidity plays an important part in the premature mortality seen in epilepsy.

Pasternak et al. (Pasternak et al. 2013) investigated if oral fluoroquinolone use is associated with an increased risk of retinal detachment. They used a nationwide, register-based cohort study in Denmark from 1997 through 2011, using linked data on participant characteristics, filled prescriptions, and cases of retinal detachment with surgical treatment (scleral buckling, vitrectomy, or pneumatic retinopexy). The Central Person Register, which is Denmark's main administrative register and includes information on date and place of birth,migration, and vital status (updated daily), was used to identify the source population for the cohort.

The National PrescriptionRegistry holds information on all prescriptions filled at all Danish pharmacies since 1995 and is considered near to complete; each new prescription generates an electronic file that is automatically transferred to this registry within minutes. Data include the Anatomic Therapeutic Chemical code of the drug and the date the prescription was dispensed. They used this register to identify prescriptions for any oral fluoroquinolone and for concomitantly used drugs. The Danish National Patient Register holds records of individual-level information from all hospitals in Denmark (inpatient admissions, emergency department visits, and outpatient visits), including physician assigned diagnoses classified according to the International Classification of Diseases, Eighth Revision (ICD-8; between 1977 and 1993) and International Classification of Diseases, 10th Revision (ICD-10; since 1994), as well as data on surgical procedures, classified according to the Nordic Medico-Statistical Committee Classification of Surgical Procedures. This register was used to identify concurrent medical conditions and cases of retinal detachment. The cohort included 748,792 episodes of fluoroquinolone use (660,572 ciprofloxacin, [88%]) and 5 520,446 control episodes of nonuse. Poisson regression was used to estimate rate ratios (RRs) for incident retinal detachment, adjusting for a propensity score that included a total of 21 variables. The risk windows were classified as current use (days 1-10 from start of treatment), recent use (days 11-30), past use (days 31-60), and distant use (days 61-180). A total of 566 cases of retinal detachment occurred, of which 465 (82%) were rhegmatogenous detachments; 72 in fluoroquinolone users and 494 in control nonusers. The crude incidence rate was 25.3 cases per 100 000 person-years in current users, 18.9 in recent users, 26.8 in past users, and 24.8 in distant users compared with 19.0 in nonusers. Compared with nonuse, fluoroquinolone use was not associated with a significantly increased risk of retinal detachment: the adjusted RRs were 1.29 (95%CI, 0.53 to 3.13) for current use; 0.97 (95%CI, 0.46 to 2.05) for recent use; 1.37 (95%CI, 0.80 to 2.35) for past use; and 1.27 (95%CI, 0.93 to 1.75) for distant use. The absolute risk difference, estimated as the adjusted number of retinal detachment cases per 1,000,000 treatment episodes, was 1.5 (95%CI, –2.4 to 11.1) for current use. So, in this cohort study based on the general Danish population, oral fluoroquinolone use was not associated with increased risk of retinal detachment.

There have been examples of linkage systems used in Australia (Sibthorpe, Kliewer and Smith 1995), some of them for special research purposes like the Maternal and Child Health Linked Database used in studies of perinatal and paediatric outcomes(Stanley et al. 1994), studies of the incidence of myocardial infarction (Martin et al. 1989), the Roadwatch Road Injury Research Database (Ferrante, Rosman and Knuiman 1993), a pilot study of linkage of Health Insurance Commission Medicare records (McCallum, Lonergan and Raymond 1993) and evaluation of the patterns of surgical treatment of breast cancer based on linked cancer registry and hospital morbidity records in New South Wales (Adelson et al. 1997; McGeechan et al. 1998)

In 1995 development of a population-based linkage of health records commenced on the Western Australia Health Services Research Linked Database

(Holman et al. 1999; Brook, Rosman and Holman 2008). Based on a best-practice protocol of separating the personal identifiers required for matching from the medical information needed for research studies, data from various health (as well as non-health) sources have been matched, and a chain of records has been generated for each individual person identified. It has summarized over 700 outputs produced by this program from 1995 to 2003. Some of the significant outcomes include improvements in health policies (like regular physical examination for mental health patients) and changes to clinical practice (like installation of shock advisory defibrillators in all ambulances and hospital wards, or community-based services for psychiatric patients at risk of suicide).

Zhang et al (Zhang et al. 2009) conducted a retrospective cohort study based on a state-wide population of patients to investigate whether or not comorbid conditions, age and other demographic factors, and drug category are associated with a repeat admission for adverse drug reactions in people aged ≥60. They used administrative data from all public and private hospitals in Western Australia, a state with a population of 2.09 million in 2007. The study population consisted of all residents aged ≥60 with a hospital admission related to an adverse drug reaction identified through the data linkage system. This system links state-wide administrative health data at the individual level using probabilistic matching of patients' names and other identifiers with clerical review of doubtful matches. It includes links between seven core datasets, of which the statutory death registry from 1969 and the hospital morbidity data system from 1970 form the main parts. They used extracts of linked hospital morbidity records and death records, with encryption to protect the identity of individual patients. The data were extracted in February 2005. The hospital morbidity data system contained information on encrypted patient identification and episode number; age, sex, indigenous status, and postcode; date of admission and date of separation (that is, transfer, discharge, or inpatient death); international classification of diseases codes for the main diagnosis and up to 19 additional diagnoses for up to four external causes (E codes), and for the main procedure and up to 10 additional procedures; type of hospital attended (public, private, other), admission type (emergency or elective), and payment classification. They used ICD-9 for 1980-7, ICD-9-CM for 1988-June 1999, and ICD-10-AM from July 1999 onwards. Data from the death records included encrypted patient identification, age, sex, indigenous status, primary cause of death, date of death, and postcode. They extracted linked hospital and death records for all patients aged ≥60 with an admission for adverse drug reaction in 1980-2003 in Western Australia. In an assessment of the technical performance of the linkage system in finding true matches between records, both the proportion of invalid links (false positives) and of missed links (false negatives) was estimated to be 0.11%. 28,548 patients aged ≥60 years with an admission for an adverse drug reaction during 1980-2000 were followed for three years using the Western Australian data linkage system. 5,056 (17.7%) patients had a repeat admission for an adverse drug reaction. Repeat adverse drug reactions were associated with sex (hazard ratio 1.08, 95% confidence interval 1.02 to 1.15, for men), first admission in 1995-9 (2.34, 2.00 to 2.73), length of hospital stay (1.11, 1.05 to 1.18, for stays

≥14 days), and Charlson comorbidity index (1.71, 1.46 to 1.99, for score ≥7); 60% of comorbidities were recorded and taken into account in analysis. In contrast, advancing age had no effect on repeat adverse drug reactions. Comorbid congestive cardiac failure (1.56, 1.43 to 1.71), peripheral vascular disease (1.27, 1.09 to 1.48), chronic pulmonary disease (1.61, 1.45 to 1.79), rheumatological disease (1.65, 1.41 to 1.92), mild liver disease (1.48, 1.05 to 2.07), moderate to severe liver disease (1.85, 1.18 to 2.92),moderate diabetes (1.18, 1.07 to 1.30), diabetes with chronic complications (1.91, 1.65 to 2.22), renal disease (1.93, 1.71 to 2.17), any malignancy including lymphoma and leukaemia (1.87, 1.68 to 2.09), and metastatic solid tumours (2.25, 1.92 to 2.64) were strong predictive factors. Comorbidities requiring continuing care predicted a reduced likelihood of repeat hospital admissions for adverse drug reactions (cerebrovascular disease 0.85, 0.73 to 0.98; dementia 0.62, 0.49 to 0.78; paraplegia 0.73, 0.59 to 0.89). They concluded that comorbidity, but not advancing age, predicted repeat admission for adverse drug reactions in older adults, especially those with comorbidities often managed in the community and that awareness of these predictors can help clinicians to identify which older adults are at greater risk of admission for adverse drug reactions and, therefore, who might benefit from closer monitoring.

Cabanillas et al (Cabanillas et al. 2012) linked retrospectively four databases of the Spanish National Health System in the region of Andalusia: a) the Users Database of the National Health System with identification and demographic data (birth date, gender, place of residence, identification number, and others); b) the Sigilum XXI database with information about sickness absences registered by physicians of the National Health System in the region of Andalusia. In this database, besides identification data there are codes for the cause and duration of sickness absence; c) DIRAYA database, with electronic health record of every user of the National Health System in the region of Andalusia in primary health care. Includes diagnoses and procedures codified using ICD-9-CM, allowing the calculation of Charlson Comorbidity Index for primary health care patients (CCIPC) (Charlson et al. 2008); d) Minimal Basic Data Set a database with information of all patients admitted to private and public hospitals of the region. Each record contains up to 83 attributes comprising the personal characteristics of the patient including, age, gender, municipality of residence; administrative information including date of admission, admission type, departure date, departure type (including in-hospital death) and medical information including up to 15 diagnoses and up to 20 procedures. Diagnoses and procedures were coded according to the ICD-9-CM codes. They analyzed the period 2007-2009 among 2,903,401 workers with 3,039,337 sickness absence episodes during that period. Using a multivariate model they predicted the length of sickness absence specific for every disease using age, gender, province of residence, hospital size and Charlson Comorbidity Index.

Relationships between CCIPC and sickness absence seems a promising issue and deserves more research, as analyzed in the next section.

1.3 Charlson Comorbidity Index for Primary Health Care Patients and Sickness Absence

Sickness absence is an important cause of lost productivity, expenses and disability (Rice, Hodgson and Kopstein 1985). Some researchers have used it as a social measure of health status and functioning (Marmot el al. 1995) and because of its relationship with all-cause mortality, it has also been suggested as a global measure of health (Kivimaki et al. 2003).

Sickness absence varies among different groups in society and over time. The rate of sickness absence shows notable differences among the countries of the European Union (Gimeno et al. 2004), but in all of them the burden of sick leave in terms of lost productivity, expenses and long term disability is huge.

One of the aims of epidemiologic research is the study of risk factors for sickness absence using information registered in administrative databases. But in order to draw valid inferences from studies based on administrative databases, it is necessary to adjust for patient risk, recognizing that the underlying nature of some patients' diseases makes them more likely than others to have poor outcomes (Iezzoni 2003).

Several comorbidity indexes have been developed and validated using administrative databases (Southern, Quan and Ghali 2004). Charlson comorbidity index (CCI) (Charlson et al. 1987) has been widely used since 1987 to measure the burden of a disease or case-mix with hospital administrative data. In its original paper Charlson et al. defined 17 comorbidities using clinical conditions recorded in charts. Deyo et al. (Deyo, Cherkin and Ciol 1992) and Romano et al. (Romano, Roos and Jollis 1993) independently developed the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) coding algorithms for the Charlson comorbidities. Deyo's and Romano's coding algorithms are similar in generating Charlson index scores and in their ability to predict outcomes (Romano, Roos and Jollis 1993; Ghali et al. 1996; Cleves, Sanchez and Draheim 1997).

Since the publication of Charlson et al. original article the CCI has been validated for its ability to predict mortality in several disease subgroups including critically ill (Poses et al. 1996; Quach et al. 2009), renal disease (Hemmelgarn et al. 2003), stroke (Goldstein 2004), heart failure (Lee et al. 2005), liver cirrhosis (Myers et al 2009) and acute myocardial infarction (Gili et al. 2011).

More recently, Charlson developed and validated an adapted comorbidity index to predict costs in primary care patients (Charlson et al. 2008). This adapted Charlson comorbidity index for primary care setting (CCIPC) added four additional comorbidities to the original CCI, and predicted costs and mortality.

In Spain, population has free access to primary care (PC) and hospital care in the National Health System (Sistema Nacional de Salud). In Andalusia, PC physicians codify the causes and conditions of everyday care of patients in their electronic medical records using ICD-9-CM codes, allowing the calculation of CCIPC. In the case of a disease causing sickness absence, every worker needs a sickness

certification completed by their PC physician, who codifies the cause of sick leave in patient electronic medical records using ICD-9-M. PC physicians codify the cause of sickness disability and other diseases selecting the ICD-9-M code in a pull-down menu. If a worker is admitted to a hospital of the National Health System, diagnostic codes, external causes and procedures are also coded, allowing a follow-up of outcomes for every sickness absence.

As far as we know, nobody has evaluated the ability of CCIPC among workers to predict sickness absence, including its number and duration, hospital admissions, its number and length of hospital stay, and in-hospital mortality.

The goal of this study was to use prospectively acquired data among a large sample of Spanish workers to determine the ability of CCIPC to predict sickness absence, hospital admissions and in-hospital mortality. Additionally, we want to analyze the promising impact of the availability of information through Big Data on this process.

2 Aims and Structure of the Chapter

The main objective of this Chapter is to use prospectively acquired data among a large sample of Spanish workers to determine the ability of Charlson Comorbidity Index for primary health care (CCIPC) to predict absenteeism, hospital admissions and in-hospital mortality, and to analyze the promising impact of the availability of information through Big Data in this process.

For achieving these objectives we have structured the Chapter in the next sections and subsections:

3. METHODS.

3.1. Population.

3.2. Data Collection.

3.3. Statistical Analysis.

4. RESULTS.

5. DISCUSSION.

5.1. Sickness absence and CCIPC.

5.2. The next step: availability of information through Big Data.

6. CONCLUSIONS.

REFERENCES.

ACKNOWLEDGMENTS.

3 Methods

3.1 Population

The study was done in Andalusia, a region situated in Southern Spain that is the most vast and populated region of Spain. To qualify for the study, workers had to have been registered as workers in databases of the National Health System in Andalusia during the period between January 1, 2007 and December 31, 2009. Workers continuously unemployed during the 3 years of the study period are not labeled as active workers in PC databases thus they were not included in the study.

3.2 Data Collection

Data from PC were captured by DIRAYA©, a practice management system developed by the Andalusian public health care system, which is used as an information and care management support, with a coverage of 94% of the Andalusian population.

Physicians used the electronic medical record in the everyday care of patients to code outpatient diagnoses, laboratory tests, radiology tests, procedures, consultations and write prescriptions. Physicians assigned diagnostic codes using ICD-9-CM at the conclusion of each visit, choosing the codes in a pull-down menu. The diagnostic list was cumulative over time with new diagnoses added as they occurred. Therefore, DIRAYA© provided a prospectively collected database of demographics, appointments and ICD-9-CM diagnoses.

DIRAYA© is also used by PC physicians to code the cause of sickness absence longer than 3 days of workers using ICD-9-CM, the dates of sick leave and discharge, and the types of sick absence: common medical disease, occupational disease or work accident. Thus, DIRAYA© also provided a prospectively collected database of sick leaves and their length.

Data for hospital admissions were captured by the administrative minimal basic data set (MBDS) of the hospitals of Andalusia during the period 2007-2009. From written information that is provided by the hospital physician who signs the discharge report, each patient's diagnosis, external causes and procedures are codified according to ICD-9-CM codes. Codification and data entry in the electronic database are performed by dedicated administrative personnel who have completed in-depth training on medical data registration. This administrative database has demographic data, admission and discharge dates, type of admission and type of discharge, diagnostic codes for principal cause and secondary diagnoses, external causes and procedures using ICD-9-CM codes (AHS 2009).

This study was approved by the Medical Ethics Committee of our hospital. Given the retrospective nature of our study, informed consent could not be obtained from each participant.

The CCIPC was used to assess the prognostic burden of comorbid diseases. This index assigns weights for specific diseases. Comorbid conditions with a

weight of one include myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, ulcer disease, mild liver disease, diabetes without end organ damage, depression, use of warfarin and arterial hypertension. Hemiplegia, moderate or severe renal disease, diabetes with end organ damage, any tumor, leukemia, or lymphoma, and skin ulcer or cellulitis have a weight of two. Moderate or severe liver disease has a weight of three. Metastatic solid tumor and AIDS have a weight of six. The total score is calculated by adding the weights.

3.3 Statistical Analysis

We estimated CCIPC for each worker in our dataset. Relationship among CCIPC, as calculated with information in DIRAYA© administrative dataset, and sickness absence, hospital admission and in-hospital mortality were evaluated with univariate logistic regression analysis. We also categorized CCIPC (0, 1-2, 3-4, 5 or more) to determine its impact on sickness absence, on the number of episodes of sickness absence, on the cumulative duration of sickness absence, on hospital admissions, on the number of hospital admissions, on the cumulative length of hospital stays, and on in-hospital mortality. Stratified analysis was done using the Mantel and Haenszel method. Analysis of covariance were used to evaluate the relation among CCIPC, age, gender and the number and duration of sickness absence, and among CCIPC, age, gender and the number and length of stay of hospital admissions. Sick leaves caused by pregnancy, child birth and puerperium were excluded from these analysis.

Subsequently, unconditional logistic regression analyses and multinomial logistic regression analyses were performed to evaluate if the predictive power of CCIPC index with outcomes could be improved by adding age, gender, province of residence and size of hospital to the predictive model, excluding pregnancy, childbirth and puerperium causes of sickness absence. Age, gender, province of residence and hospital size were added to the models predicting sickness absence, number of sick leaves, duration of sick leaves and hospital admission, number of hospital admissions, length of hospital stay and in-hospital mortality. Adjusted odds ratios and corresponding 95% confidence intervals were calculated. There are 8 provinces in Andalusia. The size of the hospital was classified in four categories: less than 200 beds, 200-499 beds, 500-999 beds and 1,000 or more beds. The performance of the risk adjustment models for predicting sickness absence, the number of episodes of sickness absence, the cumulative duration of sickness absence, hospital admission, the number of hospital admissions, the cumulative length of hospital stays and in-hospital mortality were determined measuring areas under the receiver operating characteristic curve (AUC) as a measure of model discrimination, following the methods recommended by Hanley and McNeil (Hanley and McNeil 1982).

Statistical analysis was done using STATA version 12 MP (StataCorp LP, College Station, TX).

4 Results

During the period of study, 1,884,033 workers were identified but 57,682 were excluded because of unavailability or incompleteness of data. Among the remaining 1,826,190 workers 1,103,484 had one or more sick leaves. Most (82.5%) of the workers were never hospitalized. Of the 319,194 of workers who were hospitalized, 231,453 (72.5%) were hospitalized once. Among hospitalized patients 6,154 died.

A total of 2,447,436 episodes of sickness absence were registered during the study period. A majority of them, 2,239,010 episodes, were caused by common diseases (91.5%), followed by 179,345 non-occupational accidents (7.3%), 24,463 occupational accidents (1.0%) and 2,618 occupational diseases (0.1%).

Relationships among demographic characteristics of workers, comorbidities and outcomes are exposed in Table 1. Among all workers, 1,354,881 were aged 16-49 years (75.2%). Men accounted for 55.8% of the sample. Among workers with CCIPC>0 mean CCIPC was 1.7 but with remarkable differences among groups of age. In this group, mean CCIPC was 1.20 for workers aged 16-19 years old, 1.25 for those aged 20-29 years, 1.38 for those aged 30-39 years, 1.66 for those aged 40-49 years, 1.93 for those aged 50-59 years, 2.31 for those aged 60-69 years, and 2.75 for those aged 70 or more.

Among the group of workers with CCIPC>0 the most frequent comorbidities were arterial hypertension (25.8%), depression (21.6%) and chronic pulmonary diseases (20.4%) and the least frequent comorbidities were dementia (0.1%), AIDS (0.2%) and diabetes with end organ damage (0.3%), but with remarkable differences among both genders.

During the 3-year period, there were 2,293,570 episodes of sickness absence among 1,103,484 workers. The median age of workers with these episodes was 38 (interquartile range 29-49) and 52.7% of them were men.

During the same period there were 411,048 hospital admissions among 319,194 workers with sickness absence. The median age of the hospitalized workers was 45 years (interquartile range 35-56), 59.8% of hospital admissions were men and in-hospital mortality was 1.9% among hospitalized (0.3% of all workers).

Sickness absences and hospital admissions were more frequent among female workers, but CCIPC with a score of 2 or more were more prevalent in male workers (6.3% vs. 4.7% among females) as well as in-hospital mortality (0.5% vs. 0.2% among females), emphasizing the higher severity of diseases and comorbidities among male workers.

CCIPC was stratified in four levels (0, 1-2, 3-4 and 5 or more) in order to analyze dose-effect relationships between CCIPC score and outcomes. Using stratified analysis we found a direct causal association with sickness absence, number of episodes of sickness absence, cumulative duration of sickness absences, hospital admission, number of hospital admissions, cumulative length of hospital stays and in-hospital death in terms of Odds Ratios, 95% confidence intervals, and statistical signification for trend, after adjusting for age and gender (Table 2) and

excluding cases of sickness absence caused by pregnancy, child birth and puerperium. The magnitude of the dose-effect relationship between CCIPC and sickness absence, in-hospital death and hospital admission was the highest.

Multivariate analysis was done using unconditional regression analysis and a direct causal association was found between the model including CCIPC, age, gender, province of residence and hospital size with sickness absence, number of episodes of sickness absence, cumulative duration of sickness absence, hospital admission, the number of hospital admissions and in-hospital death, as depicted in Table 3. Cases of sickness absence caused by pregnancy, childbirth and puerperium were excluded from the analysis. Areas under ROC curves were calculated for every predictive model and the biggest AUC were for in-hospital death (0.9254), number of hospital admissions (0.7695) and cumulative length of hospital stays (0.7650).

Finally, non-dichotomic outcomes were stratified in order to analyze dose-effect relationships between models including CCIPC, age, gender, province of residence, hospital size and these outcomes by means of multinomial logistic regression analysis. Outcomes also excluded cases of sick leave caused by pregnancy, childbirth and puerperium. Results are exposed in Table 4, finding a direct causal association and positive dose-effect between the model and the outcomes. The magnitude of the dose-effect relationship between CCIPC and the cumulative number of hospital admissions and cumulative length of hospital stays was the highest.

Table 1 Demographic features, Charlson comorbidity index, sickness absence, hospital admission and in-hospital deaths among the sample of workers during the period 2007-2009.

Variables	Gender		
	Male (n= 1,019,328) Number (%)	Female (n=806,862) Number (%)	All (n=1,826,190) Number (%)
Age in years			
16-20	22,443 (2.2)	11,956 (1.5)	34,399 (1.9)
20-29	182,936 (17.9)	170,992 (21.2)	353,928 (19.4)
30-39	266,215 (26.1)	258,104 (32.0)	524,319 (28.7)
40-49	250,222 (24.5)	192,013 (23.8)	442,235 (24.2)
50-59	187,423 (18.4)	118,896 (14.7)	306,319 (16.8)
60-69	105,870 (10.4)	51,509 (6.4)	157,379 (8.6)
≥70	4,219 (0.4)	3,392 (0.4)	7,611 (0.4)
Charlson comorbidity index			
0	862,286 (84.6)	677,311 (83.9)	1,539,597 (84.3)
1	93,875 (9.2)	91,785 (11.4)	185,660 (10.2)
2	37,367 (3.7)	25,770 (3.2)	63,137 (3.5)

Table 1 (*continued*)

3	11,895 (1.2)	6,376 (0.8)	18,271 (1.0)
4	4,974 (0.5)	1,767 (0.2)	6,741 (0.4)
≥5	8,931 (0.9)	3,853 (0.5)	12,784 (0.7)
Comorbidities[1]			
Myocardial infarction	6,005 (3.8)	737 (0.6)	6,742 (2.4)
Congestive heart failure	3,431 (2.2)	870 (0.7)	4,301 (1.5)
Peripheral vascular disease	3,013 (1.9)	572 (0.4)	3,585 (1.3)
Cerebrovascular disease	5,189 (3.3)	1,848 (1.4)	7,037 (2.5)
Dementia	190 (0.1)	143 (0.1)	333 (0.1)
Chronic pulmonary disease	31,438 (20.0)	27,168 (21.0)	58,606 (20.4)
Connective tissue disease	828 (0.5)	1,812 (1.4)	2,640 (0.9)
Ulcer disease	5,505 (3.5)	2,129 (1.6)	7,634 (2.7)
Mild liver disease	8,302 (5.3)	2,497 (1.9)	10,799 (3.8)
Diabetes without end organ damage	23,597 (15.0)	9,447 (7.3)	33,044 (11.5)
Depression	22,066 (14.1)	39,932 (30.8)	61,998 (21.6)
Use of warfarin	1,800 (1.1)	763 (0.6)	2,563 (0.9)
Hypertension	46,392 (29.5)	27,630 (21.3)	74,022 (25.8)
Hemiplegia	883 (0.6)	361 (0.3)	1,244 (0.4)
Moderate or severe renal disease	2,544 (1.6)	851 (0.7)	3,395 (1.2)
Diabetes with end organ damage	579 (0.4)	170 (0.1)	749 (0.3)
Any tumour, leukemia or lymphoma	12,436 (7.9)	9,526 (7.4)	21962 (7.7)
Skin ulcer / cellulitis	5,355 (3.4)	2,530 (2.0)	7,885 (2.8)
Moderate severe liver disease	1,444 (0.9)	179 (0.1)	1,623 (0.6)
Metastatic cancer	2,464 (1.6)	1,488 (1.1)	3,952 (1.4)
AIDS	522 (0.3)	124 (0.1)	646 (0.2)
At least 1 sickness absence[2]	609,232 (59.8)	494,252 (61.3)	1,103,484 (60.4)
At least 1 hospital admission[2]	166,842 (16.4)	152,352 (18.9)	319,194 (17.5)
In-hospital death	4,647 (0.5)	1,507 (0.2)	6,154 (0.3)

[1] Number and percentage among workers with CCIPHC>0.

[2] Excluding pregnancy, child birth and puerperium.

Table 2 CCIPC and sickness absence, number of episodes of sickness absence, cumulative duration of sickness absences, hospital admission, number of hospital admissions, cumulative length of hospital stays and in-hospital death. Mantel and Haenszel Odds Ratios adjusted for age and sex (excluding pregnancy, child birth and puerperium).

Variable	CCIPC	Odds Ratio	95% Confidence Interval	p	p for trend
Sickness absence (at least one episode of sickness absence vs none)	0	1	-	-	<0.0001
	1 - 2	59.4	57.4 - 61.5	<0.0001	
	3 - 4	73.3	66.8 - 80.3	<0.0001	
	5 or more	153.5	128.3 - 183.6	<0.0001	
Number of episodes of sickness absence (10 or more vs 9 or less)	0	1	-	-	<0.0001
	1 - 2	3.8	3.6 - 4.0	<0.0001	
	3 - 4	8.0	7.1 - 8.9	<0.0001	
	5 or more	11.3	9.9 - 12.9	<0.0001	
Cumulative duration of sickness absence (40 or more days vs 39 or less)	0	1	-	-	<0.0001
	1 - 2	1.8	1.8 - 1.9	<0.0001	
	3 - 4	2.2	2.1 - 2.3	<0.0001	
	5 or more	2.3	2.2 - 2.4	<0.0001	
Hospital admission (at least one hospital admission vs none)	0	1	-	-	<0.0001
	1 - 2	1.5	1.5 - 1.5	<0.0001	
	3 - 4	5.2	5.0 - 5.2	<0.0001	
	5 or more	39.1	36.1 - 42.4	<0.0001	
Number of hospital admissions (4 or more vs 3 or less)	0	1	-	-	<0.0001
	1 - 2	3.6	3.4 - 3.8	<0.0001	
	3 - 4	13.8	12.8 - 15.0	<0.0001	
	5 or more	29.6	27.2 - 32.2	<0.0001	
Cumulative length of hospital stays (14 or more days vs 13 or less)	0	-	-	-	<0.0001
	1 - 2	3.3	3.2 - 3.4	<0.0001	
	3 - 4	10.3	9.9 - 10.8	<0.0001	
	5 or more	28.6	26.9 - 30.3	<0.0001	
In-hospital death (yes vs no)	0	1	-	-	<0.0001
	1 - 2	6.7	5.9 - 7.6	<0.0001	
	3 - 4	21.2	18.4 - 24.5	<0.0001	
	5 or more	133.3	114.4 - 155.3	<0.0001	

Table 3 CCIPC and sickness absence, number of episodes of sickness absence, cumulative duration of sickness absences, hospital admission, number of hospital admissions, cumulative length of hospital stays and in-hospital death. Unconditional logistic regression analysis adjusted for age, sex, province of residence and hospital size (excluding pregnancy, child birth and puerperium).

Variable	Adjusted Odds Ratio	95% Confidence Intervals	p	Area under ROC curve
Sickness absence (at least one episode of sickness absence vs none)	44.00	42.90 – 45.13	<0.0001	0.7616
Number of episodes of sickness absence (10 or more vs 9 or less)	2.55	2.48 – 2.62	<0.0001	0.6903
Cumulative duration of sickness absence (40 or more days vs 39 or less)	1.59	1.58 – 1.60	<0.0001	0.6848
Hospital admission (at least one hospital admission vs none)	3.69	3.66 – 3.72	<0.0001	0.6711
Number of hospital admissions (4 or more vs 3 or less)	3.05	2.99 – 3.10	<0.0001	0.7695
Cumulative length of hospital stays (14 or more days vs 13 or less)	3.07	3.03 - 3.11	<0.0001	0.7650
In-hospital death (yes vs no)	5.32	5.17 – 5.48	<0.0001	0.9254

Table 4 CCIPC and cumulative number of sickness absence, cumulative duration of sickness absence, cumulative number of hospital admissions and cumulative length of hospital stay. Multinomial logistic regression analysis adjusted by age, sex, province of residence and hospital size (excluding pregnancy, birth and puerperium).

CCIPC and Cumulative number of sickness absence	Cumulative number of sickness absence	Adjusted Odds Ratios	95% Confidence Limits	p
	1-3	1	-	-
	4-10	2.17	2.15 - 2.19	<0.0001
	11-19	3.09	2.98 - 3.20	<0.0001
	20+	3.44	3.04 - 3.90	<0.0001
CCIPC and cumulative duration of sickness absence in days	Cumulative duration of sickness absence			
	1-30	1	-	-
	30-119	1.38	1.37 - 1.40	<0.0001
	120-199	1.60	1.58 - 1.62	<0.0001
	200+	2.10	2.08 - 2.12	<0.0001
CCIPC and cumulative number of hospital admissions	Cumulative number of hospital admissions			
	1-2	1	-	-
	3-4	2.31	2.27 - 2.34	<0.0001
	5-6	3.60	3.49 - 3.71	<0.0001
	7+	4.88	4.66 - 5.11	<0.0001
CCIPC and cumulative length of hospital stays in days	Cumulative length of hospital stays in days			
	1-6	1	-	-
	7-12	1.72	1.70 - 1.75	<0.0001
	13-18	2.26	2.21 - 2.31	<0.0001
	19+	3.60	3.54 - 3.67	<0.0001

5 Discussion

5.1 Sickness Absence and CCIPC

In this study using administrative data of 1,826,190 Spanish workers, we describe the ability of CCIPC, a comorbidity index developed at PC settings to predict sickness absence and its associated outcomes.

In our results CCIPC was predictive of sickness absence, the number of episodes of sickness absence, the duration of sickness absence, hospital admission, the number of hospital admissions, the length of hospital stays and in-hospital

mortality. This predictive ability persisted after including in the predictive models other variables such as age, gender, province of residence and hospital size.

A dose-effect relationship was found between CCIPC and these outcomes when this comorbidity index was stratified in four categories and analysis was adjusted by age and gender. Furthermore, the predictive ability of CCIPC was shown by multinomial logistic regression models when non-dichotomic outcomes were stratified (number of sick leaves, duration of sick leaves, number of hospital admissions and length of hospital stays), and the predictive models included other variables such as age, gender, province of residence and hospital size.

The inclusion of hospital size in the model was determined because of the different case-mix and prognosis of workers admitted to hospitals. Tertiary care centers tend to see sicker patients and perform more sophisticated procedures, so the number and duration of sick leaves, the number and length of hospital stays and the risk of in-hospital mortality may be influenced both by workers comorbidities but also by the characteristics and quality of hospital care.

CCIPC was predictive of duration of sickness absence for a pool of different diseases and conditions causing sick leave. Probably CCIPC will be more or less predictive depending of the disease or condition that causes sickness absence, but in this study it was predictive regardless of the specific disease causing the sick leave. This predictive ability could be helpful, at least theoretically, for the development of predictive models for the individual duration of sickness disability for specific diseases, including age, gender, province of residence, hospital size and CCIPC as independent variables in the models.

This study has some limitations. The data used were those contained in the DIRAYA and MBDS databases and were not complemented with additional data from patients. The analysis was limited to mortality during hospital stay because data on medium to long patient outcomes were not available.

Data that we used were collected for administrative purposes and not for the purposes of this study by clinicians, both at PC and hospital settings, using standardized data collection forms. We had to rely on the information that was provided by the clinicians who took care of the patients during everyday clinical practice.

PC physicians codify the cause of sickness disability and other diseases selecting the ICD-9-M code in a pull-down menu. Errors in the selection of disease codes may influence the CCIPC score but we have not analyzed the validity of these diagnoses and if these kind of errors have a repercussion on CCIPC scores. In Spanish hospitals, coding is performed by professional coders working from patient discharge reports, which are in turn completed by the discharging physician. The regulations are fully explained in various publications aimed at coders, and follow-up reduces coder-produced information bias, but cannot completely eliminate it. There were no major changes in the codes during the period 2007-2009 and no significant changes in coder work patterns during that period.

We included the classic demographic variables of age, gender, and place of residence that are mentioned as important determinants of sickness absence,

regardless of diagnosis or underlying disease, and they are often treated as confounders in studies of sickness absence (Allebeck and Mastekaasa 2004).

We did not take into account variables unavailable in our databases such as marital status (Mastekaasa 2000), changes in marital status (Eriksen W, Natvig and Bruusgaard 1999; Bratberg , Dahl and Risa 2002), the number of children living at home (Leigh 1986; Vistnes 1997) and its impact as causes of sick leave. Other hypothetical confounders were not available in data, such as occupation or some social class indicator (economical, occupational or cultural), variables that have been directly related to sickness absence (Chevalier et al. 1987; Eyal, Carel and Goldsmith 1994; Feeney et al. 1998; Moncada et al. 2002; Fuhrer et al. 2002), and avoidable readmissions (Pappas et al. 1997).

At the same time, many of those working illegally ("black economy") are not registered in the social security and their data are not recorded as those of workers in PC databases. Those workers with long term unemployment during the 3 years of the study period were not labeled as registered workers and were not included in the study.

Administrative databases such as DIRAYA and MBDS also have clear advantages. The data collected are usually completed in all PC settings and hospital admissions and, as they include virtually all cases, provide reasonably accurate estimates of the incidence, prevalence, comorbidity, and mortality of diseases treated in the PC and hospital settings. Data can be analyzed retrospectively, unlike other designs that require prospective data collection, and data from long periods or from large numbers of patients can be gathered quickly and easily. Since data are collected systematically, costs are considerably reduced. In studies based on these databases, there may be less selection bias due to the refusal of patients or their legal representatives to sign consent forms which would allow the patient to participate in the study.

The goal of risk adjustment was to ensure that the severity of disease among patients was accounted for in the assessment of sickness absence outcomes. This risk adjustment for sick leave outcomes can be used for quality assessment across PC centers and hospitals in different regions and for comparison of procedure outcomes (Marshall et al. 2000; Romano and Zhou 2004).

In further studies of sickness leave and outcomes associated to sickness disability the use of a comorbidity index calculated at PC setting as CCIPC seems worthy of prospective testing. But more promising is the availability of information through Big Data.

5.2 The Next Step: Availability of Information through Big Data

With increasingly large electronic health databases connected in clouds, generalizability (external validity) of the results of studies of sickness absence and its relationship with demographic variables, social factors and comorbidities will be enhanced.

Epidemiologic data resources are being consolidated into increasingly large clusters. The resulting collections of databases, cohorts, and case populations are often so large that they become unique, never-to-be-replicated resources. Although these colossal epidemiologic projects offer unprecedented research opportunities, they also create new challenges. For example, researchers need to coordinate their work within collaborative teams, standardize data collection and analysis procedures, foster the career development of junior investigators, secure funding for each of the participating teams, and provide epidemiologists outside the consortium with access in order to ensure maximum benefit from the resource (Hernan and Savitz 2013).

A recent assessment of drugs that target the renin-angiotensin-aldosterone system and angioedema risk drew from a source population of more than 100 million people and 350 million person-years of observation time (Toh and Platt 2013). The assessment identified 3.9 million eligible new users of angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), the direct renin inhibitor aliskiren, or the common referent group beta-blockers (a class of drugs not thought to affect the risk of angioedema). More than 4,500 outcome events were observed. The assessment replicated a well-known association between ACEIs and angioedema, but the risk estimates were much more precise than those from prior studies. The assessment also generated new evidence for ARBs and aliskiren. Not so long ago, an assessment of such scale existed only in our imaginations. Secondary uses of routinely collected electronic health information now enable us to conduct research using data from hundreds of thousands or even millions of patients. But certain studies or surveillance activities, especially those with rare exposure or outcome, demand data larger than any single extant source. Combining data from multiple sources would help solve the sample size problem, but sharing data has always been a challenge because of privacy, security, regulatory, legal, and proprietary concerns.

The amalgamation of data from disparate sources as genomics, molecular, clinical, epidemiologic, environmental, and digital information has the potential to alter medical and public health decision making. In 2012, the U.S. government unveiled the "Big Data" Initiative with $200 million committed to research across several agencies (Mervis 2012).

Epidemiologists have traditionally been involved in the collection and analysis of large data sets, and therefore should play a central role in directing the use of financial resources and institutional/organizational investment to build infrastructures for the storage and analysis of massive datasets. Critical to the implementation of Big Data science is the need for high-quality biomedical informatics, bioinformatics, and mathematics and biostatistics expertise (Khoury, Lam and Ioannidis 2013).

The development of systematic approaches to robustly manage, integrate, analyze, and interpret large complex data sets is crucial. Overcoming the challenges of developing the architectural framework for data storage and management may benefit from the lessons learned and the knowledge gained from other disciplines (Birney 2012).

Adaptation of technological advancements like cloud-computing platforms, already in use by private industries (e.g., Amazon Cloud Drive and Apple iCloud), can further facilitate this virtual infrastructure and transform biomedical research and health care (Pechette 2012).

The tasking challenges for integration of multiscale data to promote progress in research lies more in the realm of bioinformatics and in the details related to data sharing and to adopt standards and metrics that can cross studies and disciplines. The National Institute of Standards and Technology (NIST) is sponsoring the "Cloud Computing and Big Data Workshop" precisely to deliberate on some of these pressing challenges (NIST 2013).

For data acquired from disparate sources, harmonization of definitions can be a challenge. The epidemiology community and funding agencies can integrate the insights gained from this NIST workshop toward better integration of big data science in future epidemiologic studies. There is an urgent need for a systematic approach to manage and synthesize large amounts of information (Galea, Riddle and Kaplan 2010).

Generalizability (external validity) of epidemiologic studies of sickness absence will be enhanced with the increased availability of data through Big Data, but it is essential to avoid biases (selection, information and confounding) that may threaten the internal validity and interpretation of study results.

6 Conclusions

With electronic health databases researchers must overcome several analytical challenges. These may be particularly important in the case of studies analyzing cause-effects relationships, such as whether certain public health interventions and procedures have had a positive impact (effective) and at the same time they are safe. Electronic health data is generally observational and hence treatments and exposures are not assigned randomly. This makes it much more difficult to ensure that causal inferences are not distorted by systematic biases such as selection, information or confounding biases. Several consortia have created criteria for the reporting of clinical trials, observational studies, and studies of diagnostic accuracy, but no such criteria or guidelines exist yet for the creation or reporting of studies using health administrative data. An international symposium has assessed priorities of methodological research using administrative data, with special emphasis on reliability and validity of these data. These include assessment of internal consistency of identification algorithms, identification of reliable reference standards against which to validate data, the creation of training standards for coders, development of chart-database comparison studies, and international cross-validation of classification of diseases codes.

Record linkage involves bringing together records derived from different sources, but relating to the same individual. Several applications of record linkage have been reviewed in this article, as public health surveillance, identification of carcinogens, occupational health surveillance, external causes surveillance,

adverse effects of drugs, studies of perinatal and paediatric outcomes, studies of the incidence of myocardial infarction, evaluation of the patterns of surgical treatment, studies on comorbid conditions and hospital readmissions, and prediction of the length of sickness absence specific for every disease, among others.

In our study using record linkage, the adapted Charlson Comorbidity Index for Primay Healthcare was predictive of all outcomes related to sickness absence, hospital admissions, and in-hospital mortality among a large sample of Spanish workers. If the goal is to compare outcomes across centers and regions for specific diseases and causes of sickness absence, this comorbidity indicator is a promising option worthy of prospective testing. The future availability of information through Big Data can increase the external validity of these results if at the same time biases that threaten the internal validity of the results are avoided.

Acknowledgments. This Research was funded by the Spanish Delegación del Gobierno para el Plan Nacional Sobre Drogas (DGPNSD) (grant number 2009I017, project G41825811). DGPNSD had no further role in study design; in the collection; analysis and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

References

1. Acheson, E.D.: Medical Record Linkage. Oxford University Press, London (1967)
2. Adelson, P., Lim, K., Churches, T., Nguyen, R.: Surgical treatment of breast cancer in New South Wales 1991, 1992. Aust. N. Z. J. Surg. 67(1), 9–14 (1997)
3. Ahmed, F., Janes, G.R., Baron, R., Latts, L.: Preferred provider organization claims showed high predictive value but missed substantial proportion of adults with high-risk conditions. J. Clin. Epidemiol. 58(6), 624–628 (2005)
4. Allebeck, P., Mastekaasa, A.: Chapter 5. Risk factors for sick leave-general studies. Scand. J. Public Health 32(suppl. 63), 49–108 (2004)
5. Andalusian Health Service. Instruction Manual of the Minimal Basic Data Set of Andalusia, 2009. Health Department of the Andalusian Government, Seville (2008), http://www.juntadeandalucia.es/servicioandaluzdesalud (accessed December 25, 2013)
6. Begg, C., Cho, M., Eastwood, S., et al.: Improving the quality of reporting of randomized controlled trials. The CONSORT Statement. JAMA 276(8), 637–639 (1996)
7. Benchimol, E.I., Guttmann, A., Griffiths, A.M., et al.: Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. Gut 58(11), 1490–1497 (2009)
8. Benchimol, E.I., Manuel, D.G., To, T., Griffiths, A.M., Rabeneck, L., Guttmann, A.: Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J. Clin. Epidemiol. 64(8), 821–829 (2011)
9. Birney, E.: The making of ENCODE: lessons for big-data projects. Nature 489(7414), 49–51 (2012)
10. Black, C.D., Burchill, C.A., Roos, L.L.: The population health information system: Data analysis and software. Med. Care 33(12 suppl.), DS127–DS131 (1995)

11. Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., et al.: Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 326(7379), 41–44 (2003)
12. Bratberg, E., Dahl, S.A., Risa, A.E.: "The double burden" – Do combinations of career and family obligations increase sickness absence among women? Eur. Sociol. Rev. 18, 233–249 (2002)
13. Brook, E., Rosman, D., Holman, C.D.J.: Public good through data linkage: measuring research outputs from the Western Australian data linkage system. Aust. N. Z. J. Public Health 32(1), 19–23 (2008)
14. Brown, S.H., Fischetti, L.F., Graham, G., et al.: Use of Electronic Health Records in Disaster Response: The Experience of Department of Veterans Affairs After Hurricane Katrina. Am. J. Public Health 97(1), S136–S141 (2007)
15. Cabanillas, J.L., Gili, M., Luanco, J.M., Villar, J.: Tiempo óptimo personalizado de incapacidad temporal por diagnóstico. Sevilla, Consejería de Salud y Bienestar Social de la Junta de Andalucía (2012)
16. Charlson, M.E., Pompei, P., Ales, K.L., et al.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic. Dis. 40(5), 373–383 (1987)
17. Charlson, M.E., Charlson, R.E., Peterson, J.C., Marinopoulos, S.S., Briggs, W.M., Hollenberg, J.: The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. J. Clin. Epidemiol. 61(12), 1234–1240 (2008)
18. Chen, G., Faris, P., Hemmelgarn, B., Walker, R.L., Quan, H.: Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Med. Res. Methodol. 9, 5 (2009)
19. Chevalier, A., Luce, D., Blanc, C., Goldberg, M.: Sickness absence at the French National Electric and Gas Company. Br. J. Ind. Med. 44(2), 101–110 (1987)
20. Chretien, J., Tomich, N.E., Gaydos, J.C., Kelley, P.W.: Real-Time Public Health Surveillance for Emergency Preparedness. Am. J. Pub. Health 99(8), 1360–1363 (2009)
21. Christen, P.: Data Matching. Concepts and techniques for Record Linkage, entity resolution and duplicate detection. Springer, Berlin (2012)
22. Cleves, M.A., Sanchez, N., Draheim, M.: Evaluation of two competing methods for calculating Charlson's comorbidity index when analyzing short-term mortality using administrative data. J. Clin. Epidemiol. 50(8), 903–908 (1997)
23. Copeland, K.T., Checkoway, H., McMichael, A.J., Holbrook, R.H.: Bias due to misclassification in the estimation of relative risk. Am. J. Epidemiol. 105(5), 488–495 (1977)
24. Cousens, S., Hargreaves, J., Bonell, C., et al.: Alternatives to Randomization in the Evaluation of Public-Health Interventions: Statistical Analysis and Causal Inference. J. Epidemiol. Comm. Health 65(7), 576–581 (2011)
25. De Coster, C., Quan, H., Finlayson, A., et al.: Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. BMC Health Serv. Res. 6, 77 (2006)
26. Delgado-Rodriguez, M., Llorca, J.: Bias. J. Epidemiol. Commun. Health 58(8), 635–641 (2004)
27. Deyo, R.A., Cherkin, D.C., Ciol, M.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol. 45(6), 613–619 (1992)
28. Eriksen, W., Natvig, B., Bruusgaard, D.: Marital disruption and long-term work disability. A four-year prospective study. Scand. J. Public Health 27(3), 196–202 (1999)

29. Eyal, A., Carel, R.S., Goldsmith, J.R.: Factors affecting long-term sick leave in an industrial population. Int. Arch. Occup. Environ. Health 66(4), 279–282 (1994)
30. Fazel, S., Wolf, A., Långström, N., Newton, C.R., Lichtenstein, P.: Premature mortality in epilepsy and the role of psychiatric comorbidity: a total population study. Lancet 382(9905), 1646–1654 (2013)
31. Feeney, A., North, F., Head, J., Canner, R., Marmot, M.: Socioeconomic and gender differentials in reason for sickness absence from the Whitehall II Study. Occup. Environ. Med. 55(2), 91–98 (1998)
32. Ferrante, A.M., Rosman, D.L., Knuiman, M.: The construction of a road injury database. Accid. Anal. Prev. 25(6), 659–665 (1993)
33. Fox, A.J., Goldblatt, P.O., Jones, D.R.: Social class mortality differentials: Artefact, selection or life circumstances? J. Epidemiol. Comm. Health 39(1), 1–8 (1985)
34. Fuhrer, R., Shipley, M.J., Chastang, J.F., et al.: Socioeconomic position, health, and possible explanations: a tale of two cohorts. Am. J. Public Health 92(8), 1290–1294 (2002)
35. Galea, S., Riddle, M., Kaplan, G.A.: Causal thinking and complex system approaches in epidemiology. Int. J. Epidemiol. 39(1), 97–106 (2010)
36. Garrett, N., Mishra, N., Nichols, B., Staes, C., Akin, C., Safran, C.: Characterization of Public Health Alerts and Their Suitability for Alerting in Electronic Health Record Systems. J. Pub. Health Manag. Practice 17(1), 77–83 (2011)
37. Ghali, W.A., Hall, R.E., Rosen, A.K., Ash, A.S., Moskowitz, M.A.: Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J. Clin. Epidemiol. 49(3), 273–278 (1996)
38. Gili, M., Sala, J., López, J., et al.: Impact of Comorbidities on In-Hospital Mortality From Acute Myocardial Infarction, 2003-2009. Rev. Esp. Cardiol. 64(12), 1130–1137 (2011)
39. Gill, L., Goldacre, M., Simmons, H., et al.: Computerised linking of medical records: Methodological guidelines. J. Epidemiol. Comm. Health 47(4), 316–319 (1993)
40. Gimeno, D., Benavides, F.G., Benach, J., Amick, B.: Distribution of sickness absence in the European Union countries. Occup. Environ. Med. 61(10), 867–869 (2004)
41. Goldacre, M., Shiwach, R., Yeates, D.: Estimating incidence and prevalence of treated psychiatric disorders from routine statistics: The example of schizophrenia in Oxfordshire. J. Epidemiol. Comm. Health 48(3), 318–322 (1994)
42. Goldstein, L.B., Samsa, G.P., Matchar, D.B., Horner, R.D.: Charlson index comorbidity adjustment for ischemic stroke outcome studies. Stroke 35(8), 1941–1945 (2004)
43. Greenland, S.: Quantifying Biases in Causal Models: Classical Confounding vs. Collider-Stratification Bias. Epidemiology 14(3), 300–306 (2003)
44. Guend, P., Engholm, G., Lynge, E.: Laryngeal cancer in Denmark A nationwide longitudinal study based on register linkage data. Br. J. Ind. Med. 47(7), 473–479 (1990)
45. Guttmann, A., Nakhla, M., Henderson, M., et al.: Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. Pediatr. Diabetes 11(2), 122–128 (2010)
46. Hanley, J., McNeil, B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1), 29–36 (1982)
47. Hemmelgarn, B.R., Manns, B.J., Quan, H., Ghali, W.A.: Adapting the Charlson comorbidity index for use in patients with ESRD. Am. J. Kidney Dis. 42(1), 125–132 (2003)
48. Hernan, M.A., Savitz, D.A.: From “Big Epidemiology” to “Colossal Epidemiology”: When all eggs are in one basket. Epidemiology 24(3), 344–345 (2013)

49. Hoffman, S., Podgurski, A.: Big bad data: law, public health, and biomedical databases. J. Law Med. Ethics 41(suppl. 1), 56–60 (2013)
50. Holman, C.D.J., Bass, A.J., Rouse, I.L., Hobbs, M.S.T.: Population-based linkage of health records in Western Australia: development of a health services research linked database. Aust. N. Z. J. Public Health 23(5), 453–459 (1999)
51. Hux, J.E., Ivis, F., Flintoft, V., Bica, A.: Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care 25(3), 512–516 (2002)
52. Iezzoni, L.I.: Reasons for risk adjustment. In: Iezzoni, L.I. (ed.) Risk Adjustment for Measuring Health Care Outcomes, 3rd edn., pp. 1–16. Health Administration Press, Chicago (2003)
53. Kendrick, S., Clarke, J.: The Scottish record linkage system. Health Bull. (Edinb.) 51(2), 72–79 (1993)
54. Khoury, M.J., Lam, T.K., Ioannidis, J.P.A., et al.: Transforming Epidemiology for 21st Century Medicine and Public Health. Cancer Epidemiol. Biomarkers Prev. 22(4), 508–516 (2013)
55. Kivimaki, M., Head, J., Ferrie, J.E., Shipley, M.J., Vahtera, J., Marmot, M.G.: Sickness absence as a global measure of health: evidence from mortality in the Whitehall II prospective cohort study. BMJ 327(7411), 364 (2003)
56. Lee, D.S., Donovan, L., Austin, P.C., et al.: Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med. Care 43(2), 182–188 (2005)
57. Leigh, J.: Correlates of absence from work due to illness. Human Relations 39(1), 81–100 (1986)
58. Levy, G., Blumberg, N., Kreiss, Y., Ash, N., Merin, O.: Application of Information Technology within a Field Hospital Deployment Following the Haiti Earthquake Disaster. J. Am. Med. Inf. Assoc. 17(6), 626–630 (2010)
59. Lynge, E., Thygesen, L.: Use of surveillance systems for occupational cancer: data from the Danish national system. Int. J. Epidemiol. 17(3), 493–500 (1988)
60. Manuel, D.G., Lim, J.J., Tanuseputro, P., Stukel, T.A.: How many people have had a myocardial infarction? Prevalence estimated using historical hospital data. BMC Public Health 7, 174 (2007)
61. Marmot, M., Feeney, A., Shipley, M., North, F., Syme, S.: Sickness absence as a measure of health status and functioning: from the UK Whitehall II study. J. Epidemiol. Comm. Health 49(2), 124–130 (1995)
62. Marshall, M.N., Shekelle, P.G., Leatherman, S., Brook, R.H.: The public release of performance data: what do we expect to gain? A review of the evidence. JAMA 283(14), 1866–1874 (2000)
63. Martin, C.A., Hobbs, M.S.T., Armstrong, B.K., de Klerk, N.H.: Trends in the incidence of myocardial infarction in Western Australia between 1971 and 1982. Am. J. Epidemiol. 129(4), 665–668 (1989)
64. Mastekaasa, A.: Parenthood, gender and sickness absence. Soc. Sci. Med. 50(12), 1827–1842 (2000)
65. McCallum, J., Lonergan, J., Raymond, C.: The NCEPH record linkage pilot study: a preliminary examination of individual Insurance Commission records with linked data sets. National Centre for Epidemiology and Public Health, Canberra (1993)
66. McGeechan, K., Kricker, A., Armstrong, B., Stubbs, J.: Evaluation of linked cancer registry and hospital records of breast cancer. Aust. N. Z. J. Public Health 22(7), 765–770 (1998)

67. Melton III, L.J.: History of the Rochester Epidemiology Project. Mayo Clin. Proc. 71(3), 266–274 (1996)
68. Mervis, J.: U.S. science policy. Agencies rally to tackle big data. Science 336(6077), 22 (2012)
69. Moncada, S., Navarro, A., Cortes, I., Molinero, E., Artazcoz, L.: Sickness leave, administrative category and gender: results from the "Casa Gran" project. Scand. J. Public Health 30(1), 26–33 (2002)
70. Myers, R.P., Quan, H., Hubbard, J.N., Shaheen, A.A.M., Kaplan, G.G.: Predicting in-hospital mortality in patients with cirrhosis: results differ across risk adjustment methods. Hepatology 49(2), 568–577 (2009)
71. National Institute of Standards and Technology Workshop. Cloud computing and big data (2013), http://www.nist.gov/itl/math/cloud-112912.cfm (accessed December 19, 2013)
72. Newgard, C.D., Zive, D., Jui, J., Weathers, C., Daya, M.: Electronic versus manual data processing: evaluating the use of electronic health records in out-of-hospital clinical research. Acad. Emerg. Med. 19(2), 217–227 (2012)
73. Overpeck, M.D., Hoffman, H.J., Prager, K.: The lowest birth-weight infants and the US infant mortality rate: NCHS 1983 linked birth/infant death data. Am. J. Public Health 82(3), 441–444 (1992)
74. Pappas, G., Hadden, W.C., Kozak, L.J., Fisher, G.F.: Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups. Am. J. Public Health 87(5), 811–816 (1997)
75. Pasternak, B., Svanstrom, H., Melbye, M., Hviid, A.: Association between oral fluoroquinolone use and retinal detachment. JAMA 310(20), 2184–2190 (2013)
76. Pechette, J.: Transforming health care through cloud computing. Health Care Law Mon. 2012(5), 2–12 (2012)
77. Poses, R.M., McClish, D.K., Smith, W.R., Bekes, C., Scott, W.: Prediction of survival of critically ill patients by admission comorbidity. J. Clin. Epidemiol. 49(7), 743–747 (1996)
78. Quach, S., Hennessy, D.A., Faris, P., Fong, A., Quan, H., Doig, C.: A comparison between the APACHE II and Charlson index score for predicting hospital mortality in critically ill patients. BMC Health Serv. Res. 9, 129 (2009)
79. Rice, D., Hodgson, T.A., Kopstein, A.N.: The economic costs of illness: a replication and update. Health Care Financ. Rev. 7(1), 61–80 (1985)
80. Romano, P.S., Roos, L.L., Jollis, J.G.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J. Clin. Epidemiol. 46(10), 1075–1079 (1993)
81. Romano, P.S., Roos, L.L., Jollis, J.G.: Further evidence concerning the use of a clinical comorbidity index with ICD-9-CM administrative data. J. Clin. Epidemiol. 46(10), 1085–1090 (1993)
82. Romano, P.S., Zhou, H.: Do well-publicized risk-adjusted outcomes reports affect hospital volume? Med. Care 42(4), 367–377 (2004)
83. Roos, N.P., Black, C.D., Frohlich, N., et al.: A population-based health information system. Med. Care 33(12 suppl.), DS13–DS20 (1995)
84. Ryan, D.H.: A Scottish record linkage study of risk factors in medical history and dementia outcome in hospital patients. Dementia 5(6), 339–347 (1994)
85. Sibthorpe, B., Kliewer, E., Smith, L.: Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential. Aust. J. Public Health 19(3), 250–256 (1995)

86. Smith, M.E., Newcombe, H.B.: Use of the Canadian Mortality Data Base for epidemiological follow-up. Can. J. Public Health 73(1), 39–46 (1982)
87. Southern, D.A., Quan, H., Ghali, W.: Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. Med. Care 42(4), 355–360 (2004)
88. Spasoff, R.A.: Epidemiologic methods for health policy. Oxford University Press, New York (1999)
89. Stanley, F.J., Croft, M., Gibbins, J., Read, A.W.: A population data base for maternal and child health research in Western Australia using record linkage. Paediatr. Perinatal Epidemiol. 8(4), 433–447 (1994)
90. Thomas, J.W., Holloway, J.J.: Investigating early readmission as an indicator of quality of care studies. Med. Care 29(4), 377–394 (1991)
91. To, T., Dell, S., Dick, P.T., et al.: Case verification of children with asthma in Ontario. Pediatr. Allergy Immunol. 17(1), 69–76 (2006)
92. Toh, S., Platt, R.: Is size the next big thing in Epidemiology? Epidemiology 24(3), 349–351 (2013)
93. Tyndall, R.M., Clarke, J.A., Shimmins, J.: An automated procedure for determining patient numbers from episodes of care records. Med. Inform. 12, 137–146 (1987)
94. Van der Brandt, P.A., Schouten, L.J., Goldbohm, R.A., et al.: Development of a record linkage protocol for use in the Dutch Cancer Registry for Epidemiological research. Int. J. Epidemiol. 19(3), 553–558 (1990)
95. Vistnes, J.P.: Gender differences in days lost from work due to illness. Ind. Labor Rel. Rev. 50, 304–323 (1997)
96. von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., Vandenbroucke, J.P.: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 4(10), e296 (2007)
97. Zhang, M., Holman, C.D.J., Price, S.D., Sanfilippo, F.M., Preen, D.B., Bulsara, M.K.: Comorbidity and repeat admission to hospital for adverse drug reactions in older adults: retrospective cohort study. BMJ 338, a2752 (2009)

Classification of ECG Cardiac Arrhythmias Using Bijective Soft Set

S. Udhaya Kumar and H. Hannah Inbarani

Abstract. This paper presents the new automated classification method for electrocardiogram (ECG) arrhythmia. Electrocardiogram datasets are generally called as big data. Big Data are the group of huge volumes of unstructured data. Big Data means enormous amounts of data, such large that it is difficult to collect, store, manage, analyze, predict, visualize, and model the data. Electrocardiography deals with the electrical movement of the heart. The order of cardiac health is given by ECG and heart rate. A study of the nonlinear dynamics of electrocardiogram (ECG) signals for arrhythmia characterization is considered in this work. Cardiac problems are considered to be the most deadly disease in medical world. Cardiac arrhythmia is abnormality of heart rhythm, in fact refers to disorder in electrical conduction system of the heart. In this paper, computerized ECG interpretations are used to identify arrhythmias. It is a process of ECG signal acquisition, eliminating noise (De-noising) from ECG signal, detecting wave parameters (P, Q, R, S and T) and rhythm classification. Substantial progress has been made over the years in improvising techniques for signal conditioning, extraction of relevant wave parameters and rhythm classification. However, many problems and issues, especially those related to detection of multiple arrhythmic events using soft computing techniques is still need to be addressed in a broader manner to improve the prospect of commercial automated arrhythmia analysis in mass health care centres. The main objective of this paper is to present a classifier system based on Bijective soft set in order to classify ECG signal data into five classes (Normal, Left bundle branch blocks, Right bundle branch blocks, premature ventricular contractions and Paced rhythm class). To complete this objective, an algorithm for detection of P, QRS and T waves are applied followed by IBISOCLASS Classifier. The experimental results are acquired by examining the proposed method on ECG data from the MIT-BIH arrhythmia database. The proposed algorithm is also compared with the well-known standard classification algorithms namely Back propagation network (BPN), Decision table, J48 and Naïve Bayes.

H. Hannah Inbarani · S. Udhaya Kumar
Department of Computer science, Periyar University,
Salem-636011
e-mail: hhinba@gmail.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_11

Keywords: Soft Set, Bijective Soft Set, Classification, PQRST detection, De-noising ECG signal, Pan-Tompkins algorithm and Cardiac arrhythmia.

1 Introduction

Research in Big data system for medical diagnosis is a significant and sensational domain (Azar and Hassanien 2014). Medical data feature selection and classification is one of the most important problems in many decision-making tasks (Inbarani et al. 2014). Decision-making tasks are instances of classification problem that can be easily formulated into a prediction or forecasting tasks, diagnosis tasks, and pattern recognition (Azar 2014). Cardiac diseases are still in a high level for death reason among human and they are huge problem which need to be resolved. Unexpected death of the patient from heart diseases can be prevented by using early diagnose, computer aided tools, and medical treatment methods. One of the ways to diagnose the cardiovascular diseases is to use electrocardiogram (ECG) (Özbay 2009). Electrocardiogram (ECG) denotes the electrical movement of the heart showing the regular contraction and relaxation of heart muscle. The pattern recognition and classification of the ECG beats is a most essential task in the coronary intensive unit, wherever the classification of the ECG beats is important tool for the diagnosis. ECG deals cardiologists with suitable information about the rhythm and working of the heart. The analysis of ECG rhythm is efficient way to detect the different heart irregularities. Still now many algorithms have been established for the feature selection (Azar et al. 2013) and classification Heart rate signals (Osowski and Linh 2001). The ECG signal may be different for the same patient to such extent that they are dissimilar each other and at the same time similar for different types of beats (Saxena et al. 2002). An ECG waveform contains five basic waves P, QRS and T waves. Fig. 1 shows a typical ECG Wave form. The Q, R and S wave is called as QRS complex which represents the ventricular depolarization. The P wave represents atrial depolarization and T wave represents the repolarization of ventricle (Maglaveras et al. 1998).

In a clinical setting, such as an intensive care unit, it is necessary for automated system to exactly detect and classify electrocardiographic signals on a real time basis. Early recognition and action of the heart diseases can save the patient's life or avoid permanent damages on tissues of the heart. Until now, several studies for automatic arrhythmia detection and classification have been implemented for increasing the ECG beat classification accuracy. In ECG beats classification most of these systems are developed in two steps: feature extraction and pattern classification. The first step which is ECG features extraction has been accomplished by either in the time domain to obtain morphological of features (such as QRS complex, P wave and T wave detection, heart-rate variability etc...) (Chazal et al. 2000; Giovanni et al. 2001), or in the frequency domain to discover variations in QRS-complex power spectra between normal and arrhythmia

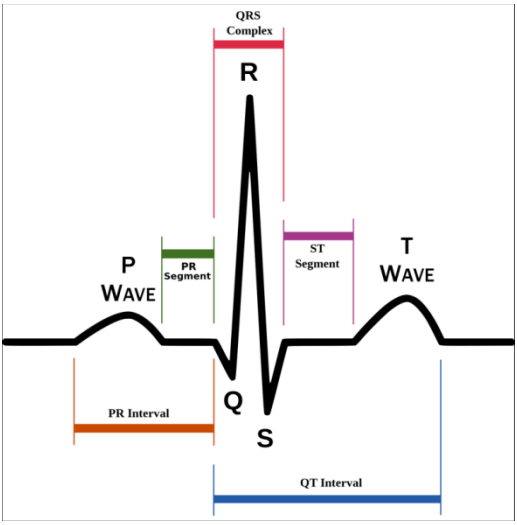

Fig. 1 Basic Structure of ECG Waveform

waveforms (Minami et al. 1999; Qin et al. 2003), time frequency domain (Lin et al. 2008; Dickhaus and Heinrich 1996) to show concurrently ECG frequency and time features. The second step: classification has been established by several methods, such as the Artificial Neural Network (ANN) and neuro-fuzzy based abnormal beat classification, self-organizing map (Marcel et al. 1997), Wavelet coefficient (Chazal et al. 2000) and RBF Neural Networks.

In this paper, we propose an automated method for ECG heartbeats classification. Five different heartbeats are considered: N (Normal), PVC (Premature ventricular contractions), LBBB (Left bundle branch blocks) RBBB (Right bundle branch blocks) and PR (paced rhythm). For the feature extraction step, we used Pan – Tompkins algorithm for identify the morphological features P, QRS complex and T waves. This algorithm is based on the technique introduced by Pan and Tompkins (Pan & Tompkins 1985). For the classification step, Improved Bijective soft set classification algorithm (IBISOCLASS) is applied for classification of five different beats. The developed algorithms are implemented and evaluated on MIT-BIH arrhythmia database. The obtained results are compared with other existing algorithms.

The rest of this paper is organized as follows: In section 2 provides a review of related work. Section 3 presents materials and methods used in cardiac arrhythmias classification. Section 4 explained about the Pan – Tompkins algorithm. Section 5 describes the basic concepts of soft set and Bijective soft set. Section 6 explains other classification algorithms used in this work for comparative analysis. In Section 7, experimental results have been reported. Finally the conclusion has been addressed in Section 8.

2 Related Work

In the literature, several approaches have been proposed for the automatic classification of ECG signals. In this section, a brief survey of the most recently published works are reported in Table 1 focusing on the classification of ECG signals using various feature extraction and classification methods.

Table 1 Related Work for this study

Authors	Purpose	Techniques
Mai et al. 2011	Classification	Multilayer Perceptron and Radial basis function neural network for ECG classification. Multilayer Perceptron accuracy – 98% Radial basis function accuracy – 97%
Melgani and Yakoub Bazi 2008	Classification	Hybrid techniques of Support vector machine and particle swarm optimization for classification of signals. Based on 20 patients' records the classification accuracy PSO-SVM – 89.72%, SVM – 85.98%, KNN – 83.70% and RBF – 82.34%
Rai et al. 2013	Feature Extraction and classification	ECG signal processing for abnormalities detection using Multi - resolution wavelet transform and Artificial Neural Network. Normal and Abnormal classes are used. The classification accuracy MLP – 100% , BPN – 97.8% and FFN – 97.8%
Issac Niwas et al. 2005	Classification	In this work, Classification of ECG signal into normal beat and nine different arrhythmias. The overall accuracy was 99.02% of the proposed approach
Benali et al. 2012	Classification	In this paper, automatic classification based on wavelet neural network can be considered as an effective tool for cardiac arrhythmias classification with high accuracy of over 98.78%. Wavelet Neural Network (WNN)

Table 1 (*continued*)

Nazmy et al. 2010	Classification	This work focused on six different ECG signals and the proposed ANFIS model gives more than 97% classification accuracy. Adaptive Neuro-Fuzzy Inference System (ANFIS)
Mitra et al. 2006	Rules Generation	Rule-based rough-set decision system is obtained for the improvement of an inference engine for ECG beat disease identification. 10 ECG morphological features Stationary wavelet Transform (SWT) Multilayer Perceptron accuracy(MLP)
Inan et al. 2006	Classification	Robust Neural-Network-Based Classification approach applied for Premature Ventricular Contractions and the accuracy was 95.16% over from all 40 files and 96.82% over the 22 files. Time interval features Wavelet transform
Hassan et al. 2011	Classification	In this paper, the authors propose a novel method of Fuzzy C- Means clustered PNN to differentiate eight types of ECG signals. Probabilistic Neural Network (PNN) Multi Layered Feed Forward Network (MLFFN)
Homaeinezhad et al. 2012	Feature extraction and classification	The following classifiers are used in this work. K-Nearest Neighbors classification (KNN) Radial basis based support vector machine (RBF-SVM) Neuro-SVM–KNN fusion classification algorithm
Wen et al. 2009	Classification	In this study, unsupervised Self-organizing Cerebellar Model Articulation controller (SOCMAC) network to design an ECG classifier by observing the QRS complex of each heartbeat.

Table 1 (*continued*)

Yu and Chou 2008	Feature selection and classification	In this paper, authors proposed a scheme to combine independent component analysis (ICA), RR interval and Neural network classifiers for ECG beat classification. Probabilistic Neural network(PNN) Back – Propagation neural network(BPNN)
Prasad et al. 2003	Classification	The Proposed method is classified as the normal sinus rhythm and 12 different arrhythmias. The overall accuracy of ANN classification of the proposed approach is 96.77%
Pan and Tompkins 1985	Feature Extraction	In this paper, author extracted QRS complex of ECG signal.
Özbay 2009	Feature extraction and classification	The following methods are used for feature extraction and classification of MIT-BIH databases. Complex wavelet transform (CWT) Complex valued ANN(CVANN)
Özbay et al. 2006	Classification	This study presents a proportional study of the classification accuracy of ECG signals using a NN architecture of multi-layered perceptron (MLP) with back propagation training algorithm, and a new fuzzy clustering NN architecture (FCNN) for early identification.
Gacek and Pedrycz 2006	Classification	In this paper, authors established a broad outline of a granular representation of ECG signals. Granules – Fuzzy set Fuzzy C-Means
Mehmet Engin 2004	Classification	The application on the fuzzy-hybrid neural network is applied for electrocardiogram (ECG) beat classification. Fuzzy C - Means classifier MLP neural network
Chazal et al. 2000	Classification	This study investigates the automatic classification of the Frank lead electrocardiogram (ECG) into different pathophysiological disease categories. Wavelet coefficients classifier
Karpagachelvi et al. 2012	Classification	In this paper, Extreme Learning Machine (ELM) is presented and compared with support vector machine (SVM)

Table 1 (*continued*)

Senthilkumar et al. 2014	Classification	In this paper, authors proposed a novel approach for medical data classification based on Modified soft rough set based classification approach.
Inbarani et al. 2013	Feature selection	In this paper, Authors proposed hybridized method for digital mammogram images. Tolerance Rough Set - PSO based Quick Reduct (STRSPSO-QR) Tolerance Rough Set - PSO based Relative Reduct (STRSPSO-RR),

3 Materials and Methods

Cardiac arrhythmia is any changes in regular rhythmic beating of the heart. The heart beat may be regular or irregular, and too fast or too slow. In this paper, the cardiac classes that take in both the normal and other cardiac abnormalities are categorized into five types (Hassan et al. 2011). They are

(i) Normal Beat (N)
(ii) Left Bundle Branch Block (LBBB)
(iii) Right Bundle Branch Block (RBBB)
(iv) Premature Ventricular Contraction (PVC)
(v) Paced Beat (PB)

The following section springs a brief explanation of the different cardiac classes.

(i) **Normal**: It happens due to the continuous, intervallic performance of the pacemaker and the integrity of the neuronal conducting pathways. In this normal beat, the QRS complex duration never goes above 0.12s and PR interval should not exceed 0.20s. The P wave has a maximum duration of 0.08s and the T wave should be at least 0.20 s. For a healthy person, the heart beat ranges from 60 to 100 BPM (Beats Per Minute), so the R-R interval should be between 0.6 to 1s (Inan et al. 2006).
(ii) **LBBB**: In this situation, initiation of the left ventricle is late, which results in the left ventricle contracting later than the right ventricle and the condition results in a QRS interval greater than 0.12 s (Inan et al. 2006).
(iii) **RBBB**: In this state, the two ventricles no longer receive the electrical impulse simultaneously. The duration of the QRS complex on the ECG is in between 0.10 and 0.11 s (for incomplete RBBB) or 0.12 s or more (for complete RBBB) and has a lengthy ventricular activation time or QR interval of 0.3 s or more.
(iv) **PVC**: The heart beat ensues before than expected and interrupts the regular rhythm of the heart. Irregularity of the rhythm, the P waves obscured by the QRS complex or the T wave, widening of the QRS complex, the opposite polarity of the T wave with respect to the R wave are the significant characteristic features of PVC (Inan et al. 2006).

(v) **PB**: An area in the excitable ventricular musculature try to control the heart beat conduction and results in slower heart rate that ranging between 30 and 50 bpm. Slow heart beat can lead to weakness, confusion, dizzying, collapsing, shortness of breath and death.

The methodology adopted in this work is shown in Fig. 2. It can be realized that the whole procedure is divided into three basic parts that is Signal Acquisition, Preprocessing (De-noising), Feature extraction and classification. The ECG signal is acquired from the MIT-BIH arrhythmia database. The novel ECG signal must be pre-processed with the purpose of eliminating noises in ECG signal. After the elimination of noise, the features are identified as P, QRS and T. Final Step of the method is to classify the signal into the five different classes.

3.1 Signal Acquisition

This is first stage of signal processing; database collection is the one of the most significant task of signal processing. For this research work, MIT–BIH Arrhythmia database directory of ECG signals from physioNet is used. The source of the ECGs

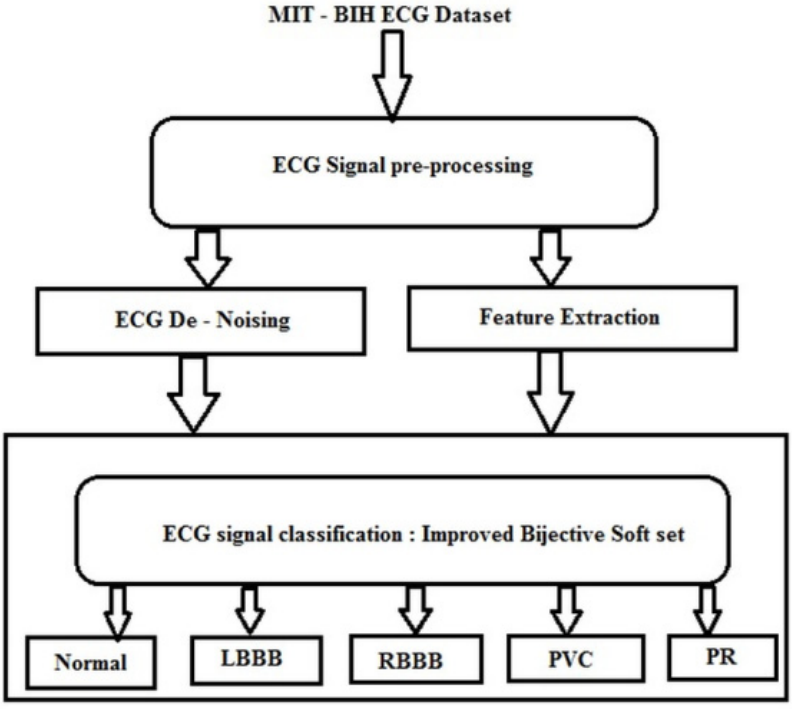

Fig. 2 Proposed Method

of MIT–BIH Arrhythmia was obtained by the Beth Israel Hospital Arrhythmia Laboratory. This database contains 48 files divided into two portions first one is of 23 files (Records number 100 to 124 inclusive with some of missing records) selected at random from this set, and another one contains 25 files (numbered from 200 to 234 inclusive with some numbers of lacking). Each of the 48 records is slightly over 30 min long (Mark and Moody 1988; Moody and Mark 2001).

3.2 Signal Pre-processing and Feature Extraction

This is the next stage of ECG signal processing, it's necessary to remove noises from input signals using Pan and Tompkins algorithm. For pre-processing of the ECG signal, noise removal involves different approaches for various noise sources. Before feature extraction, signal must be De-noised for increasing the system classification accuracy. Fig. 3 shows the original or noisy ECG signal taken from MIT-BIH laboratory.

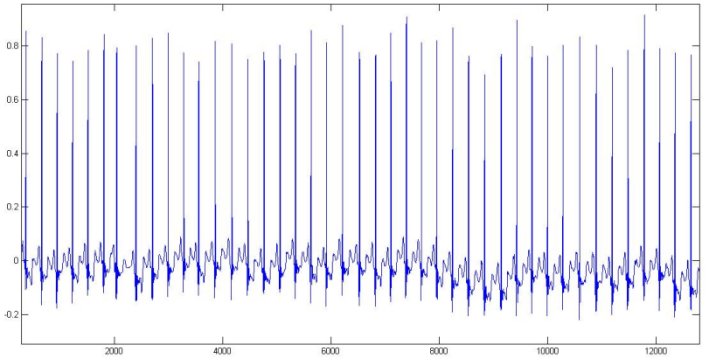

Fig. 3 Noisy MIT-BIH ECG signal

Pan and Tompkins algorithm (Pan and Tompkins 1985) filters the ECG signal with help of Low-pass filter and High-pass filters. Fig. 4 illustrates the filtered ECG signal. After filtering, the signal moves to Derivation part, in this phase the signal is differentiated to provide the QRS complex slope information.

After differentiation, the signal is squared point by point in Squaring function. The purpose of moving-window combination is to get waveform feature information in totaling to the slope of the R wave. The QRS complex matches to the increasing edge of the combination waveform. The time interval of the increasing edge is equivalent to the width of the QRS complex. A Fiducial mark of Pan and Tompkins algorithm detects QRS complex and we find P and T wave based on adjusting threshold values of P and T. The desired waveform features to be marked such as the maximal rise or the peak of the R wave according to the increasing edges of temporal locations of QRS complex. Fig. 5 shows detected values of P, QRS and T.

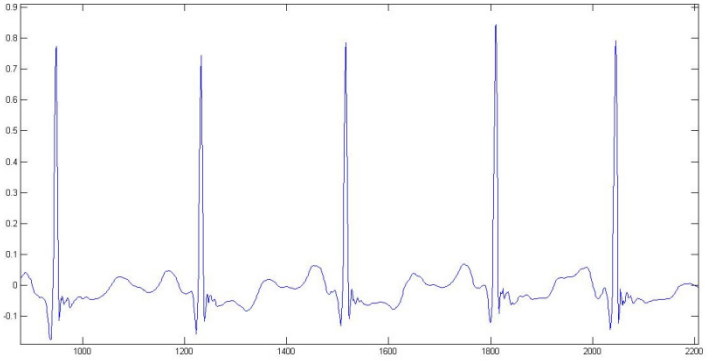

Fig. 4 Filtered MIT-BIH ECG signal

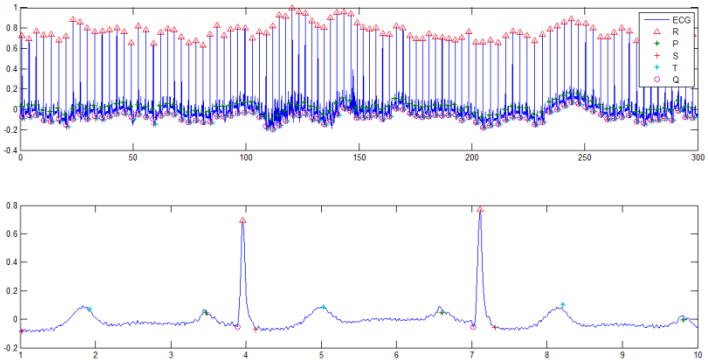

Fig. 5 P, QRS complex and T detection

3.3 Improved Bijective Soft Set – Proposed and Applied

Improved Bijective soft set is applied for signal classification. There are two types of rules generated by using improved bijective soft set classification algorithm. First type of rule is Deterministic rule (certain rule), the certain rules are generated using AND & restricted AND Operation. Second type of rule is Non-Deterministic rule (Possible rule), possible rules are achieved by using AND & Relaxed AND Operations. For each Non-Deterministic rule, support is computed (Udhayakumar et al. 2013). Improved Bijective soft set based classification approach is presented in algorithm 1. The basic concepts of soft sets and Bijective soft sets are explained in section 5.

Algorithm 1. Improved Bijective Soft set based classification

Input: Given Dataset with conditional attributes 1, 2 . . . n-1 and Decision attribute n.

Output: A set of Rules R

Step 1: Construct Bijective soft set for all conditional attributes (F_i, E_i) for i=1 to n-1, n is the number of Attributes.

Step 2: Construct Bijective soft set for decision attribute (G, B).

Step 3: Apply AND on the Bijective soft set (F_i, E_i). Result is stored in (H, C).

Step 4: Generate deterministic rules using Restricted AND operation.

$$U_{e \in E} \{F(e) : F(e) \subseteq X$$

Step 5: Generate Non-Deterministic rules using Relaxed AND operation.

$$U_{e \in E} \{F(e) : F(e) \cap X \neq \emptyset\}$$

Step6: Compute the support value for each non-deterministic rule using

$$\boldsymbol{support} = \frac{\boldsymbol{support(A \wedge B)}}{\boldsymbol{support(U)}}$$

Table 1 represents a sample of the data set as an example in order to extract the rules. Let A= {A_1, A_2} be the set of condition attributes and D is the decision attribute. In decision attribute {B, M} stands for Benign and Malignant (Udhayakumar et al. 2014).

Table 2 Sample data set

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X10
A_1	1	1	1	1	2	2	2	2	2	1
A_2	1	2	1	2	1	1	2	2	2	1
D	B	B	B	B	M	M	B	M	M	B

The proposed approach is explained with an example given in Table 1.

Step 1: Construct Bijective Soft set from Conditional attributes.

$(F_1, A_1) = \{X_1, X_2, X_3, X_4, X_9, X_{10}\}\ \{X_5, X_6, X_7, X_8\}$

$(F_1, A_2) = \{X_1, X_3, X_5, X_6, X_9, X_{10}\}\ \{X_2, X_4, X_7, X_8\}$

Step 2: Construct Bijective Soft set from Decision attribute.

D (Benign) = $\{X_1, X_2, X_3, X_4, X_7, X_9, X_{10}\}$

D (Malignant) =$\{X_5, X_6, X_8\}$

Step 3: Apply AND operation on the Bijective soft set (F_i, A_i). Table 3 shows tabular form of this AND operation.

Table 3 The tabular form of $(F_1, A_1) \wedge (F_1, A_2)$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X10
E_1	1	0	1	0	0	0	0	0	1	1
E_2	0	1	0	1	0	0	0	0	0	0
E_3	0	0	0	0	1	1	0	0	0	0
E_4	0	0	0	0	0	0	1	1	0	0

$(F_1, A_1) \wedge (F_1, A_2) = (H, C) = \{\{ X_1, X_3, X_9 X_{10}\}, \{ X_5, X_6\}, \{X_7, X_8\}, \{X_2, X_4\} \}$

Step 4: Generate deterministic rules by using $U_{e \in E}\{F(e) : F(e) \subseteq X$

$(H, C)\hat{\sim}D$ (Benign) = $\{\{X_1, X_3, X_9, X_{10}\}, \{X_2, X_4\}\}$

$(H, C)\hat{\sim}D$ (Malignant) = $\{X_5, X_6\}$

$(F, B) = \{\{X_1, X_3, X_9, X_{10}\}, \{X_5, X_6\}, \{X_2, X_4\}\}$

If $A_1 = 1$ and $A_2 = 1 \Rightarrow$ d = Benign

If $A_1 = 1$ and $A_2 = 2 \Rightarrow$ d = Benign

If $A_2 = 2$ and $A_2 = 1 \Rightarrow$ d = Malignant

Step 5: Generate non-deterministic rules using $U_{e \in E}\{F(e): F(e) \cap X \neq \emptyset\}$

$(F, B) = \{\{X_1, X_3, X_7, X_{10}\}, \{X_2, X_4\}, \{X_5, X_6, X_8\}\}$

If $A_1 = 1$ and $A_2 = 1 \Rightarrow$ d = Benign

If $A_1 = 1$ and $A_2 = 2 \Rightarrow$ d = Benign

If $A_2 = 2$ and $A_2 = 1 \Rightarrow$ d = Malignant

If $A_1 = 2$ and $A_2 = 2 \Rightarrow$ d = Benign

If $A_1 = 2$ and $A_2 = 2 \Rightarrow$ d = Malignant

Step 6: Support (Benign, Malignant) = $(\frac{1}{10} \wedge \frac{2}{10}) = 0.2$

4 Pan – Tompkins Algorithm

A graphical representation of the Pan – Tompkins algorithm is illustrated in figure 6. The signal is passed in Pan – Tompkins algorithm through filtering, derivatives, squaring and integration phase before thresholds are set and QRS complexes are detected. Pan – Tompkins algorithm detects only QRS complexes but in this paper, along with QRS complexes, P and T features are also detected.

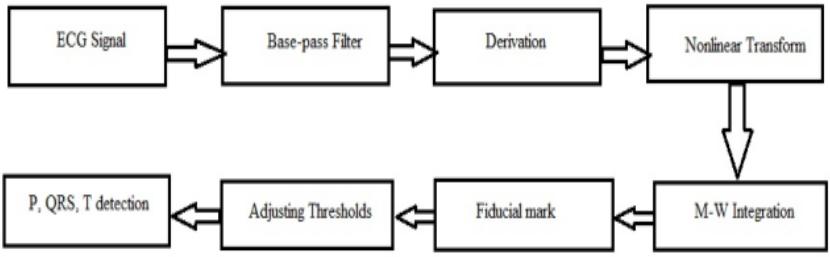

Fig. 6 Graphical representation of Pan – Tompkins algorithm

4.1 Band-Pass Filter

The band-pass filter is two filters spilled, one is low pass filter and the other one is high pass filter. Low pass filter is applied to eliminate noise such as the EMG and 50Hz power line noise. The transfer function of the second-order low pass filter used is

$$H(z) = \frac{(1-z^{-6})^2}{(1-z^{-1})^2} \tag{1}$$

The difference equation of the low pass filter is

$$y(nT) = 2y(nT\text{-}T) + y(nT\text{-}2T) + x(nT) - 2x(nT\text{-}6T) + x(nT\text{-}12T) \tag{2}$$

Where $x(n)$ is the input differentiated ECG and $y(n)$ is the band-passed ECG. T is the sampling period. This filter has purely linear phase response. The power line noise is significantly attenuated by this filter. The transfer function of the high pass filter is given by

$$H_{lp}(z) = \frac{(-1+32z^{-16}+z^{-32})}{(1+z^{-1})} \tag{3}$$

The difference equation of the High pass filter is

$$y[n] = y[n\text{-}1] - x[n]/32 + x[n\text{-}16] - x[n\text{-}17] + x[n\text{-}32]/32 \tag{4}$$

4.2 Derivatives

Differentiation of the filtered signals is to provide the slope information of QRS complex since there are quick rise and fall times of the QRS complex in the ECG signals, taking the derivative of the ECG would make it easier to detect when the QRS complex occurs.

The transfer function of the five-point differentiation equation is given by

$$H(Z) = \left(\frac{1}{8T}\right)(-z^{-2} - 2z^{-1} + 2z^{1} + z^{2}) \tag{5}$$

4.3 Nonlinear Transform

When differentiation was completed, the signal is squared point by point. The equation of this process is

$$Y(nT) = [x(nT)]^2 \quad (6)$$

This makes all data points positive and does nonlinear amplification of the output of the derivative highlighting the greater frequencies (i.e., predominantly the ECG frequencies).

4.4 Moving-Window Integration

The squared waveform passes over a moving window integrator. The determination of moving-window combination is to get waveform feature information in addition to the slope of the R-wave. It is calculated from following equation and N is the width of integration window.

$$Y(T) = \left(\frac{1}{N}\right)[x(nT-(N-1)T) + x(nT-(N-2)T) + \ldots\ldots\ldots + x(nT)] \quad (7)$$

4.5 Fiducial Marks

The QRS complex is related to the increasing edge of the integration waveform. Width of the QRS complex and rising edges are of the same time duration. The desired waveform features to be marked such as the maximal rise or the peak of the R wave are marked according to the increasing edges of temporal locations of QRS complex. Based on the noise thresholds, which are automatically adjusted, Low threshold values are possible because of the enhancement of the signal–to–noise ratio via band-pass filter (Portet et al. 2005).

5 Basic Concept – Soft Set and Bijective Soft Set

In this section, we describe the basic notions of soft sets and Bijective soft sets. Let U be initial universe of objects and E be a set of parameters in relation to objects in U. Parameters are often attributes, characteristics or properties of objects.

5.1 Soft Set Theory

Molodtsov (Molodtsov 1999) presented the concept of soft set theory which can be used as general mathematical tools for management with uncertainty.

Definition 1:
A pair (F, E) is called a soft set (over U) if and only if F is a mapping of E into the set of all subsets of the set U, where F is mapping given by

$$F: E \rightarrow P(U) \tag{8}$$

In other words, the soft set is a parameterized family of subsets of the set U. Every set F (ε) ($\varepsilon \in$E), from this family may be considered as the set of ε-elements of the soft sets (F, E), or as the set of ε-approximate elements of the soft set.

5.2 Bijective Soft Set Theory

Ke Gong et al. introduced a new type of soft set known as Bijective soft set. Form the notion of Bijective soft set, every element can be only mapped into one parameter and the union of partition by parameter set is universe. Based on the notion of Bijective soft set, he proposes some of its operations to study the relationship between Bijective soft sets (Gong et al. 2008).

Throughout this section U refers to an initial universe, E is a set of parameters; P (U) is the power set of and A ⊆ E (Gong et al. 2008).

Definition 2:
Let (F, B) be a soft set over a common universe U, where F is a mapping $F: B \rightarrow P(U)$ and B is nonempty parameter set. We say that (F, B) is a Bijective soft set, if (F, B) such that

(i) $U_{\varepsilon \in B} F(e) = U$.

(ii)For any two parameters $e_i, e_j \in B$, $e_i \neq e_j$, $F(e_i) \cap F(e_j) = \emptyset$

In other words, suppose Y ⊆ P (U) and Y = {F (e_1), F (e_2) . . . F (e_n)}, e_1, e_2. . .$e_{n \in B}$. From **Definition 2**, the mapping $F: B \rightarrow P(U)$ can be transformed to the mapping $F: B \rightarrow Y$, which is a Bijective function. i.e. for every $y \in Y$, there is exactly one parameter $e \in B$ such that F (e) = y and no unmapped element remains in both B & Y.

Definition 3 (AND operation):
AND operation on two soft sets. If (F, A) and (G, B) are two soft sets then ''(F, A) AND (G, B)'' denoted by $(F, A) \wedge (G, B)$ is defined by $(F, A) \wedge (G, B) = (H, A \times B)$, where $H(\alpha, \beta) = F(\alpha) \cap G(\beta)$, $\forall (\alpha, \beta) \in A \times B$.

Definition 4 (Restricted AND Operation):
Let $U = \{x_1, x_2, \ldots, x_n\}$ be a common universe, X be a subset of U, and (F, E) be a Bijective soft set over U. The operation of ''(F, E) restricted AND X'' denoted by $(F, E)_{\sim}^{\wedge} X$ is defined by $U_{e \in E} \{F(e): F(e) \subseteq X$.

Definition 5 (Relax AND operation):
Let U = { x_1, x_2, … ,x_n } be a common universe, X be a subset of U, and (F , E) be a Bijective soft set over U. The operation of ''(F, E) relaxed AND X'' denoted by (F, E) $\widetilde{\wedge}X$ is defined by $U_{e \in E}$ {$F(e)$: $F(e) \cap X \neq \emptyset$}.

6 ECG Signal Comparative Classification Algorithms

After performing noise elimination and feature extraction, it is necessary to classify the ECG signal in order to predict the signal type. Back propagation neural network, Naïve Bayes, Decision table, J48 and Improved Bijective soft set classification algorithm are applied for signal classification.

6.1 Back Propagation Neural Network

Back propagation algorithm is used to train the Multilayer neural networks (Alejo et al. 2012). As functional signals flow in forward direction and error signals propagate in backward direction, this is also known as back propagation network. Computation between hidden and output neurons, different activation function is applied like sigmoidal activation function. The algorithm is based on error-correction technique. The rule for updating synaptic weights for training neural network follows a generalized delta rule (Jinkwon et al. 2009). In general we consider three layered network i.e. network consisting of one input layer, one hidden layer and one output layer for classification.

Hidden neurons:

$$net_i^{(1)} = \sum_{j=1}^{n0} w_{ij}^{(1)} x_j \tag{9}$$

$$a_i^{(1)} = f(net_i^{(1)}), i = 1,2,\ldots.,n1 \tag{10}$$

Output neurons:

$$net^{(2)} = \sum_{i=1}^{n1} w_{1i}^{(2)} a_i^{(1)} \tag{11}$$

$$a^{(2)} = f(net^{(2)}) \tag{12}$$

The activation functions f (net) can be linear ones or Fermi functions of type

$$f(net) = \frac{1}{1+e^{-4\sigma(net-\sigma)}} \tag{13}$$

Analogous to the writing and reading phases, there are also two phases in the supervised learning BP network. There is a training phase when a training data set is used to define the weights that describe the neural model. So the task of the BP algorithm is to find the optimal weights to minimise the error between the target value and the actual response. Then the trained neural model will be used later in the retrieving phase to process and evaluate real patterns. Let the pattern's output

equivalent to the vector of pattern's inputs X be called the target value t (Jing et al. 2012). Then the learning of the neural network for training pattern (X, t) is performed in order to minimize the squared-error between the target and the actual response

$$E = \frac{1}{2}(t - a^{(2)})^2 \tag{14}$$

The weights are changed according to the following formula:

W new

$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij} \tag{15}$$

Where

$$\Delta w_{ij=n\delta_i^{(l)}}\, a_j^{(l)} \tag{16}$$

Output layer

$$\delta^{(2)} = f'(net^{(2)})(t - a^2) \tag{17}$$

Where the derivation of the activation function

$$f'(net) = 4net(1 - net) \tag{18}$$

Hidden layer

$$\delta_i^{(1)} = f'(net_i^{(1)}\delta^{(2)}w_{1i}^{(2)} \tag{19}$$

Algorithm 2. Back propagation neural network Algorithm

Initialize network weights (often small random values)
do
for each training data
Prediction = neural-net-output (network, ex) *// forward pass*
Actual = teacher-output(ex)
Compute error (prediction - actual) at the output units
Compute $\Delta \boldsymbol{W}$ for all weights from hidden layer to output layer
Compute $\Delta \boldsymbol{W}$ for all weights from input layer to hidden layer
update network weights
until all examples are classified correctly or another stopping criterion satisfied
return the network

6.2 Naïve Bayesian Classifier

A naïve Bayesian classifier based on Bayes theorem is a probabilistic statistical classifier. Here, the term "naïve" designates conditional independence among features or attributes (Dong et al. 2011).

Naïve Bayes classifier is very easy to build, and does not require any problematical iterative parameter estimation schemes. This means it may be gladly applied to large amount of data sets. It is easy to interpret, so inexperienced users also understand how to classify the dataset. Naïve Bayesian classifier may not Probabilistic approaches to classification typically involve shows the conditional probability distribution $P(C|D)$, where C ranges over classes and D over descriptions, in some language, of objects to be classified. Given a description d of a particular object, we assign the class $argmax_c\ P(C = c|D = d)$. A Bayesian approach splits this posterior distribution into a prior distribution $P(C)$ and likelihood $P(D|C)$:

$$argmax_c\ P(C = c|D = d) = \text{argmax}_c \frac{P(D = d|C = c)\text{P(C=c)}}{P(D=d)} \tag{20}$$

The denominator $P(D = d)$ is a normalizing factor that can be passed over when determining the maximum a posteriori class, as it does not depend on the class. The key term in Equation (20) is $P(D = d|C = c)$, the probability of the given explanation of the class. A Bayesian classifier estimates these likelihoods from training data, but this typically needs some additional simpler assumptions. For instance, in an attribute-value demonstration (also called *propositional* or single-table representation), the individual is described by a vector of values $a1, \ldots,$ $an$for a fixed set of attributes $A1, \ldots, An$. Determining $P(D = d|C = c)$ here requires an estimate of the joint probability $P(A1 = a1, \ldots, An = an|C = c)$, abbreviated to $P(a1, \ldots, an|c)$.

This joint probability distribution is problematic for two reasons:

(1) Its size is exponential in the number of attributes n, and
(2) It requires a complete training set, with several examples for each possible description. These problems vanish if we can assume that all attributes are independent given the class: n

$$P(A_1 = a_1, \ldots, A_n = a_n|C = c) = \prod_{i=1}^{n} P(A = a|C = c) \tag{21}$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the naive Bayesian classifier, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class (Dong et al. 2011).

6.3 Decision Tree - J48

J48 is a type of Decision tree algorithm developed by J. Ross Quinlan, the very common C4.5. Decision trees are a standard way to denote information from a machine learning algorithm, and offer a firm and powerful method to express structures in data (Charfi and Kraiem 2012).

It is important to understand the variety of options available when using this algorithm, as they can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in others, each choice may require some consideration.

Algorithm 3. J48 Classification Algorithm

```
INPUT  :  D //Training data
OUTPUT:  T //Decision tree
DTBUILD (*D)
{
  T = φ;
  T = Create root node and label with splitting attribute;
  T = Add arc to root node for each split predicate and label;
For each arc do
   D = Database created by applying splitting predicate to D;
   If stopping point reached for this path, then
    T'= create leaf node and label with appropriate class;
Else
    T'= DTBUILD (D);
    T= add T' to arc;
}
```

J48 algorithm deals with a number of options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more simply interpreted results. More essentially, pruning can be used as a tool to correct for possible over fitting. The basic algorithm recursively classifies up to each leaf is pure, meaning that the data has been characterized as nearby to perfectly as possible. This method offers extreme accuracy on the training data, but it may create extreme rules that only describe particular features of that data. Pruning always decreases the accuracy of a model on training data. The overall concept is to slowly generalize a decision tree until it gains a balance of flexibility and accuracy. While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample (Minghao et al. 2012).

6.4 Decision Table Algorithm

Given a training set of labelled instances, an induction algorithm builds a classifier. We describe two variants of decision table classifiers based conceptually on a simple lookup table. The first classifier, called DTMaj (Decision Table Majority) returns the majority of the training set if the decision table cell matching the new instance is empty, i.e., it does not contain any training instances (Liu et al. 2008).

The second classier, called DTLoc (Decision Table Local), is a new option that searches for a decision table entry with fewer matching attributes (larger cells) if the matching cell is empty. This variant therefore returns an answer from the local neighborhood, which we hypothesized, will generalize better for real datasets that tend to be smooth, i.e., small changes in a relevant attribute do not result in changes to the label.

The Functional Definition

A decision table has two components:

- ✓ A schema, which is a list of attributes
- ✓ A body, which is a multi-set of labelled instances.

Each feature contains a value for each of the attributes in the schema and a value for the label. The set of instances with the same values for the schema attributes is called a cell. Given an unlabelled instance, x, the label assigned to the instance by a decision table classier and is computed as follows. Let I be the set of labelled instances in the cell that exactly matches the given instance x, where only the attributes in the schema are required to match and all other attributes are ignored. If $I \neq 0$, return the majority class in I, breaking ties arbitrarily. Otherwise ($I = 0$), the behaviour depends on the type of decision table used:

- ➢ A DTMaj returns the majority class in the decision table.
- ➢ A DTLoc removes attributes from the end of the list in the Schema and tries to match based on fewer attributes until one or more matches are found and their majority label is returned. This increases the cell coverage until training instances match x. Unknown values are treated as distinct values in the matching process. Given a dataset and a list of attributes for the schema, a decision table is a well-defined functionally (Liu et al. 2008).

7 Experimental Analysis and Results

The experimental results are carried out in MATLAB software package 2012b. MIT – BIH arrhythmia dataset has 48 ECG recording each of length 30mins. In this paper, we take 24 ECG records and 24 ECG records are divided into five separate classes, there are N (Normal), LBBB (Left bundle branch blocks), RBBB (Right bundle branch blocks), PVC (Premature ventricular contractions) and PR (paced rhythm). Table 4 shows the records numbers and corresponding decision classes from MIT–BIH arrhythmia database. The five morphological features are extracted using Pan – Tompkins algorithm.

Table 4 Records numbers and decision classes from MIT–BIH arrhythmia

Class	Records Number.
Normal	101, 105,112, 113,114,115,117,121, 122, 202, 205, 230, 234
LBBB	109,
RBBB	124, 212,232
PVC	106, 203, 209, 219, 228
PR	102, 217,

For the experimental analysis, all the patient's records are given to Pan – Tompkins algorithm. Fig. 7, 8, 9 and 10 show the signals after band pass filtering, differentiating, squaring and moving average filter sample ECG signal. Extracted features are applied to proposed Improved Bijective soft set algorithm and the results are compared with BPN, Naïve Bayes, J48 and Decision table approaches. The results obtained using various classification algorithms are validated based on classification accuracy measures. Validation is important for the developments in data mining, and it is especially vital when the area is still at the initial stage of its development. There are many validation methods available (Inbarani et al. 2013). In this paper, we enumerated our classifier performance using the most familiar metrics precision, Recall and F-Measure. These measures are explained as given below:

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{22}$$

$$\text{Recall /Sensitivity} = \frac{TP}{(TP+FN)} \tag{23}$$

$$\text{F-Measure} = \frac{(2*Precision*Recall)}{(Precision+Recall)} \tag{24}$$

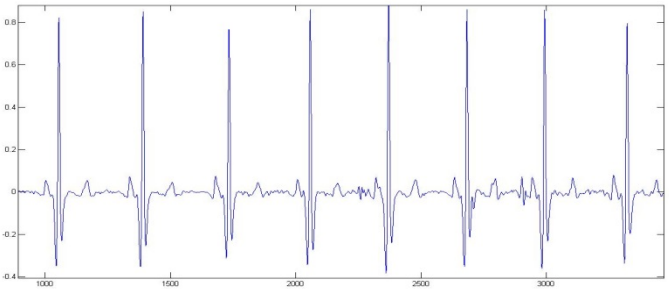

Fig. 7 After band pass filter

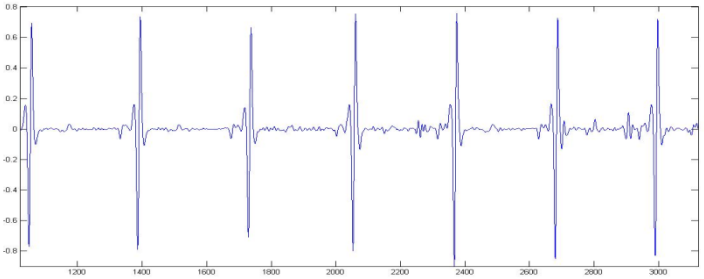

Fig. 8 After band pass filter and differentiating

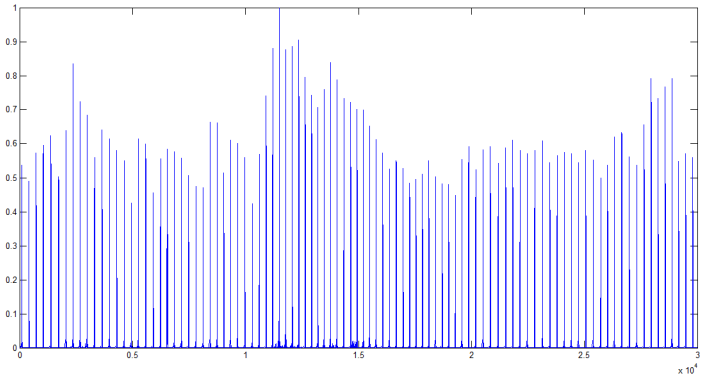

Fig. 9 After band pass filter, differentiating and squaring

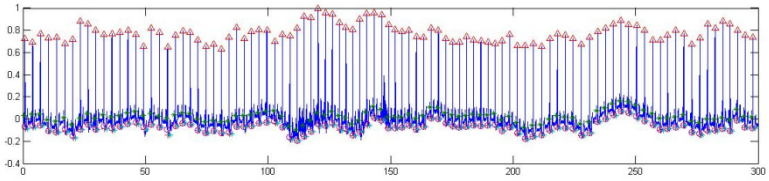

Fig. 10 The final process; after Band pass filtering, Differentiating, Squaring, Moving average filter and PORST detection

Precision is a measure of the accuracy provided that a specific class has been set. Recall is a measure of the ability of a prediction model to select instances of a certain class from a dataset. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score where TP is the number of true positive samples, TN the number of true negative samples, FN the number of false negative samples and FP is the number of false positive samples. the performance of classifiers based on the recognition

of beats, the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are well-defined more suitably as follows:

TP: LBBB class classifies as LBBB. FP: LBBB class classifies as Normal.

FN: LBBB classifies as normal. TN: Normal class classifies as normal.

The most essential metric for defining complete system performance is generally accuracy. We defined the overall accuracy of the classifier for each file as given in:

$$\text{Accuracy} = \frac{\text{Exactly classified samples}}{(\text{Total number of samples})} \tag{25}$$

Classification result is shown in terms of confusion matrix for 2469 Instances of five ECG beats. The confusion matrix demonstrating the classification results of the Improved Bijective soft set is shown in Table 5. According to Table 5, all five classes are classified properly and there is no misclassification in both of the classes using proposed algorithm. The classification performance was achieved with 100% accuracy and it is measured in terms of precision, recall and F-Measure.

Table 5 The confusion matrix of classification methods and five classes

Classification Methods	Output classes	Normal	PB	RBBB	LBBB	PVC	Classification Accuracy %)
IBISOCLASS	Normal	1022	0	0	0	0	**100**
	PB	0	203	0	0	0	**100**
	RBBB	0	0	559	0	0	**100**
	LBBB	0	0	0	124	0	**100**
	PVC	0	0	0	0	561	**100**
BPN	Normal	825	28	99	25	45	80.72
	PB	12	179	7	0	4	88.17
	RBBB	73	5	477	0	4	85.33
	LBBB	15	1	0	106	2	85.48
	PVC	75	0	8	0	478	85.20
Naïve Bayes	Normal	596	17	338	32	39	58.31
	PB	90	70	34	3	6	34.48
	RBBB	178	1	360	0	20	64.4
	LBBB	8	0	6	103	7	83.06
	PVC	211	0	98	1	251	44.74
Decision Table	Normal	904	8	79	0	31	88.45
	PB	36	149	17	0	1	73.39
	RBBB	93	5	447	0	14	80
	LBBB	35	1	0	86	2	69.35
	PVC	75	0	8	0	478	85.20
J48	Normal	1004	1	8	2	7	98.2
	PB	4	193	2	0	4	95
	RBBB	15	1	533	1	9	95.3
	LBBB	4	2	2	116	0	93.5
	PVC	12	0	3	0	546	97.3

Performance of classification algorithms are evaluated using accuracy measures for the 24 ECG records from MIT-BIH arrhythmia datasets. Table 6 represents the performance analysis of the proposed classification algorithm on the 24 ECG records from MIT-BIH arrhythmia datasets. Figure 11 shows the comparative analysis of classification algorithms for the dataset. Figure 11 illustrates that the

classification accuracy of J48 is higher than that of BPN, Decision table and Naïve Bayes. It also shows the effectiveness of Improved Bijective soft set based classification approach over the other classification approaches BPN, J48, Decision table and Naïve Bayes.

Table 6 Performance analysis of the classification algorithms for ECG signal

Accuracy Measures	Classification algorithms				
	IBISOCLSS	**BPN**	**Naïve Bayes**	**Decision table**	**J48**
Precision	100%	85.8%	59.7%	80%	90.2%
Recall	100%	85%	55.2%	78%	90.1%
F-Measure	100%	85.4%	55.2%	77.9%	90.1%

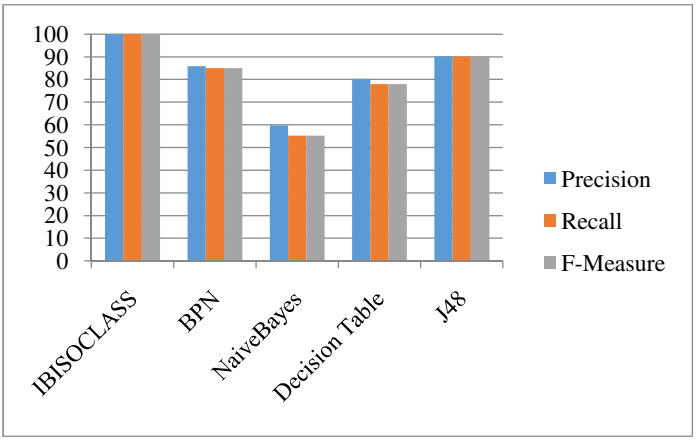

Fig. 11 Comparative analysis of classification algorithms for ECG signal

8 Conclusion

In this paper, a novel bijective soft set based classification method is proposed for ECG signal classification from MIT - BIH data base. The proposed methodology includes three modules: signal acquisition, feature extraction and classifier. In the feature extraction module, we have extracted morphological features as the effective features for differentiating various types of ECG beats. Then, for the classification stage improved bijective soft set is applied and evaluated for ECG beats recognition of five different classes of ECG signals. The acquired result

illustrates that the proposed method performs relatively better than other classification algorithms. Previous study, Hari Mohan Rai et al. achieved 100% accuracy based on two classes (Normal and Abnormal) using neural network classifier but in this study we obtain 100% accuracy based on five different classes. It can therefore be considered as an effective tool for cardiac arrhythmias classification with high accuracy of 100%.These results are very hopeful and inspire us to extend this study to other biomedical as well as non-biomedical applications.

Acknowledgement. The first author immensely acknowledges the partial financial assistance under University Research Fellowship, Periyar University, Salem. The Second author would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-41-650/2012 (SR). And also the authors wish to thank anonymous reviewers for their valuable comments to improve this article.

References

Alejo, R., Toribio, P., Valdovinos, R.M., Pacheco-Sanchez, J.H.: A Modified Back-Propagation Algorithm to Deal with Severe Two-Class Imbalance Problems on Neural Networks. In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Olvera López, J.A., Boyer, K.L. (eds.) MCPR 2012. LNCS, vol. 7329, pp. 265–272. Springer, Heidelberg (2012)

Azar, A.T.: Neuro-fuzzy feature selection approach based on linguistic hedges for medical diagnosis. International Journal of Modelling, Identification and Control (IJMIC) 22(3) (forthcoming, 2014)

Azar, A.T., Banu, P.K.N., Inbarani, H.H.: PSORR - An Unsupervised Feature Selection Technique for Fetal Heart Rate. In: 5th International Conference on Modelling, Identification and Control (ICMIC 2013), Egypt, August 31-September 1-2 (2013)

Azar, A.T., Hassanien, A.E.: Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier. Soft Computing (2014), doi:10.1007/s00500-014-1327-4

Benali, R., Reguig, F.B., Slimane, Z.H.: Automatic Classification of Heartbeats Using Wavelet Neural Network. Journal of Medical System 36(2), 883–892 (2012)

Charfi, F., Kraiem, A.: Comparative Study of ECG Classification Performance Using Decision Tree Algorithms. International Journal of E-Health and Medical Communication 3(4), 102–120 (2012)

De Chazal, P., Celler, B.G., Rei, R.B.: Using wavelet coefficients for the classification of the electrocardiogram. In: Proceedings of the 22nd Annual International Conference of the IEEE, vol. 1(1), pp. 64–67 (2000), http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7218

Dickhaus, H., Heinrich, H.: Classifying bio-signals with wavelet networks-a method for noninvasive diagnosis. IEEE Engineering in Medicine and Biology 15(5), 103–111 (1996)

Dong, T., Shang, W., Zhu, H.: Naïve Bayesian Classifier Based on the Improved Feature Weighting Algorithm. Advanced Research on Computer Science and Information Engineering 152(1), 142–147 (2011)

Gacek, A., Pedrycz, W.: A granular description of ECG signals. IEEE Transaction on Biomedical Engineering 53(10), 1972–1982 (2006)

Giovanni, B., Christian, B., Sergio, F.: Possibilities of using neural networks for ECG classification. Journal of Electrocardiology 29(1), 10–16 (2001)

Gong, K., Xiao, Z., Zhang, X.: The Bijective soft set with its operations. An International Journal on Computers & Mathematics with Applications 60(8), 2270–2278 (2008)

Hari, M.R., Anuragm, T., Shailja, S.: ECG signal processing for abnormalities detection using multi-resolution wavelet transform and Artificial Neural Network classifier. Science Direct 46(9), 3238–3246 (2013)

Hassan, H.H., Paul, K.J., Abraham, T.M.: Classification of Arrhythmia Using Hybrid Networks. Journal of Medical Systems 35(6), 1617–1630 (2011)

Homaeinezhad, M.R., Atyabi, S.A., Tavakkoli, E., Toosi, H.N., Ghaffari, A., Ebrahimpour, R.: ECG arrhythmia recognition via a neuro-SVM–KNN hybrid classifier with virtual QRS image-based geometrical features. An International Journal of Expert Systems with Applications 39(2), 2047–2058 (2012)

Inan, O.T., Giovangrandi, L., Kovacs, G.T.: A Robust Neural-Network-Based Classification of Premature Ventricular Contractions Using Wavelet Transform and Timing Interval Features. IEEE Transactions on Biomedical Engineering 53(12), 2507–2515 (2006)

Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Computer Methods and Programs in Biomedicine 113(1), 175–185 (2014)

Inbarani, H.H., Banu, P.K.N., Azar, A.T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Computing and Applications (2013), doi:10.1007/s00521-014-1552-x

Inbarani, H.H., Jothi, G., Azar, A.T.: Hybrid Tolerance-PSO Based Supervised Feature Selection For Digital Mammogram Images. International Journal of Fuzzy System Applications (IJFSA) 3(4), 15–30 (2013)

Issac Niwas, S., Shantha Selva Kumari, R., Sadasivam, V.: Artificial neural network based automatic cardiac abnormalities classification. In: Proceedings of the 6th International Conference on Computational Intelligence and Multimedia Applications, pp. 41–46 (2005)

Jing, L., Cheng, J., Shi, J., Huang, F.: Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In: Jin, D., Lin, S. (eds.) Advances in CSIE, Vol. 2. AISC, vol. 169, pp. 553–558. Springer, Heidelberg (2012)

Jinkwon, K., Hang, S.S., Kwangsoo, S., Myoungho, L.: Robust algorithm for arrhythmia classification in ECG using extreme learning machine. BioMedical Engineering OnLine (2009)

Karpagachelvi, S., Arthanari, M., Sivakumar, M.: Classification of electrocardiogram signals with support vector machines and extreme learning machine. Neural Computing and Applications 21(6), 1331–1339 (2012)

Lin, C.H., Du, Y.C., Chen, T.: Adaptive wavelet network for multiple cardiac arrhythmias recognition. Expert Systems with Applications 34(4), 2601–2611 (2008)

Liu, H., Feng, B., Wei, J.: An Effective Data Classification Algorithm Based on the Decision Table Grid. In: Seventh IEEE/ACIS International Conference on Computer and Information Science, pp. 306–311 (2008)

Maglaveras, N., Stamkopoulos, T., Diamantaras, K., Pappas, C., Strintzis, M.: ECG pattern recognition and classification using nonlinear transformations and neural networks: a review. International Journal of Medical Informatics 52(1-3), 191–208 (1998)

Mai, V., Khalil, I., Meli, C.: ECG biometric uses multilayer perceptron and radial basis function neural networks. In: Proceedings of the 33rd Annual International Conference of the IEEE EMBS, pp. 2745–2748 (2011)

Marcel, R.R., Jamil, F.S., Philip, J.: Beat Detection and Classification of ECG using self-organizing maps. In: Proceedings of the 19th International Conference of the IEEE EMBS, vol. 1(1), pp. 89–97 (1997)

Mark, R., Moody, G.: MIT–BIH arrhythmia database directory, http://ecg.mit.edu/dbinfo.html

Engin, M.: ECG beat classification using neuro – fuzzy network. Pattern Recognition Letters 25(15), 1715–1722 (2004)

Melgani, F., Bazi, Y.: Classification of Electrocardiogram Signals with Support Vector Machines and Particle Swarm Optimization. IEEE Transactions on Information Technology in Biomedicine 12(5), 667–677 (2008)

Minami, K., Nakajima, H., Toyoshima, T.: Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network. IEEE Transaction on Biomedical Engineering 46(2), 179–185 (1999)

Minghao, P., Yongjun, P., Shon, H.S., Jang-Whan, B., Ryu, K.H.: Evolutional Diagnostic Rules Mining for Heart Disease Classification Using ECG Signal Data. Advances in Control and Communication 137(1), 673–680 (2012)

Mitra, S., Mitra, M., Chaudhuri, B.B.: A Rough-Set-Based Inference Engine for ECG Classification. IEEE Transactions on Instrumentation and Measurement 55(6), 2198–2206 (2006)

Molodtsov: Soft set theory-Rough first results. Computational Mathmetics Application 37(4-5), 19–31 (1999)

Moody, G.B., Mark, R.G.: The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine 20(1), 45–50 (2001)

Nazmy, T.M., El-Messiry, H., Al-Bokhity, B.: Adaptive neuro-fuzzy inference system for classification of ECG signals. In: Proceeding of the 7th International Conference on Informatics and Systems, pp. 1–6 (2010)

Osowski, S., Linh, T.H.: ECG beat recognition using fuzzy hybrid neural network. IEEE Transaction on Biomedical Engineering 48(11), 1265–1271 (2001)

Özbay, Y.: A New Approach to Detection of ECG Arrhythmias: Complex Discrete Wavelet Transform Based Complex Valued Artificial Neural Network. Journal of Medical System 33(6), 435–445 (2009)

Özbay, Y., Ceylan, R., Karlik, B.: A fuzzy clustering neural network architecture for classification of ECG arrhythmias. Computers in Biology and Medicine 36(4), 376–388 (2006)

Pan, J., Tompkins, W.: A real-time QRS detection algorithm. IEEE Transactions Biomedical Engineering 32(3), 230–236 (1985)

Portet, F., Hernández, A.I., Carrault, G.: Evaluation of real-time QRS detection algorithms in variable contexts. Medical and Biological Engineering and Computing 43(3), 379–385 (2005)

Prasad, G.K., Sahambi, J.S.: Classification of ECG arrhythmias using multi-resolution analysis and neural networks. In: Proceedings of the IEEE Conference on Convergent Technologies, vol. 1(1), pp. 227–231 (2003)

Qin, S., Ji, Z., Zhu, H.: The ECG recording and analysis instrumentation based on virtual instrument technology and continuous wavelet transform. In: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 4(1), pp. 3176–3179 (2003)

Saxena, S.C., Kumar, V., Hamde, S.T.: Feature extraction from ECG signals using wavelet transforms for disease diagnostics. International Journal of System and Science 33(13), 1073–1085 (2002)

Senthilkumar, S., Inbarani, H.H., Udhayakumar, S.: Modified Soft Rough set for Multiclass Classification. In: Krishnan, G.S.S., Anitha, R., Lekshmi, R.S., Senthil Kumar, M., Bonato, A., Graña, M. (eds.) Computational Intelligence, Cyber Security and Computational Models. AISC, vol. 246, pp. 379–384. Springer, Heidelberg (2014)

Udhayakumar, S., Inbarani, H.H., Senthilkumar, S.: Improved Bijective-Soft-Set-Based Classification for Gene Expression Data. In: Krishnan, G.S.S., Anitha, R., Lekshmi, R.S., Senthil Kumar, M., Bonato, A., Graña, M. (eds.) Computational Intelligence, Cyber Security and Computational Models. AISC, vol. 246, pp. 127–132. Springer, Heidelberg (2014)

Udhayakumar, S., Inbarani, H.H., Senthilkumar, S.: Bijective soft set based classification of Medical data. In: International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), pp. 517–521 (2013)

Wen, C., Lin, T.C., Chang, K.C., Huang, C.H.: Classification of ECG complexes using self-organizing CMAC. Measurement 42(3), 399–407 (2009)

Wieben, O., Afonso, V.X., Tompkins, W.J.: Classification of premature ventricular complexes using filter bank features, Introduction of decision trees and a fuzzy rule-based system. Medical & Biological Engineering & Computing 37(5), 560–565 (1999)

Yu, S.N., Chou, K.T.: Integration of independent component analysis and neural networks for ECG beat classification. Expert Systems with Applications 34(4), 2814–2846 (2008)

Semantic Geographic Space: From Big Data to Ecosystems of Data

Salvatore F. Pileggi and Robert Amor

Abstract. Enhancing the physical view of a geographic space through the integration of semantic models enables a novel extended logic context for geographic data infrastructures that are modelled as an ecosystem of data in which semantic properties and relations are defined with the concepts composing the model. The signigicant capabilities of current semantic technology allow the implementation of rich data models according to an ontological approach that assures competitive interoperable solutions in the context of environments for general purpose (e.g. the Semantic Web) as well as inside more specific systems (e.g. Geographic Information Systems). Extended capabilities in terms of expressivity have strong implications also for data/information processing, especially on a large scale (Big Data). Semantic spaces can play a critical role in those processes contrasting the mostly passive role of models simply reflecting a geographic perspective. This chapter proposes a short overview of a simple model for semantic geographic space and a number of its applications, mostly focusing on the added value provided by the use of semantic spaces in different use cases.

Keywords: Semantic Technologies, Big Data, Geographic Information System, Semantic Interoperability, Semantic Reasoning, Geographic Space Modelling, Crime Map, Semantic Web.

1 Introduction

Nowadays, geographic space is involved in a surprising number of computer applications (figure 1), in a large set of domains that vary from classic fields (such

Salvatore F. Pileggi · Robert Amor
Department of Computer Science, The University of Auckland
e-mail: f.pileggi@auckland.ac.nz, trebor@cs.auckland.ac.nz

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_12

as engineering (Peachavanish et al. 2006) and architecture (Jones et al. 2004)) to relatively new contexts (e.g. mobile services (Reichenbacher 2009)), as well as a number of fields (such as social science (Goodchild et al. 2008)) that are changing their approaches and perspectives due to the enhanced technological environment provided by IT (McAfee 2006) in terms of information size (e.g.Big Data (Katina & Miller 2013)), information complexity (e.g. social networks (Pileggi et al. 2012)) and computing capability (McAfee 2006).

A renewed and constantly changing technological scenario (Brown et al. 2011) is having a strong impact on old (and mostly unsolved) problems (such as the definition of semantics for geographic space (Kuhn 2002) or the interoperability of geographic data (Manso & Wachowicz 2009)). At the same time, it is opening new interesting scenarios, including novel applications (e.g. (Jiang & Yao 2006, Tsou 2004)) as well as a complete re-design of old-ones.

Popularity is, in this case, synonymous with a heterogeneity of the reality to represent target data (Pileggi & Amor 2013) as well as with a pluralism of requirements (Couclelis 1991), constraints and perspectives (Pileggi & Amor 2014) that make a global convergence of solutions hard, if not impossible. Most efforts are currently oriented to follow the general trends of web computing to solve these problems, mainly the interoperability of data structures and the consolidation of data models (Pileggi & Amor 2013). More concretely, there are generic approaches that propose wide scale solutions, such as designing specific web languages on the model of the domain (e.g. Semantic Sensor Web (Pileggi et al. 2010)). Other approaches focus on application oriented models. In both cases, there are evident limitations due to different reasons with a common point of convergence: if last generation technology is providing empowered capabilities, applications are consequently proposing challenging requirements that want to overcome certain barriers.

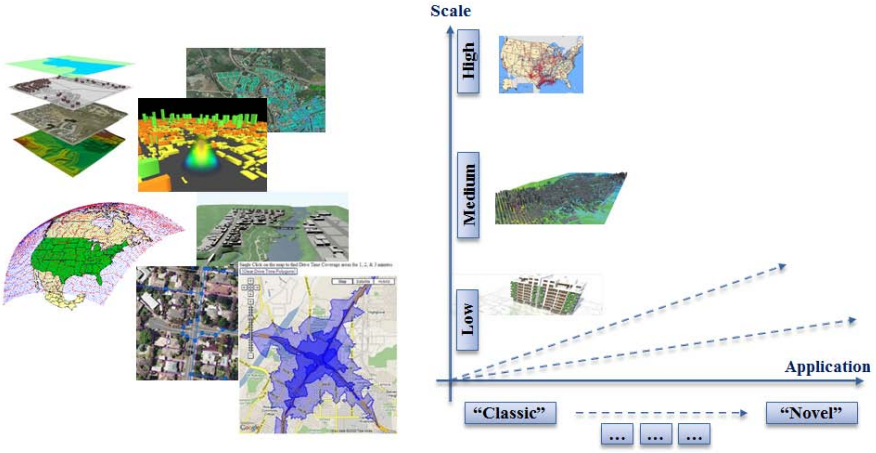

Fig. 1 Example of GIS and a simple classification by Application/Scale

The concept of Semantic Geographic Space (Pileggi & Amor 2014) is, from a theoretical point of view, closer to a generic solution since the common vision of the geographic space is integrated with semantics that, on the one hand, represent specific perspectives of the target space and, on the other hand, provide a contextual understanding of a data set. But it is the specification of these semantics (and consequently of application-specific perspectives for the space and for the related data infrastructure) that makes the data model applicable in practice as a domain-specific representation.

There are different parameters of quality for the assessment of a data model. In this case there is objectively a critical tradeoff between the need for generality and the simultaneously concrete set of requirements in terms of expressivity to accommodate. An ontological approach (Pileggi & Amor 2014) in the context of the current semantic technologies (*Web Ontology Language (OWL)* n.d.) seems to provide a valid frame for designing competitive and scalable solutions through the use of declarative data structures potentially working according to interoperable models. (Bittner et al. 2005).

Apart from the introductory part, including this section and the following section about related work, and from the conclusion section at the end of the document, the chapter is divided into two different logic parts that respectively deal with the description of a model for Semantic Geographic Space and a brief analysis of some application cases.

As previously introduced, modelling semantic spaces implies a critical convergence between generic aspects and domain/application specific concepts (Pileggi, et al. 2013, Pileggi & Amor 2014). An exhaustive analysis across the different domains would be hard to propose. In order to prioritize the understanding of the key concepts and the discussion of their applications in concrete environments, the chapter focuses only on the more generic aspects according to an abstract view of the space that does not include any domain-specific vocabulary. The description of the model is integrated with the overview of an implementation (Pileggi & Amor 2014) over OWL technology (*Web Ontology Language (OWL)* n.d.).

The scope of the second part of the chapter is the critical discussion about the benefits introduced by the model if applied to different scenarios as well as the potentialities of semantic spaces inside complex applications. Unfortunately, an exhaustive analysis would require a survey on a large set of use cases. Furthermore, most advances potentially introduced are not related exclusively to the space model but involve semantic reasoners or other components for information processing. A small number of significant applications have been selected to show specific features in concrete contexts trying to limit to the strictly needed domain specific aspects.

2 Related Work

As addressed in the previous section, a computerised vision of geographical spaces has a direct and strong relation with its application domain. Regardless of philosophical issues and implications, a simplified view (figure 2) assumes the ”real” space related to multiple machine-processable representations, each one reflecting one or more perspectives of the target space. Sometimes the space itself (and its semantics) is the object of interest, other times sets of data are associated to the geographic space through some logic.

This apparently simple scenario encloses an important number of open research issues mostly reflecting concerns and requirements from complex application domains. Due to the constant increase in the information’s size and of the information’s sources, the first and probably more widespread concern is, directly or indirectly, related to interoperability. Most recent applications try to overcome the traditional barriers through approaches based on rich data models (e.g. (N.Anh et al. 2012, M.G.Strintzis et al. 2009)), such as the already cited ontological models (e.g. (X.Mao & Q.Li 2011, Bittner et al. 2005)).

One of the most evident limitations in many semantic environments is currently a clear application-specific focus (e.g. (F.Luckel & P.Woloszyn 2009, C.Shahabi et al. 2010, H.H.Eldien 2009, D.Zheng-yu & W.Quan 2009)) that, often, does not allow an open and flexible approach for the space representation. The representation of Social Objects (Pileggi et al. 2012) is missed, even looking at the latest solutions. This lack seems to be explicitly in contrast with the last trends for data and applications that assume an increasing socialization of the information (Thompson 2013).

Semantic Geographic Spaces (Pileggi & Amor 2014) try to overcome most limitations previously addressed by providing a multi-layer semantic framework, including a generic base layer on the top of which can be designed further specific layers. It is eventually integrated with domain specific elements at different levels. This approach integrates the physical description of the geographic space with a logic-based understanding of it, focusing on semantic relations and properties.

3 Semantic Geographic Space

As previously introduced, Semantic Geographic Space is designed to overcome the current geographic view of the space integrating machine processable semantics into the data infrastructure in a context of enhanced interoperability.

Different approaches can be proposed, prioritizing the accommodation of certain requirements more than others.

The solution proposed in (Pileggi & Amor 2014) is aimed at the definition of a generic base support that can be extended and integrated in the context of concrete application domains in terms of vocabulary and structure (open model). The logic adopted to bring together the different data structures composing the model is

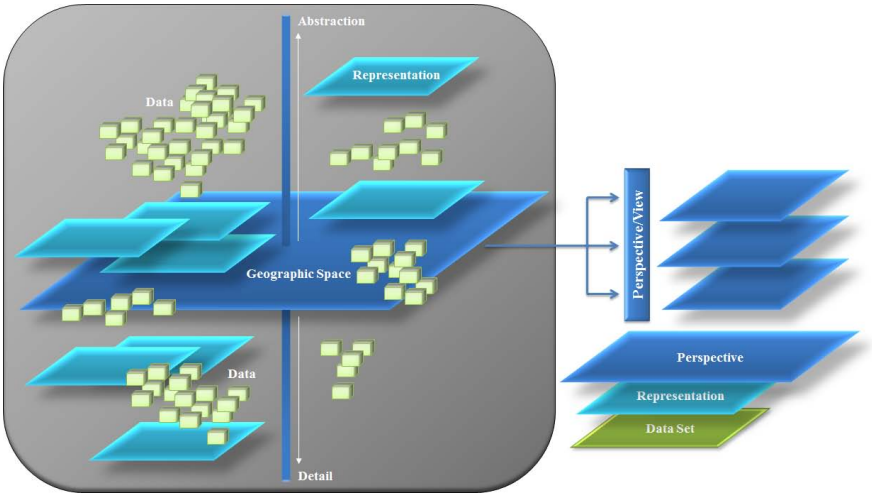

Fig. 2 A simplified view of geographical spaces from a computer system's perspective

explicitly designed in order to assure an applicability on large scale regardless of the level of abstraction or details of the target space to represent.

Even though a complex data infrastructure is usual to propose several clusters of concepts to harmonize, the key and critical asset inside the model is the geographic space model (Pileggi & Amor 2014) for its strong influence on all the other design decisions. It acts as a kind of frame in which any space of interest is defined as a "black-box" (*Container* in (Pileggi & Amor 2014)) that has to be characterized (through Semantic Properties) and related to the other spaces (through Semantic Relations). Figure 3 shows the main concepts composing the geographic space models, as they are implemented according to an ontological approach (Pileggi & Amor 2014).

Exactly the same approach is followed to define the data space (Pileggi & Amor 2014): data are independently represented and they can be characterized though properties and related to other data and to spaces through relations; data containers (*Data Layer* in the model) are built on the same schema, enabling a dynamic grouping by type, by property and by relation.

Some details about semantic relations and properties will be provided in the next two subsections.

3.1 Semantic Relations

Semantic relations are a key factor for defining semantic spaces (Pileggi & Amor 2014) since they establish a characterized relationship among the subjects composing the data model. For instance, relations are understood as a domain specific

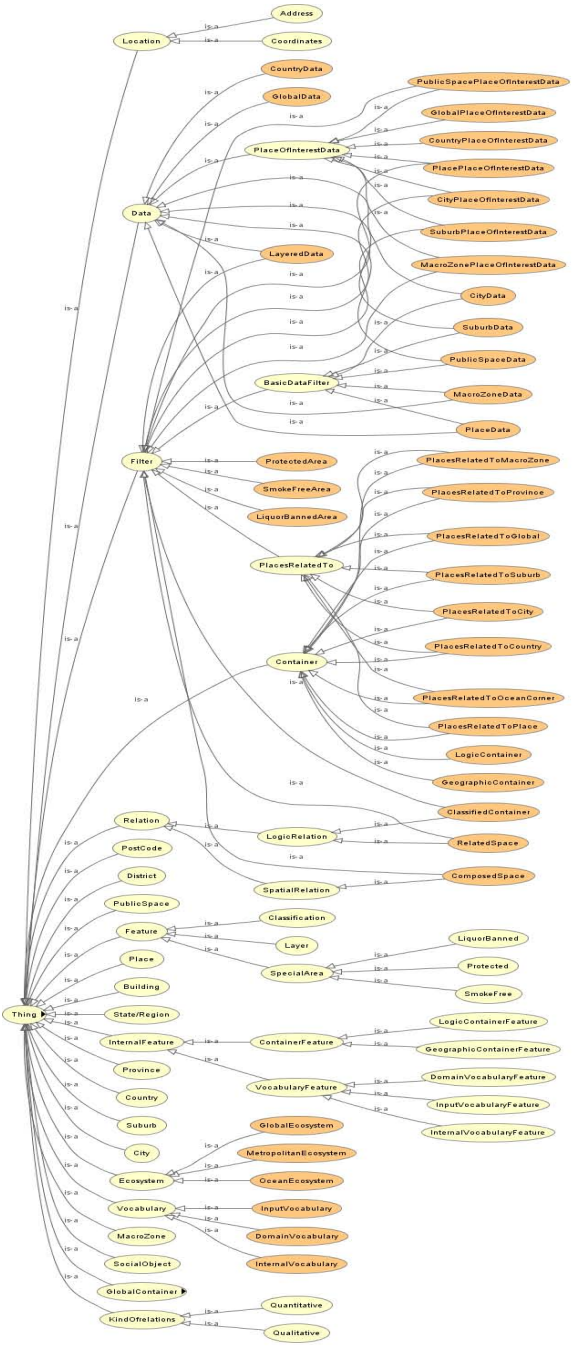

Fig. 3 MANSION's main concepts implemented according to an ontological approach (Pileggi & Amor 2014)

component of the model (Pileggi & Amor 2014) that has to exhaustively express the relationships among spaces, among data sets, between spaces and data sets as well as any other logic connection involving target objects, according to a certain perspective that the model is reflecting.

Due to its application specific focus, a significant overview of a concrete set of semantic relations could result in a domain analysis more than in a technical overview. In (Pileggi & Amor 2014) a small set of relations involving exclusively spaces is briefly discussed. Those relations (figure 4, 5, 6 and 7) are expressive enough to provide a consistent point of view of a space and, at the same time, are generic enough to reflect a domainless perspective.

Apart from the usual relations to express the physical inclusion of a space into another (Pileggi & Amor 2014) and to compose a bigger space by using smaller ones (Pileggi & Amor 2014), this set of relations allows one to distinguish between spaces geographically close (relation *N* in figure 4) and spaces with semantic dependencies (relation *D* in figure 5), as well as with semantic similarities (relation *PA* in figure 7). Assuming a multi-layer logic definition of the space according to a certain parameter (*Logic Level* in figure 6), logic spaces can be defined through the relation *P*. Logic spaces can reflect abstracted views of physical spaces (e.g. district) or can define logic environments without a clear connection with the geographic space: for example an overall "Industrial Area" could be composed of all the industrial areas distributed in a territory, as well as, in the case of a university that has more than one campus, an overall "University" could be composed of the different campuses.

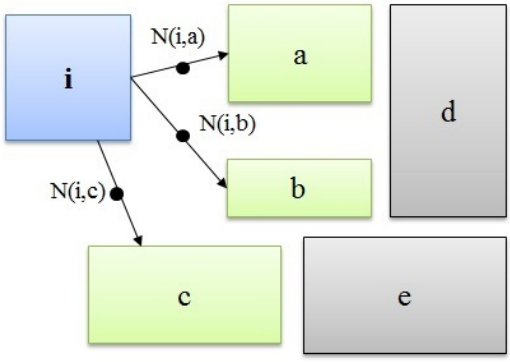

Fig. 4 Examples of generic relations among spaces: Neighbour (N)

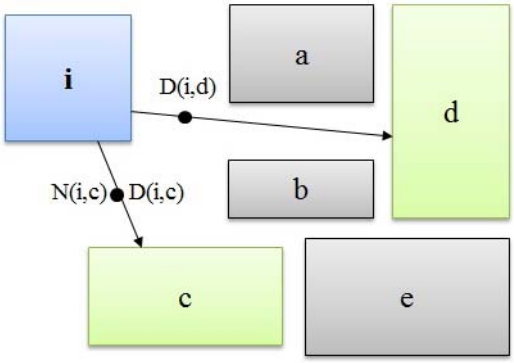

Fig. 5 Examples of generic relations among spaces: Dependent (D)

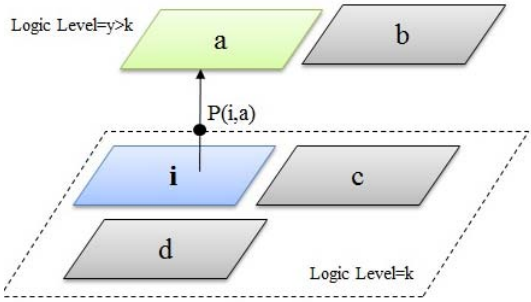

Fig. 6 Examples of generic relations among spaces: Parent (P)

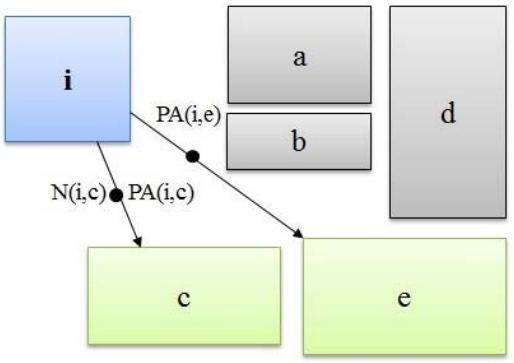

Fig. 7 Examples of generic relations among spaces: Pair (PA)

3.2 Semantic Properties and Profiling

Semantic properties are the natural complement of semantic relations since they enable the characterization of spaces (and of other subjects) according to a certain classification or vocabulary (Pileggi & Amor 2014). They are also used to enhance or specify a semantic relation (Pileggi & Amor 2014).

Extensive sets of properties to classify or characterize spaces (e.g. Residential, Commercial, Beach, Lake, Nightlife, Natural Reserve) or data layers (e.g. Tourism) are easy to figure out for generic views but can be extremely structured in certain domains.

At application level, semantic properties play a key role since they provide a contextual understanding of the associated information. Many advanced solutions apply complex techniques for contextual data processing that need extended profiling. Therefore, semantic profiles (Noulas et al. 2011, Xie et al. 2013) are often integrated and associated to one (or more) semantic properties. Profiles are complex structures that, if associated to the space (*Space Profile*), can provide a strong characterization of the environment. Furthermore, space profiles can be used both with other profiles (such as user profile (Golemati et al. 2007), service profile (Ankolekar et al. 2002), social profile (Kontaxis et al. 2011)) to implement complex semantic matchings (Giunchiglia et al. 2009).

4 Applications

The strong popularity of geographic spaces inside Information Systems and, in general, the increasing demand for geographic-based systems from Information Society address a wide range of potential application domains. In these domains, the geographic space can be involved in a different way that varies from relatively simple representations to enormously complex processes.

The application specific foci and the simultaneous growing diffusion of geographic-oriented services and applications make also an exhaustive generic classification of space models adopted by the different Geographic Information Systems (Worboys & Duckham 2004) hard to cover comprehensively. A commonly accepted criteria (figure 1) consists of an analysis by characteristic inside a concrete application domain. For example, the space models associated to the representation of territories (e.g. city) could be classified according to their scale: a low scale, at the level of a building, matches requirements from engineering/architectural tasks (such as design); an intermediate scale describing a metropolitan area (or a subset of it) is used, among others, for modelling, analyzing and solving complex problems (e.g.transport system optimization); models addressing a large scale are usual to represent extended spaces (e.g.regions) and they normally support applications that need a global understanding of the space (e.g.territorial planning).

Rather than a systematic classification, the most effective approach to evaluate a geographic space model inside its application domain consists of an analysis of the requirements in terms of expressivity (enhanced semantics). The examples of

application proposed in the following subsections are organized according to this metric and are characterized by an increasing complexity in terms of semantics to represent.

4.1 Geographic Data Ecosystems

Geographic Data Ecosystems (Pileggi & Amor 2014) differ from common geographic data infrastructures due to their enhanced semantics (normally implemented by rich data models on semantic technologies such as (*Web Ontology Language (OWL)* n.d.)) that allow the representation of the subjects and the objects composing the model both with the properties and the relations existing amongst them.

The main components of a geographic ecosystem (Pileggi et al. 2013) are the geographic space and the associated space of data.

The most relevant extensions for the space representation were discussed in the first part of the chapter and consist of a richer understanding of geographic spaces including profiling, abstractions and relations amongst the different components. In practice, enhanced semantics overcome the classic concept of physical geographic space integrating it with semantic perspectives.

Also data are represented according to a similar approach that enable abstractions (e.g. complex data layers) as well as relations among data or between data and abstractions. The two main concepts (space and data), respectively representing the objects of interest and their context, are merged in a unique data model through a further set of semantic relations. The normal federation of heterogeneous data is replaced with a consistent representation aimed at interoperability and at a high expressivity in the context of complex applications.

A Geographic Data Ecosystem is an application extremely close to the reference data model and, in fact, it is normally designed directly on the top of it. In a certain meaning, a geographic data ecosystem is the data model itself at some user level and, therefore, it reflects the expressivity of the adopted data model.

Example of Geographic Data Ecosystems according to the MANSION's model (Pileggi & Amor 2014) (previously described in the chapter) are proposed in figure 8 and 9. The uniqueness of this approach to design ecosystems consists in the technique for browsing semantic data and spaces. In fact MANSION proposes a model of the space that reflects the point of view of an observer able to look at the overall space or to "enter" the different subspaces by using the associated vocabulary of concepts or by following existing relations. From both a global or a local perspective, the observer has a semantic view of the space including relations to the other spaces, to data as well as to other abstractions existing in the model. The observer is able to move inside different levels of abstraction as if they are using some kind of semantic zoom on the model.

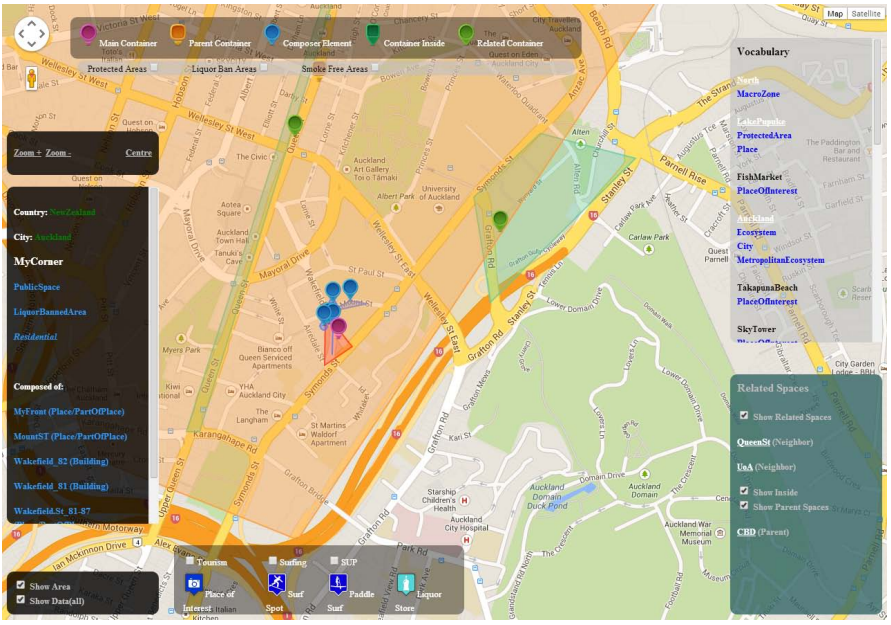

Fig. 8 An example of Geographic Data Ecosystem according to the MANSION's model

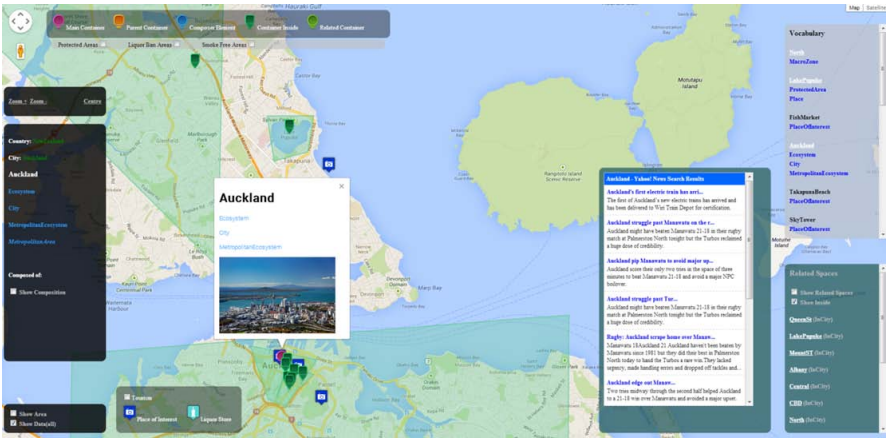

Fig. 9 An example of Geographic Data Ecosystem linked to RSS news feeds

4.2 Crime Mapping: Analysts and Citizens

Crime mapping (Ratcliffe 2010) is normally used by social scientists (Vann & Garson 2001) or analysts (Boba 2001) from law enforcement agencies to map, visualize, analyze and understand crime trends and patterns, as well as their impact on individuals and on communities. An example of professional software (*i2 COPLINK, IBM* n.d.) is showed in figure 10.

The digital era has progressively lead to a computer-based crime mapping, where the amount and the complexity of the information to process play a critical role to identify crime hot spots, along with other trends and patterns, and to design consequent sophisticated strategies for crime prevention, contrasting with those, normally based on statistical models.

The effectiveness of the techniques used by analysts to design and process crime maps depends on several factors. Amongst others, it has a clear relation with the context information, including the geographic space. It's reasonable to think that a semantically richer representation of the space can be a determinant in the context of many concrete applications.

Apart from these generic considerations, taking advantage of the enormous and quick diffusion of web services (including Geographic Information Systems that are often delivered as web services (Fu & Sun 2010)), the crime maps are becoming popular also as generic information resources for society, exactly like other kinds of information (e.g.transport). In practice, an increasing number of people without any professional relation to crime analysts is having access to the crime information, as well as to crime statistics and spatio-temporal distributions. It's pretty evident that crime maps are no longer an exclusive prerogative of the police (and indirectly of the governance) but more properly an information resource potentially available to a wider set of stakeholders (e.g.citizens, visitors).

This section focuses mostly (but not exclusively) on the perspective of the citizen to analyze the possible contributions of semantic geographic space inside current GIS for crime mapping, even though, as previously discussed, also the contribution inside professional processes could be relevant. Furthermore, the analysis is mostly referring to the tools effectively available online at today, more than to research prototypes.

The simplest way to provide a short overview of the public crime maps currently available on the web is a simple classification by level of detail of the information. An example of a low detailed map (figure 11) is provided by the UK Police (*Crime Maps, UK Police* n.d.). As showed in the figure, the information is very abstracted and focuses mostly on trends and statistics. This kind of map is clearly an official channel to show a big picture of police activities to contrast crime more than a proper support to citizens/visitors.

Switching to more detailed maps (e.g. the map shown in figure 12 from USA (*CrimeMapping* n.d.)) the information available is considerably richer, including details about the kind of event and an accurate spatio-temporal distribution. In some maps, extended and extremely detailed information about crime events is provided,

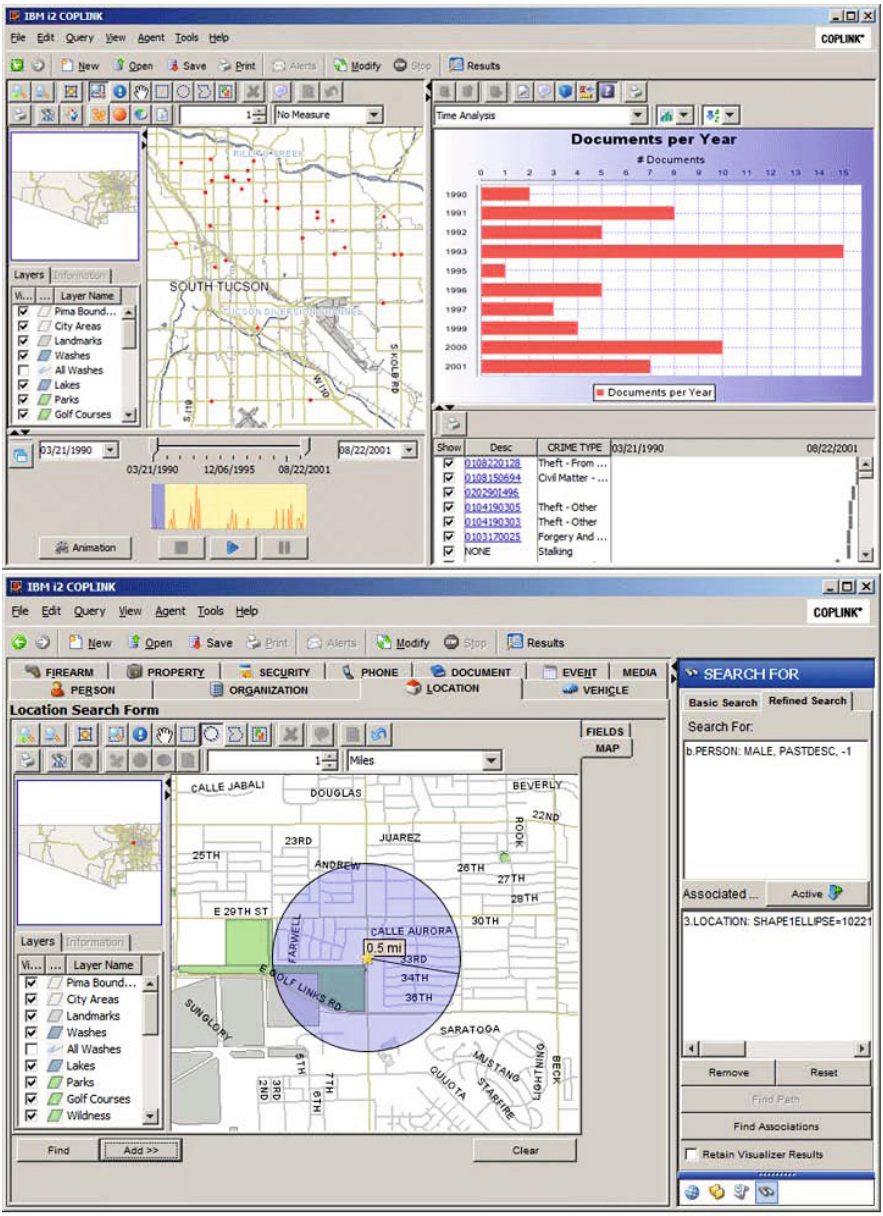

Fig. 10 An example of professional software for crime analysis involving GIS used by Police (*i2 COPLINK, IBM* n.d.)

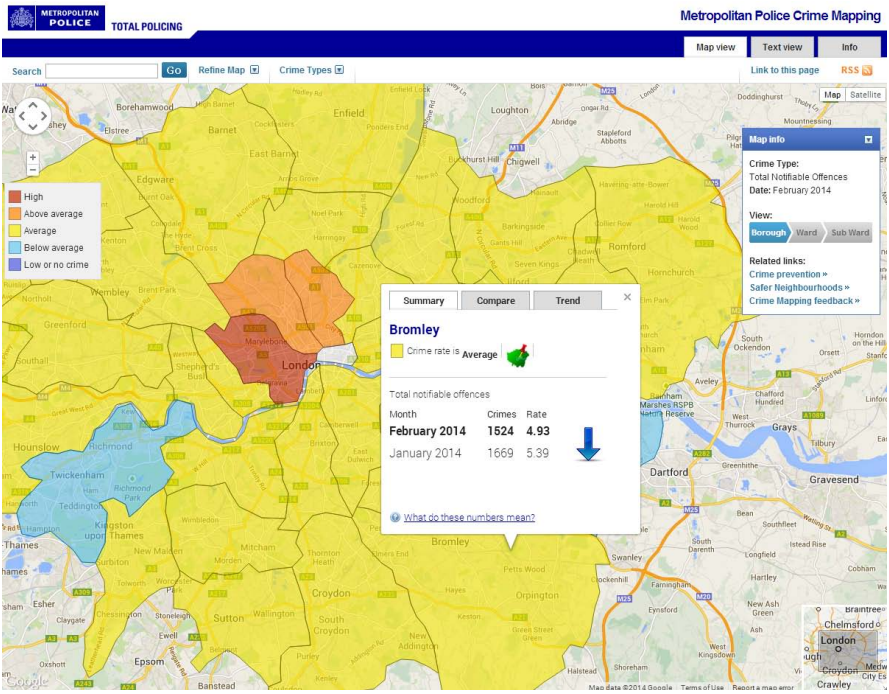

Fig. 11 An example of low detailed crime map from (*Crime Maps, UK Police* n.d.)

such as in the map showed in figure 13 from Australia (*CrimeMAP* n.d.) where also the ”state” (solved/unsolved) of the event is included.

This detailed information should support the community, as it is explicitly affirmed by many sources such as (*CrimeMapping* n.d.) that aims its service as in the follow:

”This portal is dedicated to helping law enforcement agencies provide the public with valuable information about crime activity by neighborhood. Our goal is to assist police departments in reducing crime through a better informed citizenry.”

Even considering the importance, the detail and the accuracy of the information provided, there are actually several concerns about the real usefulness to the community and about practical benefits to citizens in their everyday life. If the goal is to make aware about the crime risks in certain zones, then there are serious doubts about the need to provide a so sensitive and critical information in a context of detailed geo-localization: aiming at consciousness could realistically result in this case in a feeling of fear and a general tendency to consider the police (and indirectly the governance) as ”guilty of not doing enough”. These concerns appear strongly reinforced if the (potential) benefits are compared to collateral effects on a large scale, including, among the others, privacy issues and damages to local business. Also applications designed on top of this information are likely proposing the same

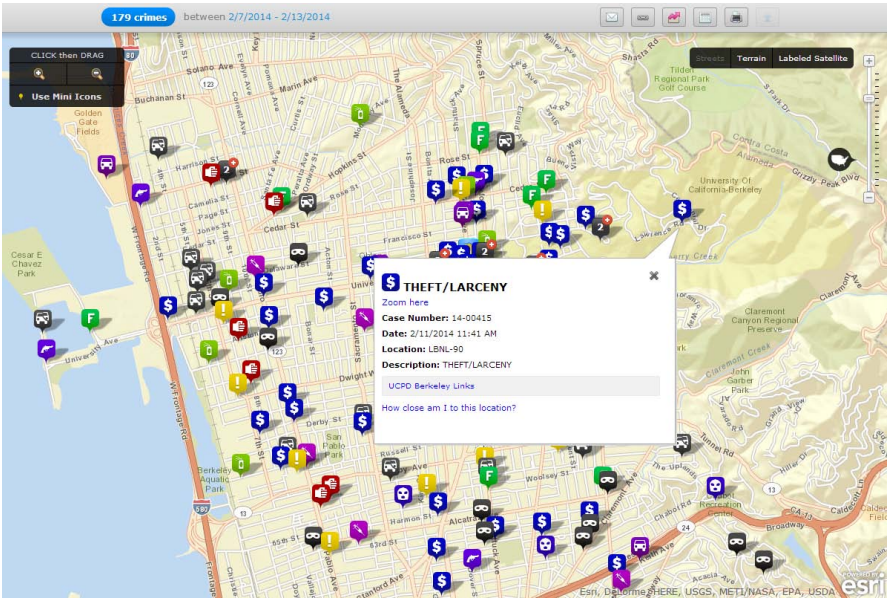

Fig. 12 An example of detailed crime map from (*CrimeMapping* n.d.)

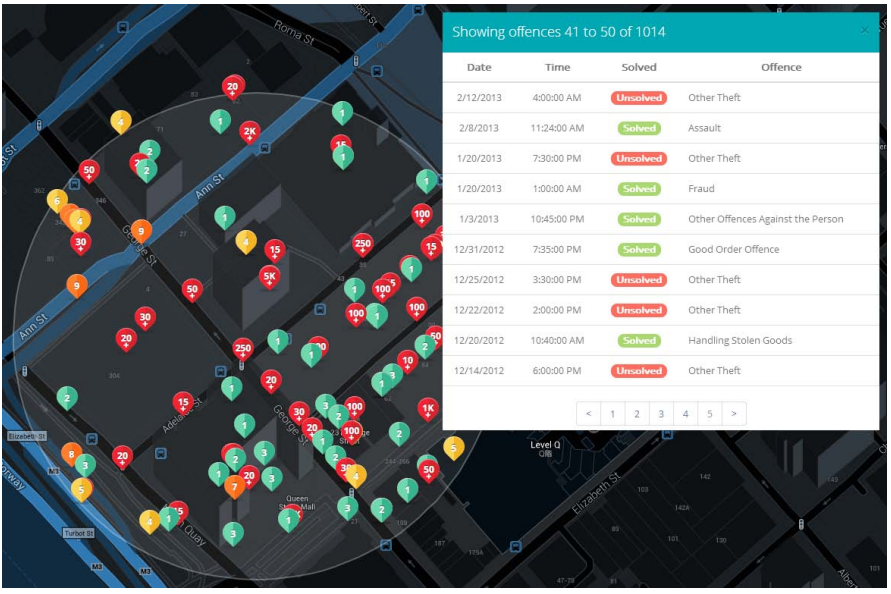

Fig. 13 Another example of detailed crime map including status information (from (*CrimeMAP* n.d.))

limitations. It's evident that approaches designed for analysts cannot be the same for the community.

Crime mapping and analysis is a current object of research in the context of several initiatives and is quickly evolving. Improved techniques and approaches for modeling and analysis of complex patterns and trends both with a more capable technologic support are resulting in a refreshed scenario that should produce tangible advances in a short time. For example, collaborative approaches (Furtado et al. 2012) seem to reflect the progressive socialization of the information on a large scale through active roles of multiple stakeholders.

The application of semantic profiling (as introduced in the first part of the chapter) is proposed as an example of the practical impact that the enriched understanding of the space context can have on crime mapping. Regardless of the level of detail of the provided information, from a non professional point of view (citizen/visitor) the main limitation of the current crime maps is the inability to correctly understand (and consequently interpret) the meaning of the information.

The semantic profile (*Crime Profile* (figure 14), in this case) associated to the different spaces can provide an abstracted and generic overview of the potential dangers that could characterize a certain kind of place (e.g. park, parking, nightlife district, red-light district, tourism attraction, stadium) without considering a specific place. This part of the profile defines, in practice, a *Class of Danger* composed of different *Risk Category* (e.g. assault, robbery) associated to a place with certain characteristics. This part of the profile has no relation with any concrete space but it is designed according to an abstracted view of a certain kind of place from a crime perspective.

The direct and natural integration for a class of danger is the evaluation or quantification (*Risk Assessment*) of each risk in that concrete place from data. The assessment is contextual information that can be modelled according to different schemas and strategies but, in order to overcome many limitations previously described, a high-level abstracted view of data (e.g. low, medium, high) that compares with other places of similar characteristics is advised. Summarized views of the danger for a place are welcome if they are explicitly associated to the context (risks in the profile).

In order to provide more consistent support, the profile has to include also a map of *Recommendations* associated to the different risks. Recommendations are structured in a similar way to risks: generic recommendations associated to generic risks are associated to specific recommendations for a concrete place as a function of the overall information available (including the context).

The crime profile is a relatively simple technique that authorities could easily manage in practical cases. It could contribute to facilitate the understanding of risks for citizens trying to provide an abstracted picture of the crime status in an attempt to not damage the image of the places. Furthermore, it can be associated with other profiles to provide an integrated support to the community. For example, a health profile could be aimed at making people aware about potential risks (e.g. pollution)

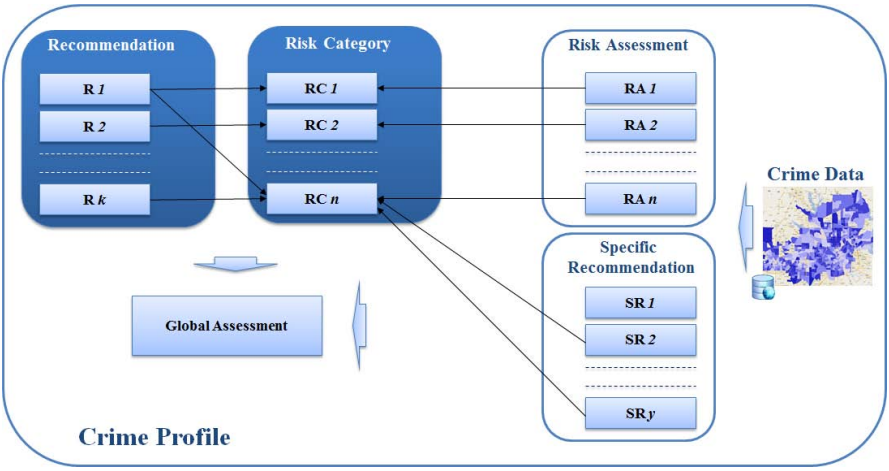

Fig. 14 An example of crime profile focusing on the citizen/visitor's perspective

on their healthy in a certain kind of place, the current status of the risk as well as the recommendations associated to that risk and that status.

4.3 Reasoning on Geographic Spaces

A metropolitan area is a complex scenario, where people live, act and perform their everyday life, as individuals and as a part of a community. The processing of the information in its geographic context applying different techniques (such as (Frank 1992) and (Nobbir & Miller 2007)) is a common practice in many application domains (e.g. (Liben-Nowella et al. 2005) and (K.Jha & Schonfeld 2000)).

The most significant novelty that is involving the processing of geographic data is the constantly increasing scale of the information (Katina & Miller 2013) as well as its complexity (e.g. Social Data (Wellman 2001, Pileggi et al. 2012)). A higher scale implies further and more consistent needs for interoperability (Winters et al. 2006) as well as for semantics (Fagin et al. 2002) and it is progressively leading researchers towards richer data models (Ontology (*Web Ontology Language (OWL)* n.d.)), including declarative structures, and towards more sophisticated computing techniques (Semantic Reasoning (Guo 2008)).

Big Data have issued to overcome in terms of size and sometimes in complexity (Pileggi et al. 2012) compared to the common understanding of a database or the normal data set, as well as the data often changes or moves too fast to be processed according to conventional techniques (Mayer-Schönberger & Cukier 2013). For example, techniques of stream computing (Yamagiwa & Sousa 2007) are applied to process streaming of data. In general, Big Data introduce a number of advantages

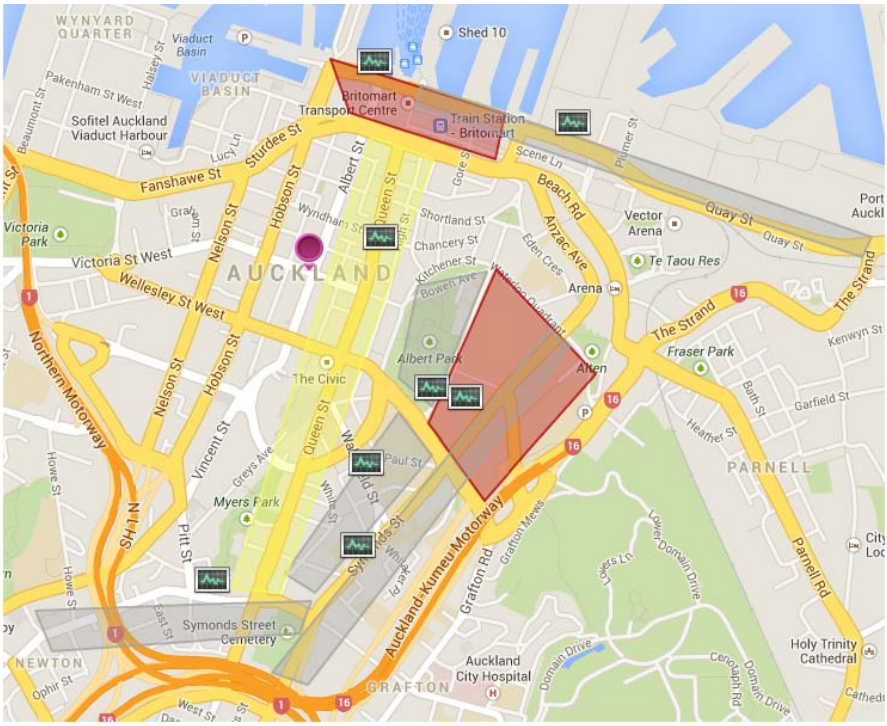

Fig. 15 An example of abductive reasoning on geographic space: scenario from direct data processing

(McAfee & Brynjolfsson 2012) mostly consisting in a great availability of heterogeneous data and data sources that allow one to model and solve problems at a large scale often according to statistical models (Chen et al. 2012).

The strong impact that Big Data are having on IS (Chen et al. 2012) does not mean that there is an "answer for all". In fact, Big Data is not necessarily a better data (Boyd & Crawford 2012) and there actually are several concerns about the reliability of so large and distributed set of data (Wei Tan & Dustdar 2013, Meeker & Hong 2014). Also assuming trustable set of (big) data, (big) errors are often detected, as recent relevant studies demonstrate (David Lazer & Vespignani 2014). Evidently, traps on Big Data can have different causes but most experts agree that, apart from the reliability of data sources, a lack of context (or a bad context (Henricksen et al. 2002)) is one of the most critical factors, strongly affecting many systems and processes. Probably, these considerations advise an integration of novel techniques for data processing with more conventional solutions, a well defined context, as well as a clear understanding of the reliability of involved sources.

This is exactly the role that a semantically enriched specification of the space wants to have in the geographic data processing: from one side, a consistent context is provided as a logic frame and, on the other hand, techniques based on semantic

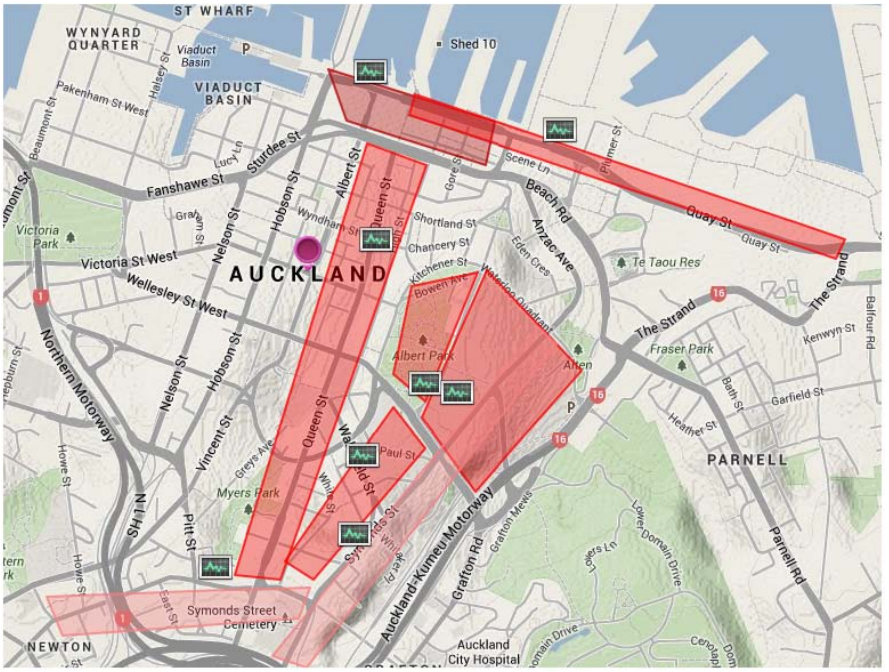

Fig. 16 An example of abductive reasoning on geographic space: distributions by processing data according to an exclusively geographic view of the space

reasoning can be integrated to the Big Data analysis in order to assure more reliable results.

Different problems require strongly different approaches to design semantic support and related reasoners. The ”normal” work mode for current web semantic technology (*Web Ontology Language (OWL)* n.d.) is oriented to deduction (Barbieri et al. 2010), since semantically higher concepts are normally inferred from semantically lower ones through a set of rules provided by model languages. An abdutive behavior of the reasoner (Paul 1993) implements a logical inference that goes from observations to hypothesis to account for the reliable data (observation) and seeks to explain relevant evidence. Other solutions are aimed at building the semantic space from a specific data set and from the physical space, supporting a reasoning model more oriented to reflect an inductive logic (Barbieri et al. 2010).

As an example of an application, a reasoner aimed at the modelling of metropolitan activities is considered. A number of public spaces are defined (figure 15, 16 and 17) at a low level of detail. They belong to the heart of the city of Auckland (New Zealand). The activity for each space is considered to be proportional to the popularity of the place inside several social networks. A significant number of sources is considered in this experiment.

In figure 15 is represented the distribution of the activities according to the direct processing of the information. Spaces in red represent the most significant activities

detected, with darker tone of red reflecting the higher values. Orange/yellow tones are used to mark average or poor levels of activity, as well as gray tones shows areas with a very poor or undetected activity. As showed in the figure, many (very central!) spaces cannot be characterized well as there are several concerns about the reliability of these distributions due mostly to the fact that the main street of the city characterized by a strong and constant activity is actually associated to a low value. There are several reasons for those errors, such as the ambiguity of social content, the need for more detailed spaces and unreliable sources.

The output of a simple reasoner that takes in input from results of the direct data processing and distribute the activity from each space to the close ones is showed in figure 16. This is objectively a better approximation of reality but, as is evident in the figure, only minor differences among spaces can be appreciated. That is because the reasoner assumed a linear decrease of the activity according to a homogeneous view of the space.

Processing the information in an extended and semantically richer context (the semantic space in this case) can contribute to overcome these limitations. The geographic representation of the space is integrated with a set of semantic relations and properties that represent the dependencies among spaces and the definition of abstract districts (e.g. commercial, education, residential) according to the perspec-

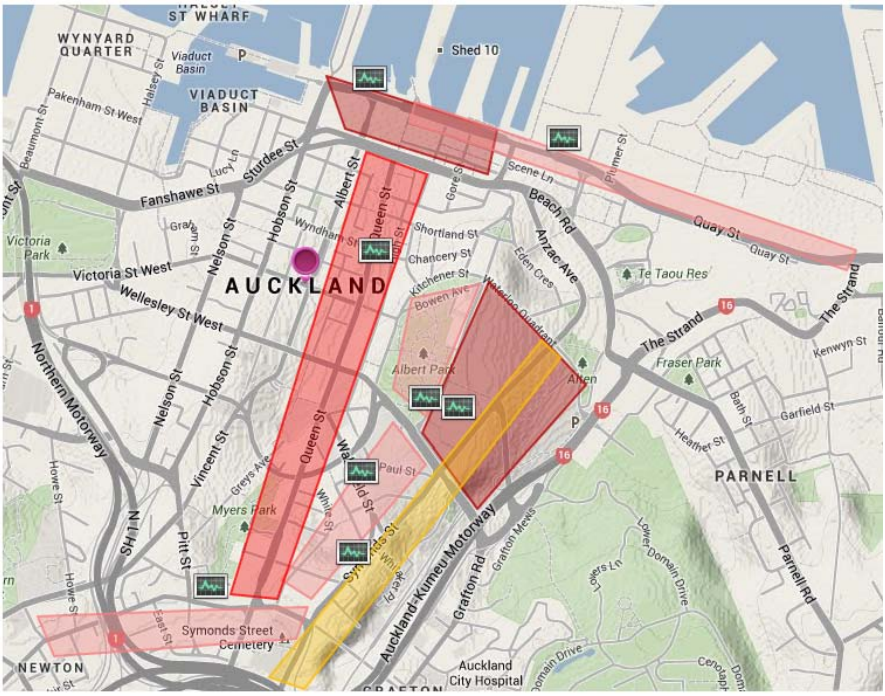

Fig. 17 An example of abductive reasoning on geographic space: reasoning on a semantically enhanced space

tive of a generic observer (e.g. a normal citizen). As showed in figure 17, in the new distribution differences are larger and they do not reflect a linear decrease of the activity as a function of the distance: the reasoner has provided a distribution of the activity coherent with both the input and the context.

5 Conclusions

As discussed in the chapter, a semantically rich understanding of geographic space can be applied in a wide range of domains in order to reach specific goals. The different features of the model can contribute in different ways to solve problems involving a complex representation of the space. Above all, the model provides a key added value anytime the information is processed in a geographical context as part of a knowledge-intensive task that needs a specific perspective of the target geographic space.

The effective application of this approach depends on the capability to define a domain-specific view of the physical space in a formal instance of the model (including the reference vocabulary, semantic relations and properties) as well as the rules to process it. This assumption is fully realistic in most cases where domain specialists are involved in the process. This is, for example, the case of social studies aimed at understanding and evaluation of complex dynamics and behaviours.

On the contrary, the model is hard to apply where the view of the space cannot be exhaustively described, such as in business processes involving virtual stakeholders. Depending on the applications' scope, generic views can be effective or could be completely inadequate.

Multiple views of the space are well supported and, in general, very useful to provide a multi-scenario analysis as well as heterogeneous views of the space (a potentially richer input for the reasoner).

References

Ankolekar, A., et al.: DAML-S: Web service description for the semantic web. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 348–363. Springer, Heidelberg (2002)

Barbieri, D., Braga, D., Ceri, S., Della Valle, E., Huang, Y., Tresp, V., Rettinger, A., Wermser, H.: Deductive and inductive stream reasoning for semantic social media analytics. IEEE Intelligent Systems 25(6), 32–41 (2010)

Bittner, T., Donnelly, M., Winter, S.: Ontology and semantic interoperability. In: Large-scale 3D Data Integration, pp. 139–160 (2005)

Boba, R.: Introductory guide to crime analysis and mapping. Community Oriented Policing Services, USA (2001)

Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society 15(5), 662–679 (2012)

Brown, B., Chui, M., Manyika, J.: Are you ready for the era of "big data". McKinsey Quarterly 4, 24–35 (2011)

Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS Quarterly 36(4) (2012)

Couclelis, H.: Requirements for planning-relevant gis: a spatial perspective. Papers in Regional Science 70(1), 9–19 (1991)

CrimeMAP (n.d.), http://www.crimemap.info/ (accessed March 17, 2014)

CrimeMapping (n.d.), http://www.crimemapping.com/ (accessed March 17, 2014)

Crime Maps, UK Police (n.d.), http://maps.met.police.uk/ (accessed March 17, 2014)

Shahabi, C., Banaei-Kashani, F., Khoshgozaran, A., Nocera, L., Xing, S.: Geodec: A framework to visualize and query geospatial data for decision-making. IEEE MultiMedia 17(3), 14–23 (2010)

Lazer, D., Kennedy, R., King, G., Vespignani, A.: The parable of google flu: Traps in big data analysis. Science 343, 1203–1205 (2014)

Zheng-yu, D., Quan, W.: Road network analysis and evaluation of huizhou city based on space syntax. In: IEEE Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009 (2009)

Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data exchange: Semantics and query answering. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) ICDT 2003. LNCS, vol. 2572, pp. 207–224. Springer, Heidelberg (2002)

Luckel, F., Woloszyn, P.: A "perlaborative" environment for sustainable cities design staff in a participative perspective. gis and knowledge database. In: International Conference on Computers and Industrial Engineering, CIE 2009. IEEE (2009)

Frank, A.: Qualitative spatial reasoning about distances and directions in geographic space. Journal of Visual Languages and Computing 3(4), 343–371 (1992)

Fu, P., Sun, J.: Web GIS: principles and applications. Esri Press (2010)

Furtado, V., Caminha, C., Ayres, L., Santos, H.: Open government and citizen participation in law enforcement via crowd mapping. IEEE Intelligent Systems 27(4), 63–69 (2012)

Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic matching. In: Encyclopedia of Database Systems, pp. 2561–2566. Springer (2009)

Golemati, M., Katifori, A., Vassilakis, C., Lepouras, G., Halatsis, C.: Creating an ontology for the user profile: Method and applications. In: Proceedings of the First RCIS Conference, pp. 407–412 (2007)

Goodchild, M.F., Anselin, L., Appelbaum, R.P., Harthorn, B.H.: Toward spatially integrated social science. International Regional Science Review 23(2) (2000)

Guo, W.: Reasoning with semantic web technologies in ubiquitous computing environment. Journal of Software 3(8) (2008)

Henricksen, K., Indulska, J., Rakotonirainy, A.: Modeling context information in pervasive computing systems. In: Mattern, F., Naghshineh, M. (eds.) PERVASIVE 2002. LNCS, vol. 2414, pp. 167–180. Springer, Heidelberg (2002)

Eldien, H.H.: Noise mapping in urban environments: Application at suez city center. In: IEEE International Conference on Computers and Industrial Engineering, CIE 2009 (2009)

i2 COPLINK, IBM, http://maps.met.police.uk/ (accessed March 17, 2014)

Jiang, B., Yao, X.: Location-based services and gis in perspective. Computers, Environment and Urban Systems 30(6), 712–725 (2006)

Jones, C.B., Abdelmoty, A.I., Finch, D., Fu, G., Vaid, S.: The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In: Egenhofer, M., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 125–139. Springer, Heidelberg (2004)

Katina, M., Miller, K.W.: Big data: New opportunities and new challenges. IEEE Computer 46(6), 22–24 (2013)

Jha, M.K., Schonfeld, P.: Integrating genetic algorithms and geographic information system to optimize highway alignments. Transportation Research Record: Journal of the Transportation Research Board 1719(1), 233–240 (2000)

Kontaxis, G., Polakis, I., Ioannidis, S., Markatos, E.: Detecting social network profile cloning. In: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 295–300. IEEE (2011)

Kuhn, W.: Modeling the semantics of geographic categories through conceptual integration. In: Egenhofer, M., Mark, D.M. (eds.) GIScience 2002. LNCS, vol. 2478, pp. 108–118. Springer, Heidelberg (2002)

Liben-Nowella, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences of the United States of America 102(3), 11623–11628 (2005)

Manso, M., Wachowicz, M.: Gis design: A review of current issues in interoperability. Geography Compass 3(3), 1105–1124 (2009)

Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt (2013)

McAfee, A.: Mastering the three worlds of information technology. Harvard Business Review 84(11) (2006)

McAfee, A., Brynjolfsson, E.: Big data: the management revolution. Harvard Business Review 90(10), 60–68 (2012)

Meeker, W.Q., Hong, Y.: Reliability meets big data: Opportunities and challenges. Quality Engineering 26(1), 102–116 (2014)

Strintzis, M.G., Mademlis, A., Kostopoulos, K., Moustakas, K., Tzovaras, D.: A novel 2d urban map search framework based on attributed graph matching. IEEE MultiMedia (2009)

Anh, N., Vinh, P.T., Duy, H.K.: A study on 4d gis spatio-temporal data model. In: IEEE 2012 Fourth International Conference on Knowledge and Systems Engineering (KSE), pp. 34–38 (2012)

Nobbir, A., Miller, H.J.: Time-space transformations of geographic space for exploring, analyzing and visualizing transportation systems. Journal of Transport Geography 15(1), 2–17 (2007)

Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. The Social Mobile Web 11 (2011)

Paul, G.: Approaches to abductive reasoning: an overview. Artificial Intelligence Review 7(2), 109–152 (1993)

Peachavanish, R., Karimi, H.A., Akinci, B., Boukamp, F.: An ontological engineering approach for integrating cad and gis in support of infrastructure management. Advanced Engineering Informatics 20(1) (2006)

Pileggi, S.F., Amor, R.: Addressing semantic geographic information systems. Future Internet 5(4), 585–590 (2013)

Pileggi, S.F., Amor, R.: Mansion-gs: semantics as the n-th dimension for geographic space. In: International Conference on Information Resource Management, Conf-IRM 2014 (2014)

Pileggi, S.F., Calvo-Gallego, J., Amor, R.: Bringing semantic resources together in the cloud: from theory to application. In: Fifth International Conference on Computational Intelligence, Modelling and Simulation, CimSim 2013 (2013)

Pileggi, S.F., Fernandez-Llatas, C., Traver, V.: When the social meets the semantic: Social semantic web or web 2.5. Future Internet 4(3), 852–864 (2012)

Pileggi, S.F., Fernandez-Llatas, C., Traver, V.: Metropolitan Ecosystems among Heterogeneous Cognitive Networks: Issues, Solutions and Challenges. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2011. CCIS, vol. 348, pp. 323–333. Springer, Heidelberg (2013)

Pileggi, S.F., Palau, C.E., Esteve, M.: Building semantic sensor web: Knowledge and interoperability. In: SSW, pp. 15–22 (2010)

Ratcliffe, J.: Crime mapping: spatial and temporal challenges. In: Handbook of Quantitative Criminology, pp. 5–24. Springer (2010)

Reichenbacher, T.: Geographic relevance in mobile services. In: ACM 2nd International Workshop on Location and the Web (2009)

Thompson, J.: Media and modernity: A social theory of the media. John Wiley & Sons (2013)

Tsou, M.-H.: Integrated mobile gis and wireless internet map servers for environmental monitoring and management. Cartography and Geographic Information Science 31(3), 153–165 (2004)

Vann, I., Garson, D.: Crime mapping and its extension to social science analysis. Social Science Computer Review 19(4), 471–479 (2001)

Web Ontology Language (OWL), `http://www-03.ibm.com/software/products/en/coplink/` (accessed March 17, 2014)

Tan, W., Blake, M.B., Saleh, I., Dustdar, S.: Social-network-sourced big data analytics. IEEE Internet Computing 17(5), 62–69 (2013)

Wellman, B.: Computer networks as social networks. Science 293(5537), 2031–2034 (2001)

Winters, L.S., Gorman, M.M., Tolk, A.: Next generation data interoperability: It's all about the metadata. In: IEEE Fall Simulation Interoperability Workshop (2006)

Worboys, M., Duckham, M.: GIS: A computing perspective. CRC Press (2004)

Xie, X., Zhu, Q., Du, Z., Xu, W., Zhang, Y.: A semantics-constrained profiling approach to complex 3d city models. Computers, Environment and Urban Systems 41, 309–317 (2013)

Mao, X., Li, Q.: Ontology-based web spatial decision support system. In: 2011 19th International Conference on Geoinformatics (2011)

Yamagiwa, S., Sousa, L.: Caravela: A novel stream-based distributed computing environment. IEEE Computer 40(5), 70–77 (2007)

Big DNA Methylation Data Analysis and Visualizing in a Common Form of Breast Cancer

Islam Ibrahim Amin, Aboul Ella Hassanien,
Samar K. Kassim, and Hesham A. Hefny

Abstract. DNA methylation is one of epigenetics mechanisms that plays a vital role in cancer research area by controlling gene expression, especially in the research of abnormally hypermethylated tumor suppressor genes or hypomethylaed oncogenes. The role of DNA methylation analysis leads to determine the significant hypermethlated or hypomethylated genes that are candidate to be cancer biomarkers also the visualization of DNA methylation status leads to discover very important relationships between hypermethylated and hypomethylated genes by using mathematical theory modeling called formal concept analysis.

Keywords: Epigenetics, DNA methylation, hypermethylated genes, Hypomethylated genes and Formal concept analysis.

Islam Ibrahim Amin
Institute Of Statistical Studies and Researches
Cairo University, Egypt
e-mail: eng.IslamAmin@gmail.com
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net

Aboul Ella Hassanien
Faculty of Computers and Information, Cairo University, Cairo - Egypt,
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
e-mail: aboitcairo@gmail.com

Samar K. Kassim
Faculty of Medicine, Ain Shams University, Cairo, Egypt

Hesham A. Hefny
Institute Of Statistical Studies and Researches
Cairo University, Egypt

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_13

1 Introduction

Epigenetics refers to heritable changes in gene that does not cause changes to the underlying DNA sequence (e.g. DNA methylation). DNA sequence consists of four bases : adenine (A), cytosine (C), guanine (G) and thymine (T). DNA wrapped around proteins to form chromosomes. Every cell of human body has 25,000 genes that are located on 46 chromosomes (Poethig, 2001). Gene is a part of DNA and is the blueprint to encode proteins. The gene expression process shows how gene can be used to synthesize a protein. Gene expression process contains two phases : transcription and translation. In the transcription process, the gene is copied to produce mRNA. In the translation process the mRNA translates to protein as shown in Figure 1. The gene expression process can be affected by the type of cell and the biological state. Some genes are responsible for regulating cell growth. Oncogenes and tumor suppressors are two broad classes of genes controlling cell cycle. Any change of the nucleotide sequence of DNA called mutation that may cause cancer, therefore identifying cancer related genes has become very interesting area of research (Li et al., 2012). Microarrays are used to monitor the complete set of mRNA in the cell (Karakach et al., 2010) for thousands of genes therefore the analysis of gene expression data leads to identify the cancer related genes. A single microarray experiment can monitor thousands of genes under several conditions (e.g. normal and cancer samples) (Kaytoue et al., 2011). The output of microarray is called gene expression data (GED) which can be represented as a table, columns refer to samples and rows refer to genes.

There are many data mining approaches that can handle gene expression data such as : clustering, bi-clustering and formal concept analysis. The old fashion method is clustering that groups similar gene expression patterns into one cluster (e.g. K-means, hierarchical clustering, and self-organizing maps), but it imposes a gene to belong to only one cluster. Bi-clustering methods were presented to overcome the clustering method limitations. Recently Formal concept analysis (FCA) has been applied to analysis GED (Amin et al., 2013a, 2012, 2013b; Kaytoue-Uberall et al., 2008).

DNA methylation plays an important role in the gene regulation. Methylation occurs in a region located near the the transcription start site of a gene (promoter). Promoter is a region located upstream of a gene that the RNA polymerase enzyme is attached to start transcription of a gene as shown in Figure 1. Methylation does not involve changes in DNA sequences but it occurs at a specific regions in the promoter called CPG sites (CpGs), the 'C' in CpG refers to cytosine, 'G' refers to guanine and 'P' refers to phosphodiester bond between the guanine(G) and the cytosine(c) as shown in Figure 2. Methylation of CpG in the promoter can inhibit the expression of genes. Recently microarrays (e.g. Illumina) used to monitor many thousands methylation level of CpGs simultaneously in one experiment (Lynch et al., 2009). Hypermethylated means that regions have become more methylated that can inhibit transcription but hypomethylated means that the regions

are less methylated that can affect the expression of these genes to be more expressed.

The rest of this chapter is organized as follows: section 2 gives a brief introduction to the biological characteristics of breast cancer tumor subtypes, DNA methylation, FCA and statistical background. Section 3 shows the proposed method. Section 4 shows experiment results. Applying FCA for breast cancer subtypes in section 5. Finally, Section 6 concludes the paper with a brief summary and future directions.

2 Preliminaries

2.1 DNA Methylation

Hypermethylated indicating that regions have become more methylated but hypomethylated indicating that the regions are less methylated. Hypomethylated CPG islands associated with genes can affect the expression of these genes to be more expressed, also hypermethylated CpGs affect the expression of genes to be less expressed. Searching for hypermethylated or hypomethylated become an important area of research. Thanks to the evolution of development high-throughput microarray (e.g. Illumina) can monitor the methylation level (Lynch et al., 2009; Smale and Kadonaga, 2003). DNA methylation is essential biochemical process for normal development in the higher organisms. DNA methylation occurs by adding a methyl group to the 5' position of cytosine pyrimidine ring, cytosine is one of DNA four bases as shown in Figure 2 (Florescu et al., 2012).

Promoters are regulatory regions of genes and are located upstream of the genes to be transcribed towards the 3' region of the anti-sense strand or template strand (Strachan and Read, 1999). They can be about 100-1000

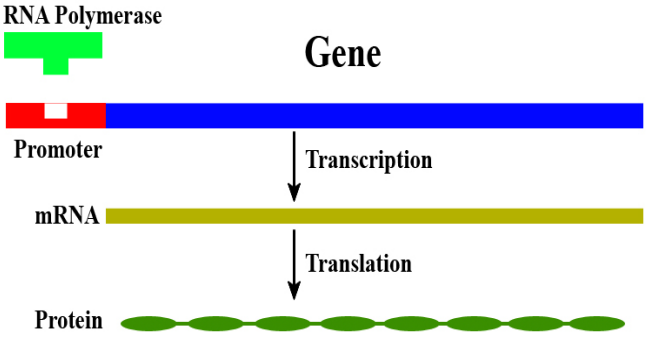

Fig. 1 Gene expression process

base pairs long. Promoters contain specific DNA sequences that provide the binding site for the transcription factors and the RNA polymerase enzyme Transcription factors have specific activator or repressor elements that attach to specific promoters and regulate gene expressions. Promoters are typically positioned upstream of their corresponding genes. They are given negative sequence numbers counting back from -1, for example -30 is a position 30 base pairs upstream (Smale and Kadonaga, 2003). Methylation in promoter can inhibit transcription of gene. Methylation located at specific genomic regions called CPG sites (CPGs). Cancer initiation and progression is stimulated by the accumulation of inherited or acquired DNA mutations. The nature of these alterations may be genetic or epigenetic. Epigenetic modifications are changes in DNA structure that do not involve changes in nucleotide sequence although stably inherited from cell to cell. These include DNA methylation, histone modifications (phosphorylation, acetylation, methylation) and microRNAs (Esteller, 2008). Methylation of cytosine located 5' to a guanosine can occur across the genome, but mostly found within 0.5-4 kb CpG dinucleotide rich regions, known as CpG islands (Esteller and Herman, 2002; Takai and Jones, 2002). Methylation is a method of controlling gene expression widely used throughout the healthy genome. Under normal conditions, the vast majority of CpG sites in the genome are methylated resulting in silencing of their corresponding genes (Jones and Baylin, 2002). The disruption of normal methylation patterns has been found to be an important event in carcinogenesis. Silencing of tumor suppressor genes through promoter hypermethylation is known to be a common event in cancer formation. It provides a selective growth advantage to tumor cells and contributes to the overall genetic instability of the tumor (Widschwendter and Jones, 2002). Hypermethylation appears to be an early event in the process of carcinogenesis, and occurs frequently so that hundreds of genes may be inactivated by DNA methylation in a single cancer (Barat and Ruskin, 2010).

Thanks to the evolution of development high-throughput microarray (e.g. Illumina), the methylation level can be determined (Amin et al., 2013b). The output of illumina experiments represented as a table, rows refer to regions (CPG sites) of genes and columns refer to tissues samples therefore, the relationship between genes and rows is one to many (one gene may have many CPG sites). The methylation level represented as continues values from zero (completely unmethylated) to one (completely methylated), the expression value 'β' reported for each CpG site represents proportion methylated which is in the [0, 1] interval (Bediaga et al., 2010).

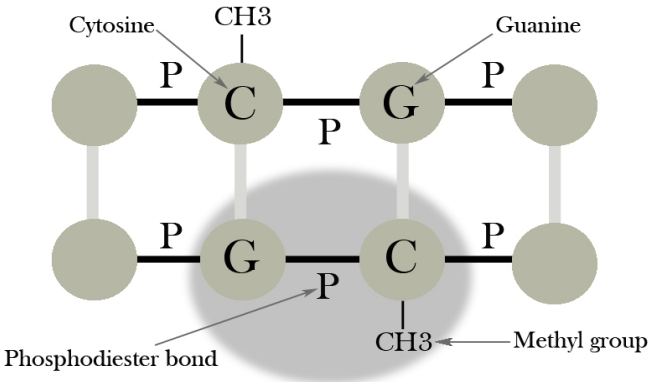

Fig. 2 Methylated CPG sites

2.2 The Biological Characteristics of Breast Cancer Subtypes

Breast cancer could be classified according to receptor status. Receptors are proteins founded on the surface of a cell, in the cytoplasm and nucleus. These receptors have a vital role in receiving chemical signals to keep it from outside into the inside of the cell. Breast cancer receptors such as estrogen receptor (ER), progesterone receptor (PR), and HER2. Breast cancer called estrogen-receptor-positive (ER+) only if it has estrogen receptors. Breast cancer can be called estrogen-receptor-negative (ER-) only if it has not estrogen receptors. Breast cancer which has not any of ER, HER2 and PR are called triple-negative. It is vital to have a test of hormone receptors to determine the most effective and efficient treatment for this cancer. There are four major breast cancer subtypes called basal-like, ERBB2+, luminal B, and luminal A (Bediaga et al., 2010). also the different characteristics of each subtypes in (Castellanos-Garzón et al., 2013), also the summarization of the relationship of breast cancer subtypes and receptors in Table 1.

2.2.1 Luminal A and Luminal B Types

The luminal types are estrogen receptor (ER) positive, usually low grade, grow slowly, and are not aggressive. These types grow from the cells lining the lumen of ducts or glands of the breast as proved by their gene expression patterns. The growth of Luminal A is slower and the prognosis is better than that of Luminal B tumors.

2.2.2 HER2 Type

This tumor type has HER2 gene over-expression and several other genes disorders. They usually are of high pathological grade appearance. They tend to grow more quickly and have a worse prognosis. Hormone therapy and anti-HER2 therapies can be effective against these types of cancers.

2.2.3 Basal Type

These cancer types are of the so-called triple-negative, that is, they show negative estrogen or progesterone receptors and have normal amounts of HER2. Their gene expression patterns are similar to cells in the basal layers of breast ducts and glands. This type is more common among women with BRCA1 gene mutations. These cancers are of high pathological grade that tend to grow quickly and have a poor prognosis. Hormone therapy and anti-HER2 therapies are not effective against these cancers, however, chemotherapy can be helpful (Goldhirsch et al., 2011; Olayioye, 2001; Yanagawa et al., 2012).

The difference between Luminal and Basal like types is that Luminal types express estrogen and progesterone receptors (ER and PR +ve) while Basal types don't.

Table 1 Hormone receptor status for each subtype

Subtype	These tumors tend to be
luminal A	ER+ and/or PR+, HER2-, low Ki67
luminal B	ER+ and/or PR+, HER2+ (or HER2- with high Ki67)
Basal-like	ER-, PR-, HER2-, cytokeratin 5/6 + and/or HER1+
HER2 type	ER-, PR-, HER2+

2.3 Statistical Background

The main task of analysis DNA methylation data is to identify significant genes whose pattern demonstrate a differential change in methylation level under a certain experimental condition. In this chapter DNA methylation data analysis implies two phases : non- specific filter and specific filter to identify significant hypermethylated genes.

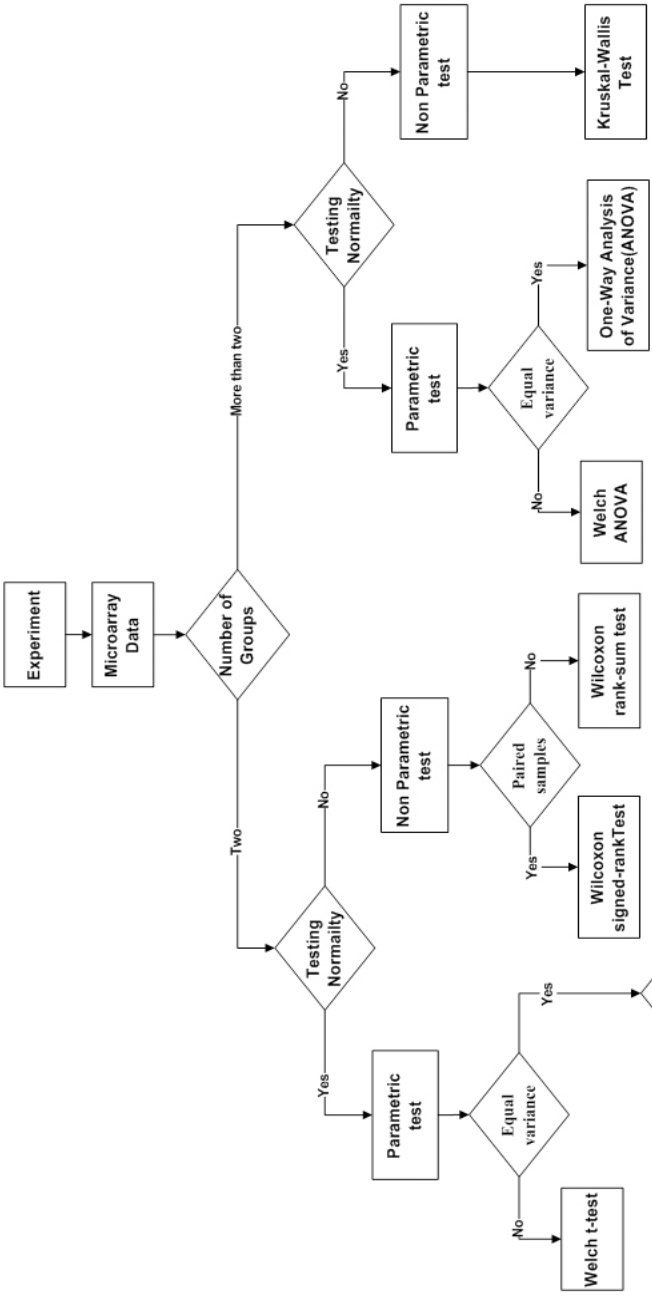

Fig. 3 Statistical Methods for Microarray Data Analysis

2.3.1 Non-specific Filtering

This phase aims to filter out CPGs (rows) which are not candidate to be hypermethylation (get rid of hypomethylated CPGs) by calculating ($\Delta\beta$), the difference between the mean of methylation level for cancer samples with the mean of methylation level for the corresponding adjacent normal tissue, negative values indicate to hypomethylated, whereas positive values indicated to hypermethylation, To select hypermethylation markers that had the lowest difference between cancer and normal tissues samples, a non-specific filter is applied that required $\Delta\beta >$ zero as shown in Figure 4.

2.3.2 Specific Filtering

This phase aims to determine the most differential hypermethylated CPGs using an appropriated statistical test after testing the normality of methylation data by using a one sample Kolmogorov-smirnov test to determine which test will be used a parametric or a non-parametric test. The paired sample t-test is used as a parametric test otherwise a Wilcoxon signed rank test is used as non-parametric test. After determining the significant genes, an additional filter is applied that required a specific minimum difference of (β value) between the two groups (normal and cancer) to reduce the false positives that arise from multiple testing. Figure 3. Shows how to select the most appropriated statistical test for you experiment data according to your experiment design.

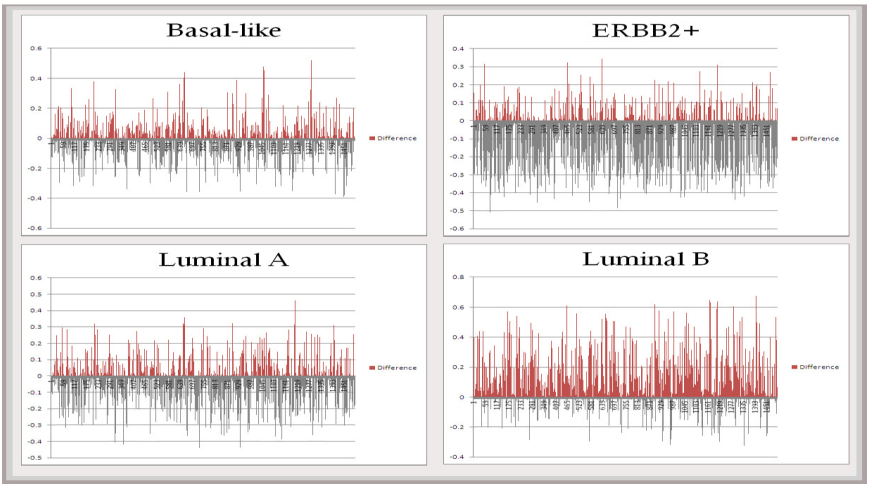

Fig. 4 Histogram of methylation differences between the cancer and normal samples ($\Delta\beta$), negative values refer to hypomethylation and positive values refer to hypermethylation

3 Proposed Method

Illumina methylation microarray can measure the DNA methylation level in 1505 CPG loci sites from the regulatory regions of 806 cancer related genes (one to five CPG sites per gene). In this paper, we analyze the the DNA methylation data from 28 breast cancer subtypes paired samples. The methylation data reported in this paper have been previously deposited in NCBIs Gene Expression Omnibus (GEO) (Omnibus, 2014) and are accessible through GEO Series accession number [GEO: GSE22135]. The experiment representing the four major breast cancer subtypes carried out in 30 paired breast tissues (normal and cancer). for further analysis, two cancer samples has been excluded because of their low methylation levels, finally there are 28 paired breast tissues (normal and cancer) and four samples from peritumoral region. The normal tissue is located at least 2 centimeters away from site of the tumor. The methylation level measured as a continuous values start from zero (completely unmethylated) to one (completely methylated) (Bediaga et al., 2010). The proposed method implies two phases to identify the most significant hypermethylated genes : non-specific filter and specific filter as shown in Figure 5.

3.1 Non-specific Filtering

This step aims to filter out hypomethlated CPGs By calculating ($\Delta\beta$), the difference between the mean of methylation level for cancer samples with the mean of methylation level for the corresponding adjacent normal tissue. Our experiment deals with 1,505 CPG loci. The final result of applying the non-specific filter ($\Delta\beta >$ Zero) to obtain the positive values which indicate to hypermethlated CPGs for each breast cancer subtypes as shown in Table 2.

3.2 Specific Filtering

The output of the non-specific filtering phase is used to be the input to the specific filtering phase. In (Amin et al., 2013a, 2012) assumed distribution normality for as is generally true for microarray data, but this chapter takes into consideration the importance of testing normality to determine which statistical test will be performed. One-sample Kolmogorov Smirnov test is used to test the normality for each breast cancer subtypes, if data follows normality distribution then the paired t-test will be used otherwise Wilcoxon singed rank is the most appropriate test for a paired samples. It is logical to suppose that CPGs with a methylation value great than 0.2 are candidate to be significant hypermethylated CPGs, therefore the additional filter is applied to reduce the false positives after performing multiple testing.

3.3 Formal Concept Analysis

Formal concept analysis was introduced as a mathematical theory modeling by WILLE (1982) (Ganter et al., 1997). Formal concept analysis is very helpful for the analysis of data, also it has been applied in many applications. visualizing the data is one of the useful objective of FCA. The concept lattice provides this visualization. The method of "interactive attribute exploration" is another important feature provided in connection with formal concept analysis, this feature allows knowledge acquisition to answer any precise questions (Burmeister, 2003).

Formal concept analysis provides a powerful framework to identify the groups of objects which sharing the common properties (Arévalo et al., 2005). Discovering hidden dependencies by using FCA as a data analysis techniques based on formal context to build a lattice. This lattice is very important for knowledge representation and knowledge discovery therefore becoming more interesting for biologists (Kaytoue-Uberall et al., 2008). The benefits of FCA in informations sciences is showing in (Priss, 2006).

The standard definition of FCA in (Ganter et al., 1997) which describe the attempt of WILLE to restructure lattice theory. This theory is depending on several theoretical model for conceptual hierarchies. This model deal with a concept as a unit thoughts with two parts: the extension and the intension. The extent contains all objects (entities) that share the given attributes. The intent contains all attributes (properties) shared by the given objects. The form of the formal context is $\mathbf{K}$: = *(G,M,I)* is one of notations of FCA, G is a set of objects and M is a set of attributes. The form of the binary relation between the sets of objects and attributes is $I \subseteq (\ \mathrm{G} \times \mathrm{M})$, also other applications refer to this relation as (g,m) $\in I$ (also known as gIm) which indicate to the object g has the attribute m. The formal context represented as a table which is displaying the relation between objects and attributes. If $\mathbf{K}$:= *(G,M,I)*, A is a subset of the set of objects (e.g. B consist of all objects belong to G) therefore $\acute{A}$ refer to all attributes from the set of attributes (M) which apply to each object in A. B is a subset of the set of attributes (e.g. B consist of all attributes belong to M) therefore $\acute{B}$ refer to all objects from the set of objects (G) to which each attribute from B applies. The pair of (A,B) which is formed the two sets of objects and attributes is called formal context if $\acute{A} = B$ and $\acute{B} = A$. The form of the concept lattice of the context *(G,M,I)* is $\mathbf{B}$ *(G,M,I)* (or $\mathbf{B} = (\mathbf{K})$) only if ($A_1$,$B_2$) and ($A_2$,$B_2$) are concepts of a formal context *(G,M,I)*, A_1, $A_2 \subseteq G$ and B_1,$B_2 \subseteq M$. (A_1,B_1) is a subconcept of (A_2,B_2) only if A_1 is a subset of A_2 and B_2 is a subset of B_1, therefore (A_2,B_2) is called superconcept of (A_1,B_1). This definition explains that how a concept always has a larger intension and a smaller extension than any of its superconcepts (Burmeister, 2003). The following are the standard definitions of FCA.

Definition 1. A formal context is a triple**K**= *(G,M,I)* consist of a set of attributes *M*, a set of objects *G* and the binary relation I between G and M. (i. e. $I \subseteq G \times M$). (g,m) $\in I$ which is read object g has attribute m". Formal concept analysis provides intension and extension, extension is a subset of objects and intension is a subset of attributes for determine a concept, which are defined in the (definition 2).

Definition 2. If a set of A $\subseteq$ G,
$\acute{A}$:= { m $\in$ M | gIm for all g $\in$ A }
$\acute{B}$:= {g $\in$ G | gIm for all m $\in$ B } .

Definition 3. Let *(G,M,I)* is a formal context consists of a pair (A,B) with B $\subseteq$ M called intension, A $\subseteq$ G called extension and $\acute{A}$ = B $\wedge$ $\acute{B}$ = A.

Definition 4. Let (A_1,B_1) and (A_2,B_2) are concepts of a formal context *(G,M,I)* and A_1, $A_2 \subseteq$ G $\wedge$ B_1,$B_2 \subseteq$ M (If $A_1 \subseteq A_2$ (where $A_1 \subseteq A_2$ equivalent to $B_2 \subseteq B_1$), therefore (A_1,B_1) is a subconcept of (A_2,B_2). (A_1,B_1) $\subseteq$ (A_2,B_2) if the (A_2,B_2) is a superconcept of (A_1,B_1). The hierarchical order of the concepts is denoted by $\in$ **B** *(G,M,I)* is called the concept lattice of the context *(G,M,I)* which is the hierarchical order of the set of all concepts of *(G,M,I)*.

4 Experimental Results and Discussion

The non-specific phase has been performed by using the GeneFilter package in R (Gentleman et al., 2011). The output of the first phase reducing CPGs to be 706, 443, 547, 1158 for basal-like, ERBB2+, luminal A and luminal B respectively after that the normality tested with One-sample Kolmogorov Smirnov test applied by using SPSS statistical package which demonstrating that each of breast cancer subtypes methylation values does not follows normal distribution, therefore the most appropriate test is Wilcoxon singed rank (non-parametric test) for the paired samples. The specific filter (Wilcoxon signed rank with P-value ≤ 0.05) applied by using the GeneSelector package in R (Boulesteix and Slawski, 2009). The output of this phase filter out CPGs to be 628, 432, 485 and 1140 for basal-like, ERBB2+, luminal A and luminal B respectively. After applying the additional filter to reducer the false positives. The final results are 39, 19, 29, 71 for basal-like, ERBB2+, luminal A and luminal B respectively which are considered as the most significant hypermethylated CPGs as shown Table 2.

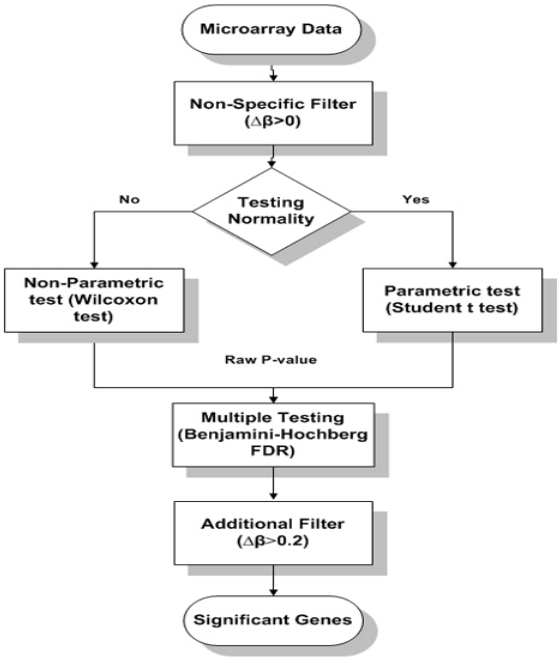

Fig. 5 The proposed model to identify the hypermethylated genes

Table 2 The result of identifying hypermethylated CPGs

Subtypes	$\Delta\beta >$**Zero**	**Wilcoxon singed rank**	$\Delta\beta >$ **0.2**	**Genes**
Basal-like	706 CpGs	628 CpGs	39 CpGs	30
ERBB2+	443 CpGs	432 CpGs	16 CpGs	16
luminal A	547 CpGs	485 CpGs	39 CpGs	30
luminal B	1158 CpGs	1140 CpGs	71 CpGs	50

5 Applying FCA for Breast Cancer Subtypes

This section proposes FCA for mining the hypermethylated genes among breast cancer molecular subtypes. The formal context is first extracted, then we can construct the concept lattice based on a formal context.

Table 3 The formal context

Genes	Basal-like	ERBB2+	luminal A	luminal B
ABCB1	x		x	
ACVR1				x
ADAMTS12			x	
ADCYAP1				x
AGTR1	x	x		
ALOX12	x			x
APC		x	x	
ASCL2			x	
BCR	x			
BDNF	x			
CCNA1				x
CD40	x			x
CD9				x
CDH13	x			x
CFTR	x		x	x
CHGA			x	
COL1A2			x	x
DAB2IP				x
DAPK1			x	
DBC1				x
DLK1	x			x
DLL1				x
EPHA3			x	
EPHB1			x	
ETS1			x	
FABP3				x
FGF2				x
FGF3		x		
FGF9		x		
FRZB	x			
FZD9			x	x
GDF10		x		
GSTM2	x	x		
GSTP1			x	x
HCK		x		
HOXA9	x	x		x
HS3ST2	x		x	x
HTR1B	x		x	x
HTR2A	x			
IGF2AS				x
IGFBP3				x
IGFBP7			x	x
IPF1	x			
ISL1			x	x
JAK3			x	x
MME	x	x	x	x
MMP14	x			
MOS	x	x		x
MYCL2				x
MYOD1			x	
NEFL	x	x	x	x
NPY		x	x	x
PAX6			x	
PDGFRA			x	
PDGFRB	x			x
PENK	x		x	x
PITX2			x	x
PLAT	x			
POMC				x
PTGS1		x		
PTGS2				x
PYCARD	x			
RASSF1			x	x
RBP1			x	
SCGB3A1		x	x	x
SERPINE1	x			
SLC22A3			x	x
SLIT2			x	x
SNCG	x			
SOX1	x		x	x
SOX17	x			x
SPARC				x
ST6GAL1				x
STAT5A				x
TAL1				x
TBX1				x
TERT				x
TMEFF2	x	x		x
TNFRSF10D				x
TNFRSF1B		x		
TSP50	x			
ZNF215			x	x

5.1 Formal Context

A formal context represents the relationship between objects (cancer subtypes) and attributes (hypermethylated genes) which can be easily represented by a cross-table as shown in Table 3.

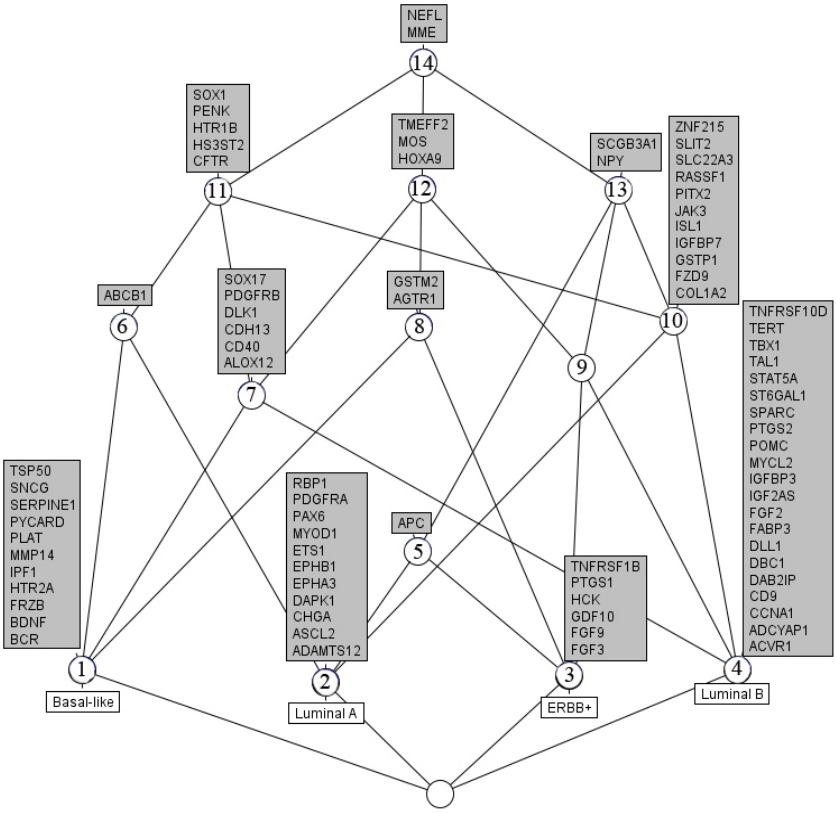

Fig. 6 The concept lattice of hypermethylated genes in breast cancer subtypes

5.2 Formal Concept Lattice

The formal context is used to construct a concept lattice. The concept lattice will construct every relationship between breast cancer subtypes and hypermethylated genes. According to the lattice in Figure 6 we can identify 14 concepts. The concept lattice makes these subtypes hierarchically grouped together according according the their common hypermethylated genes. By using the lattice, concepts(1,2,3 and 4) identify the distinct hypermethylated

Table 4 The distinct common hypermethylated genes among breast cancer subtypes

Concept	**Subtypes**	**Hypermethylated Genes**
C(14)	Luminal A, Luminal B, Basal-like, ERBB+2	MME , NEFL
C(13)	Luminal A, Luminal B, ERBB+2	NPY ,SCGB3A1
C(12)	Basal-like, Luminal B, ERBB+2	MOS, HOXA9, TMEFF2
C(11)	Luminal A, Luminal B, Basal-like,	SOX1, PENK, HTR1B, HS3ST2, CFTR
C(10)	Luminal A, Luminal B	FZD9, GSTP1, IGFBP7, ISL1, JAK3, PITX2, RASSF1 , SLC22A3, SLIT2, ZNF215, COL1A2
C(9)	Luminal B, ERBB+2	NULL
C(8)	Basal-like, ERBB+2	AGTR1, GSTM2
C(7)	Luminal B, Basal-like	ALOX12, CD40, CDH13, DLK1, PDGFRB, SOX17,
C(6)	Luminal A, Basal-like	ABCB1
C(5)	Luminal A, ERBB+2	APC
C(4)	Luminal B	ACVR1, ADCYAP1, CCNA1, CD9, DAB2IP, DBC1, DLL1, FABP3, FGF2, IGF2AS, IGFBP3, MYCL2, POMC, PTGS2, SPARC, ST6GAL1, STAT5A, TAL1, TBX1, TERT, TNFRSF10D
C(3)	ERBB+2	FGF3, FGF9, HCK, PTGS1, TNFRSF1B
C(2)	Luminal A	ADAMTS12, ASCL2, CHGA, DAPK1, EPHA3, EPHB1, ETS1, MYOD1, PAX6, PDGFRA, RBP1,
C(1)	Basal-like	BCR, BDNF, FRZB, HTR2A, IPF1, MMP14, PLAT, PYCARD, SERPINE1, SNCG, TSP50

genes for each breast cancer subtypes. The common hypermethylated genes among breast cancer subtypes can be discovered from lattice as shown in Table 4.

6 Conclusion and Future Work

In this paper, formal concept analysis (FCA) is used as data mining tool for mining the hypermethylated genes among breast cancer subtype. Firstly constructing the formal context which represents the relationships between cancer subtypes and hypermethylated genes then, the concept lattice is constructed based on formal context which let to discover a new relationship between subtype. FCA is a very powerful for identify the relationships between objects therefore these relationships between hypermethylation and hypomethylation in breast cancer subtypes need to be considered in future works.

References

Amin, I.I., Kassim, S.K., Hassanien, A.E., Hefny, H.A.: Applying formal concept analysis for visualizing dna methylation status in breast cancer tumor subtypes. In: 2013 9th International Conference on International Computer Engineering (ICENCO), pp. 37–42. IEEE (2013a)

Amin, I.I., Kassim, S.K., Hassanien, A.E., Hefny, H.A.: Formal concept analysis for mining hypermethylated genes in breast cancer tumor subtypes. In: 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 764–769. IEEE (2012)

Amin, I.I., Kassim, S.K., Hefny, H.A., et al.: Using formal concept analysis for mining hyomethylated genes among breast cancer tumors subtypes. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 521–526. IEEE (2013b)

Arévalo, G., Ducasse, S., Nierstrasz, O.: Lessons learned in applying formal concept analysis to reverse engineering. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 95–112. Springer, Heidelberg (2005)

Barat, A., Ruskin, H.J.: A manually curated novel knowledge management system for genetic and epigenetic molecular determinants of colon cancer. Open Colorectal Cancer Journal 3 (2010)

Bediaga, N.G., Acha-Sagredo, A., Guerra, I., Viguri, A., Albaina, C., Ruiz Diaz, I., Rezola, R., Alberdi, M.J., Dopazo, J., Montaner, D., et al.: Dna methylation epigenotypes in breast cancer molecular subtypes. Breast Cancer Res. 12(5), R77 (2010)

Boulesteix, A.-L., Slawski, M.: Stability and aggregation of ranked gene lists. Briefings in Bioinformatics 10(5), 556–568 (2009)

Burmeister, P.: Formal concept analysis with ConImp: Introduction to the basic features. Fachbereich Mathematik, Technische Universität Darmstadt (2003)

Castellanos-Garzón, J.A., García, C.A., Novais, P., Díaz, F.: A visual analytics framework for cluster analysis of dna microarray data. Expert Systems with

Applications 40(2), 758–774 (2013)
Esteller, M.: Epigenetics in cancer. New England Journal of Medicine 358(11), 1148–1159 (2008)
Esteller, M., Herman, J.G.: Cancer as an epigenetic disease: Dna methylation and chromatin alterations in human tumours. The Journal of Pathology 196(1), 1–7 (2002)
Florescu, A., Cojocaru, D., Fazio, V., Parrella, P.: Study of mir-200c and mir-9 methylation on patients with breast cancer. Analele Stiintifice ale Universitatii" Alexandru Ioan Cuza" din Iasi Sec. II a. Genetica si Biologie Moleculara 13(1), 1–6 (2012)
Ganter, B., Wille, R., Franzke, C.: Formal concept analysis: mathematical foundations. Springer-Verlag New York, Inc. (1997)
Gentleman, R., Carey, V., Huber, W., Hahne, F.: Genefilter: Methods for filtering genes from microarray experiments. R package version, 1(0) (2011)
Goldhirsch, A., Wood, W., Coates, A., Gelber, R., Thürlimann, B., Senn, H.-J., et al.: Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. Annals of Oncology 22(8), 1736–1747 (2011)
Jones, P.A., Baylin, S.B.: The fundamental role of epigenetic events in cancer. Nature Reviews Genetics 3(6), 415–428 (2002)
Karakach, T.K., Flight, R.M., Douglas, S.E., Wentzell, P.D.: An introduction to dna microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems 104(1), 28–52 (2010)
Kaytoue, M., Kuznetsov, S.O., Napoli, A., Duplessis, S.: Mining gene expression data with pattern structures in formal concept analysis. Information Sciences 181(10), 1989–2001 (2011)
Kaytoue-Uberall, M., Duplessis, S., Napoli, A.: Using formal concept analysis for the extraction of groups of co-expressed genes. In: Le Thi, H.A., Bouvry, P., Pham Dinh, T. (eds.) MCO 2008. CCIS, vol. 14, pp. 439–449. Springer, Heidelberg (2008)
Li, B.-Q., Huang, T., Liu, L., Cai, Y.-D., Chou, K.-C.: Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. PloS One 7(4), e33393 (2012)
Lynch, A.G., Dunning, M., Iddawela, M., Barbosa-Morais, N., Ritchie, M.: Considerations for the processing and analysis of goldengate-based two-colour illumina platforms. Statistical Methods in Medical Research 18(5), 437–452 (2009)
Olayioye, M.A.: Update on her-2 as a target for cancer therapy: intracellular signaling pathways of erbb2/her-2 and family members. Breast Cancer Res. 3(6), 385–389 (2001)
Omnibus, G.E.: Gene expression omnibus, geo [online] (2014)
Poethig, R.S.: Life with 25,000 genes. Genome Research 11(3), 313–316 (2001)
Priss, U.: Formal concept analysis in information science. ARIST 40(1), 521–543 (2006)
Smale, S.T., Kadonaga, J.T.: The rna polymerase ii core promoter. Annual Review of Biochemistry 72(1), 449–479 (2003)
Strachan, T., Read, A.: Human molecular genetics, 2nd edn. Bios Scientific (1999)

Takai, D., Jones, P.A.: Comprehensive analysis of cpg islands in human chromosomes 21 and 22. Proceedings of the National Academy of Sciences 99(6), 3740–3745 (2002)

Widschwendter, M., Jones, P.A.: Dna methylation and breast carcinogenesis. Oncogene 21(35), 5462–5482 (2002)

Yanagawa, M., Ikemot, K., Kawauchi, S., Furuya, T., Yamamoto, S., Oka, M., Oga, A., Nagashima, Y., Sasaki, K.: Luminal a and luminal b (her2 negative) subtypes of breast cancer consist of a mixture of tumors with different genotype. BMC Research Notes 5(1), 376 (2012)

Data Quality, Analytics, and Privacy in Big Data

Xiaoni Zhang and Shang Xiang

Abstract. In today's world, companies not only compete on products or services but also on how they can analyze and mine data in order to gain insights for competitive advantages and long term growth. With the exponential growth of data, companies now face unprecedented challenges, however are also presented with numerous opportunities for competitive growth. Advancement in data capturing devices and the existence of multi-generation systems in organizations have increased the number of data sources. Typically, data generated from different devices may not be compatible with each other, which calls for data integration. Although, ETL market offers a wide variety of tools for data integration, it is still common for companies to use SQL to manually produce in-house ETL tools. There are technological and managerial challenges to deal with data integration. During data integration, data quality must be embedded in it.

Big data analytics delivers insights which can be used for effective business decisions. However, some of these insights may invade consumer privacy. With more and more data related to consumer behavior being collected and the advancement in big data analytics, privacy has become an increasing concern. Therefore, it is necessary to address issues related to privacy laws, consumer protections and best practices to safeguard privacy. In this chapter, we will discuss topics related to big data in the area of big data integration, big data quality, big data privacy, and big data analytics.

Keywords: big data, data quality, privacy, data analytics.

1 Introduction

Research conducted by the McKinsey Global Institute (2011) points to big data having the capability to create substantial value and commercial impact.

Xiaoni Zhang
Northern Kentucky University, USA
e-mail: zhangx@nku.edu

Shang Xiang
KPMG
e-mail: sxiang@kpmg.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_14

McKinsey found the potential of a 60 percent increase in retailers' operating margins, 0.7 percent increase in productivity in U.S. health care, all translating into a $300 billion value per year. In addition, there's the potential increase in demand for deep analytical talent positions, estimated between 140,000 and 190,000. Currently, many companies have or plan to implement big data solutions. Businesses changed their view on traditional view of assets to include big data in addition to cash, inventory, and fixed assets (Brands 2014). Not only businesses are exploring big data. Governments with advanced ICT infrastructure have invested in big data for national security, transparency, economic development, and operational efficiency (Gang-Hoon 2014).

In 2002, the Data Warehousing Institute reported the cost of poor data quality on US economy at $600 billion annually (TDWI 2002). Flawed data cost ten times as much to complete a simple work (Everett 2012). Data quality affects business decisions. When evaluating data quality, usually several dimensions are used: accuracy, relevancy, currency and completeness. As more and more big data becomes integrated, the demand for big data analytics increases. Unlike traditional data analysis in which data types are typically numeric, big data is complex in its data types. Data types include numbers, texts, images, voices, videos, etc. Thus big data analysis is performed on the combination of structured and unstructured data. Big data requires the ability to analyze diverse data sources and data types. As new technologies and techniques are emerging in this market, the need to explore the analytic issues and practices associated increases.

Big data is a buzzword and companies from all industries have been investing in big data technologies, techniques, applications and management. In today's world, companies compete not only on products or services but also on how they can analyze and mine data in order to gain insights on competitive advantages and long-term growth. With the exponential growth of data, companies now face unprecedented challenges, conversely are also presented with numerous opportunities. Advancement in data capturing devices and the existence of multi-generation systems in organizations have created many data sources. Although the data generated from different devices may not be compatible with each other, the need for data integration. The ETL market offers a wide variety of tools for data integration; it is still common for companies to use SQL to manually produce in-house ETL tools. There are technological, managerial challenges to deal with data integration. During data integration, data quality and appropriate tools must be used to ensure data quality. Data is the most valuable asset of any organization. Business decisions depend on high quality data. High quality data creates many opportunities for improvement from daily operations to strategic planning. Data quality management is a necessity for businesses and this management needs to be constantly updated, analyzed and perfected. New technologies, techniques, and methodologies on data quality management are constantly developed. Best practices and lessons learned are good sources to improve data quality management. However, data quality issue is not a new issue.

While acknowledging that Big Data has provided benefits to consumers in creating pricing transparency (Fulgoni 2013). Targeting of Internet advertising

based on data analysis is said to offer a means to maintaining or improving brand equity.

In this book chapter we present several interesting topics on big data, as detailed below:

- Section 2 - discuss data/information quality, market trends and data management
- Section 3 - focus on privacy and security issues in general and healthcare in specific
- Section 4 - overview on big data analytics and technologies
- Section 5 - discuss markets and prior publications on big data
- Section 6 - summarize what has been covered in this book chapter and provide future research directions and technology trends

2 Data/Information Quality and Data Integration

Data quality issues are prevalent and such issues attract attention of companies of all sizes from all industries. Data quality is critical for business operations, if there are any errors in business transactions, the consequences could be detrimental – ranging from lost sales, lost customers, lost competitions, failure to market a product. Flawed data could create chain effects affecting many business activities such as the delivery of products to specific locations, traditional marketing outreach to customers and prospects via mailings.

2.1 Definition

Data quality issues have been researched for many years and it has been used interchangeably with information quality. Petter et al. (2013) define information quality as the desirable characteristics of system outputs (content, reports and dashboards) and information quality is one of the key success factors for information systems. Setia et al. (2013) consider that information quality has these dimensions: completeness, accuracy, format and currency whereas Petter et al. (2013) note that information quality contains these dimensions: relevance, understandability, accuracy, conciseness, completeness, understandability, currency, timeliness, and usability.

Data quality is context specific. The data is of high quality if their meanings are well understood by different user groups. With the increasing variations of devices that data is generated from and coupled with different applications, data consistency could represent a problem. For example, customer names and contact information may be stored differently across applications, which in turn a person's age and birth date may conflict within different portions of a database.

Given the data quality dimensions, when managing data quality, it is important to note the quality at seven sources: entry quality, process quality, identification quality, integration quality, usage quality, aging quality, and organizational quality (McKnight 2009). The American SAP User Group reported that 93% of companies experienced data problems in their most recent projects (Woods 2009).

2.2 Overview of Markets

According to ResearchMoz (2013), global data quality tools market will grow at a compounding annual growth rate of 16.78 percent between 2012-2016. Improving productivity is one of the key factors that contribute to this market growth. The Global Data Quality Tools market has been witnessing the emerging SaaS-based data quality tools. Gartner estimates that this market reached $960 million in software revenue at the end of 2012. This translates to growth of 12.3% in constant-dollar terms over 2011 (a standout year in which this market grew by 17.5%). Gartner forecasts that the growth of data quality tools market will accelerate during the next few years and will become the fastest growing market. The market is expected to increase to 16% by 2017 and reach $2 billion in constant-dollar software revenue. The data quality tools in the market of unstructured data are expected to grow significantly in the future.

Recognizing that data quality affects the performance of organizations, companies will focus their attention on data quality. With the developing cloud computing and big data markets, technologies and data quality market will be more complex. New services and vendors emerge, with two major groups. The market has data service providers and data management service providers. Data services providers offer data delivery, analysis, management, or governance-related services, whereas data management services providers operate in the area of finding, collecting, migrating, and integrating data.

With the popularity of Software as a Service (SaaS) and cloud based computing, Gartner (2013) reports that deployments reached 14% for SaaS and 6% for cloud-based tools. In the emerging data management area (big data and analytics), companies are likely to use third parties for its data management, with 69% of respondents reporting they currently use or will use a service provider (Green 2014).

Data quality initiatives for transactional, financial, location and product data are increasing and 78% of data quality projects address customer data quality (Lawson 2013). Overall, data governance drives many data quality initiatives. Data quality in the big data arena has not been taken off yet. Gartner report also shows that the current buyers do not consider data quality related issues. There are stand-alone data quality tools in the market that address the core functional requirements, such as: data profiling, data quality measurement, parsing and standardization, cleansing, matching, monitoring, and enrichment.

Table 1 shows the Gartner report on the magic quadrant for data quality tools between 2009-2013. Gartner classifies the major vendors into four categories: leaders, challengers, visionaries and niche players. In general the leaders control 50% share of the market. In the past five years, the vendors in the leaders category remain constant with such companies like Informatica, IBM, SAS, SAP and

Trillium. In the challenger category, Pitney Bowes is in this category consistently for the past five years, however Oracle joined this category in 2011. In the visionary category, Human Inference was in this category for four years, Talent and Atccama for three years, and Information Builders for the last two years. In the niche player category, five players: Red Point, Data Mentors, Datactics, Innovative Systems, Uniserv, are consistently ranked in this group for the past five years.

In terms of the market share large vendors like IBM, Informatica, Pitney Bowes, SAP and SAS takes 50% and all rest of the market is shared by mid-sized and small-sized companies. As shown in Table 1, in the past five years, the market players are stable with the exception that Oracle enters the data quality market in 2011.

Table 1 Magic Quadrant Data Quality Tools Market between 2009-2013

Year	**2013**	**2012**	**2011**	**2010**	**2009**
Leaders	Informatica	Informatica	Informatica	Informatica	Informatica
	IBM	IBM	IBM	IBM	IBM
	Triumlium	Triumlium	Triumlium	Triumlium	Triumlium
	SAP	SAP	SAP	SAP/ Business Objects	SAP
	SAS	SAS/ Dataflux	SAS/ Dataflux	Dataflux	Dataflux
Challengers	Oracle	Oracle	Oracle		
	Pitney Bowes	Pitney Bowes	Pitney Bowes	Pitney Bowes	Pitney Bowes
Visionaries	Neopost/ Human Inference	Human Inference		Human Inference	Human Inference
	Talend	Talend	Talend		
	Information Builders	Information Builders/ iWay		Datanomic	Datanomic
	Atccama	Atccama	Atccama		
	X88				
Niche Players	Red Point	Red Point (DataLever)	Human Inference		Netrics
	Data Mentors	Data Mentors	Data Mentors	Data Mentors	Data Mentors
	Datactics	Datactics	Datactics	Datactics	Datactics
	Innovative Systems	Innovative Systems	Innovative Systems	Innovative Systems	Innovative Systems
	Uniserv	Uniserv	Uniserv	Uniserv	Uniserv
			DataLever	DataLever	DataLever

2.3 Data/Information Quality Management

Poor data quality creates detrimental effects on businesses. Data flaws increase the costs for organization. In turn many functional areas are affected, including bids and proposals, research and development, human resources, and customer relationship management. Data management plays in many cost-saving initiatives. Good data management and data governance practices are helpful in ensuring data quality.

Ten years ago, Data Warehousing Institute reported that poor data quality costs six hundred billion dollars annually. Recently, English noted organizations spent 20-35% of operating revenue in recovery from process failure, information scrap and rework caused by poor data quality (2009). Poor data quality is costly for any organization. To resolve data quality issues, organizations invest in technologies to address data errors. Gartner's survey (2013) show more than two-thirds of organizations will increase their spending on data management services in the next year.

Total information quality management (TIQM) methodology developed by English (2003) is practical and useful in guiding and managing data quality. It follows the six sigma's define-measure-analyze-improve-control methodology. TIQM requires the establishment and deployment of roles, responsibilities, policies, and procedures concerning the acquisition, maintenance, dissemination, and disposition of data. Successful enforcement of TIQM depends on the partnership between business users and the technology group. The business users define business rules and meanings for data elements whereas the technology group builds architecture, databases, and applications to ensure the effective life span of data elements.

2.4 Big Data Quality

Master data management is crucial in data quality management. Master data management and data governance initiatives are critical programs organizations should have. In the big data era, with data volumes growing explosively, data strategy should focus on collecting the right and needed data. In the data quality tools market, there is a trend for data quality vendors and data integration vendors to converge. It is typical for companies with technologies of many generations made by various vendors to offer various methodologies.

Master data is the most important data for business. Commonly used master data includes customer, employee, products etc. All business transactions involve master data. Master data is used by many business applications and its definitions, standardization, and management is crucial in order to depend on the master data. Master data is used to generate reports for operational purposes, analyze trends, and identify anomalies for strategic purpose.

Data warehouses, data quality and data integration technologies work together to better safeguard data quality. Big data quality involves a variety of data types

with increased velocity. Assessing big data quality is more complicated than before with Hadoop or NoSQL. Technologies are the essential for managing big data; however new technologies are not mature and still need to evolve.

3 Data Privacy and Security

Privacy is of interest to each of us. However, as an individual, we do not know how to protect privacy. Technology advancements make "zero privacy" possible. Privacy means the right to be left alone. Information privacy or data protection laws prohibit the disclosure or misuse of information held on private individuals. Many countries in the world have enacted data privacy laws. Compared to Europe, the data privacy law is less regulated in the U.S.

The legal framework regulating the big data business model is built upon existing principles of intellectual property, confidentiality, contract and data protection law. Big data could potentially lead to big legal battles without proper security and privacy measures in place. Companies buy, sell and share data with other entities. It is important to know the sources of the data, and the extent to which the data can be re-used. When using big data, first and foremost companies must be certain that the data from various sources are anonymized.

Privacy invasions occur from time to time and the consequences of privacy invasions in healthcare are particularly salient. In this chapter, we focus on data privacy in healthcare due to the fact that the U.S. healthcare system is undergoing major changes and healthcare has more privacy regulations and requirements. Information privacy or data protection laws prohibit the disclosure or misuse of information held on private individuals. Laws, regulations, technologies, and hospital mergers all have played interwoven roles in privacy.

3.1 Healthcare Big Data

Most healthcare executives have high expectations for big data, but talent shortage and lack of resource are roadblocks to realization of big data's potential. Society of actuaries (2013) report that 87% of healthcare decision makers perceive big data as impactful in future; 84% find it difficult to find skilled people in optimizing big data; 45% plan to hire more skilled people in the next year. In terms of current value of big data, 45% states substantial benefits; 27% says that big data provides some benefits but not much; 22% of healthcare decision makers claims no benefits.

One-third of IT decision makers in healthcare say they have concerns about the potential of big data (35%); they consider big data as a double-edged sword with both opportunity and risk (34%); they also thinks that payers (55%) are more likely than providers (47%) to perceive big data as an opportunity. Only half of decision makers say their organization is very well-prepared to take full advantage of data growth (51%). Based on Society of Actuaries's survey, it seems that at health insurance companies are better prepared for big data than hospitals and health systems.

Decision makers at health insurance companies, hospitals and health systems set high expectations for big data. The benefits to be delivered by big data will include not only help set better financial decisions, in addition to population health and clinical outcomes. Yet, currently many of the expected benefits are not realized and healthcare is awaiting further developments in big data. As big data is such a new field, it is difficult to find people with the necessary big data experience and skill. In a few years, when big data technologies become increasingly mature and more people learn the skills, tangible benefits of implementing big data solutions will be realized.

3.2 Data Privacy in Healthcare

Data is generated at a compounding rate. With the increased healthcare data created at an astounding rate by advanced diagnostic equipment, PDAs, tablets, health monitoring devices etc. Data security and privacy issues have always been concerns in all industries. As consumers, we care about our identity and privacy. As patients, we want our medical records to be secured and the history of our medical records to be known only to those we deem important. Laws are important in restricting organizations and individuals' behavior and play vital roles in safeguarding privacy and security.

Health Insurance Portability and Accountability Act (HIPPA) was enacted in 1996. The Health Information Technology for Economic and Clinical Health (HITECH) Acts were signed into law in 2009. HIPPA and HITECH are designed to address the concerns associated with the electronic transmission of health information, the implementation of electronic information systems and the prevalence of cloud computing in healthcare. These new challenges have introduced unprecedented privacy and security challenges (Birk, 2013).

3.3 Data Security Overview

Experian reports "the healthcare industry, by far, will be the most susceptible to publicly disclosed and widely scrutinized data breaches in 2014" (Carr, 2014). In healthcare, social security numbers are stored and used as a unique identifier for many systems. Ponemon Institute survey reports that 94% of the respondents have at least one data breach in the last two years (2014). The 2010 HIMSS Analytics Report on Security of Patient Data shows that healthcare organizations are having major obstacles securing patient information, with efforts being largely reactive rather than proactive. 19% of the health organizations surveyed had a security breach as compared to 13% in 2008, of these, 84% were as a result of lost or stolen laptops, improper disposal of documents, stolen backup tapes, etc.; 87% of organizations note their data access, sharing, and security policies.

Today's hackers are more tech savvy and can infiltrate networks by exploiting vulnerabilities. IBM (2012) reports the commonly used avenues for breaches include exploiting default or easily guessed passwords, backdoor malware, use of

stolen credentials, exploiting backdoor or command and control channels, key loggers spyware, and SQL injection attacks.

Privacy and security risks can also come from connected medical devices and consumer health-monitoring gadgets. These devices play critical and important roles in remote patient monitoring and personal health management. Yet many of these devices do not have the basic security functions and do not provide privacy protections because the wireless data links used to transmit data and instructions are not encrypted (Carr 2013). Recently, many people are concerned with the HealthCare.gov website given the fact many entities can access the information on the website. The potential for misuse of healthcare information is great. In addition, patient portals present another source of risks. Patient Portals become a popular interface for patients to communicate with their healthcare providers. With patient portals, patients can leave a message for a doctor, make an appointment, request refills and enter health history data. Given so much personal health information available on the patient portal, security and privacy becomes a major concern.

It is important to note that data security and data privacy are different but interrelated issues. Data security prevents or grants access to data based upon authorization. Data privacy ensures that only those who have a valid need to view and/or utilize data can access data and that they work within the organization's policies. Privacy and security of patient health information must be at the core of any organization's policies. Policies and procedures should clearly define who has the right to access what part of health information and how data is protected, stored, and secured during transmission from one site to another. Security and privacy policies must be adhere to both federal and state laws.

3.4 Management and Policies

With the government mandating the application of electronic healthcare records systems, over half of the hospitals have implemented electronic medic records by 2012. As the data breach risk is quickly increasing, security and privacy takes the front stage. Data breaches are costly to any organizations. When such incidents happen, organizations face lawsuits, penalties and lost customers. Organizations must be vigilant about data security and privacy protocols. Organizations need to develop and implement various measures to strengthen its security and privacy. Organizations should set up a budget for security and privacy. Most of the health IT initiatives focus on meeting regulatory requirements, managing digital patient data, reducing costs, improving care, increasing clinical efficiency and collaboration. In the past security is not on top of the investment list, however, it should be a priority since the aftermath is far costlier. The average cost of encryption for a single device is $150. The average cost of an enterprise encryption system is between a quarter million to half a million dollars. Several vendors offer privacy-monitoring software to specifically meet the healthcare industry's needs. For high tech breaches, investment in encryption software is a necessity.

To manage security and privacy in the big data era, organizations should address not only external threats but also internal threats. In healthcare, traditionally many people do not think they are personally responsible for data management. In fact, technologies in healthcare are typically many years behind the financial industry and people in healthcare are less technology savvy and have limited knowledge of data security. Thus education and trainings are important in improving the understanding of privacy and security issues. Staff education is effective in dealing with low tech data breaches. Part of maintaining data security is educating end users in basic security measures and awareness of company policy. Educating employees regarding security dangers such as: clicking on email links, taping the password to the computer desk, or surfing unauthorized websites will help avert unintentional sabotage. Again, organizations need to install monitoring and assessment systems in place to ensure that employees are working within the boundaries of the organization's policies. According to Health Insurance Portability and Privacy Act providers may be responsible for their employees' data breaches in certain circumstances. With new threats coming in and continuous education on employees is the key to safeguard security.

Healthcare organizations develop policies to safeguard against security breaches and clearly define steps to deal with violations. Some of the measures use to protect patient data include security policy, data access monitoring, physical security, formal education among others. Data security should be a responsibility for the entire organization and not just specific departments. Developing this enterprise view on security will help reduce e so-called "low tech" security breaches and strengthen current policies and efforts. Therefore, organizations must utilize tactics that support such an initiative such as "rotating privacy and security audits to spot problems, including password on sticky notes and computers left unattended which displaying sensitive information; compliance with privacy and security rules; completion of education and training as metrics in performance evaluations; and daily hurdles to discuss patients on the floor who might spark curiosity – and inappropriate intrusion into their health information" (Birk, 2013).

It is important to have plans in action and be proactive. Reactive behaviors such as playing follow-up after the breach occurs are not effective. Healthcare data is ideal for the identity thief. Therefore, it is critical that the healthcare organizations setup risk management teams. The reality is that many organizations do not implement proper safety risk management when moving to web-based systems and the cloud. Organizations must always assess and reassess all systems and situations. It is crucial that proper steps are taken from the beginning to protect the organization against security and privacy breaches.

3.5 Big Security Data

Organizations of any size are exposed to the unprecedented number and variety of threats and risks to cyber security. Big Data will transform intelligence-driven models (Kar 2014). Research firm Gartner predicts that more than 25 percent of

global firms will adopt big data analytics for at least one security and fraud detection use by 2016.

The variety, volume and speed of security data have increased rapidly. As a result, analytic on big security data has become ever more important, especially as hacking algorithms and methods has become unexpectedly more sophisticated. The sheer volume of security related data makes it extremely difficult to identify a threat. However, big security data are useful in the safeguard of organizations security and the mining of security data making them proactive in defending their own networks and protecting organizations. McAfee (2014) security survey of IT decision makers show that 35% of organizations can detect data breaches within minutes of happening; 22% of organizations would need a day to identify a breach; 5% of the organizations would take up to a week. On average, it takes 10 hours for an organization to recognize a security breach. Given the serous security issues, big data analytics can play a critical role and allows organizations to access data, gain a complete view of business, perform effective security analysis, and detect advanced threats.

3.6 Security Products

Security experts have been predicting trends and challenges in the security activities. For example, Schwartz (2012) notes seven security trends for 2013 including: 1) mainstream cloud and mobile adoption seeks security;,2) businesses begin sandboxing smartphone apps; 3) cloud offers unprecedented attack strength; 4) post-flashback, cross-platform attack increase; 5) destructive malware targets critical infrastructure; 6) hackers target QR codes, TecTiles; 7) digital wallets become cybercrime targets. Hurst predicts ten security challenges for 2013 as follows: 1) state-sponsored espionage; 2) distributed denial of service (ddos) attacks; 3) cloud migration; 4) password management; 5) sabotage; 6) botnets; 7) insider threat; 8) mobility; 9) internet; 10) privacy laws.

With so many challenges identified and more to be discovered, security products can be effective in addressing challenges above. Table 2 shows Security Readers' Choice awards for top security products in 19 categories. In application security category, readers were asked to vote on Static and dynamic vulnerability scanners, and other source code analysis products and services used during development. In authentication category, readers voted on digital identity verification products, services, and management systems, including PKI, hardware and software tokens, smart cards, knowledge-based systems, digital certificates, biometrics, cell phone-based authentication. In Cloud security category, readers voted on Services and products designed to secure business use of cloud computing, including data encryption, identity and access management and network security.

In Data loss prevention category, readers voted on Network, client and combined data leakage prevention software and appliances for enterprise and midmarket deployments, as well as "DLP lite" email-only products. In email security category, readers voted on Antispam, antiphishing, email antivirus and

antimalware filtering, software and appliance products, as well as hosted "in-the-cloud" email security services. Includes email archiving and e-discovery products and services. In encryption category, readers voted on Hardware and software-based file and full-disk encryption, and network encryption products. In Endpoint security category, readers voted on business-grade desktop and server antimalware and endpoint protection suites that include antivirus and antispyware, using signature-, behavior- and anomaly-based detection, whitelisting, host-based intrusion prevention and client firewalls. In Enterprise firewalls category, readers voted on enterprise-caliber network firewall appliances and software, stateful packet filtering firewalls with advanced application layer and protocol filtering. In Identity and access management category, readers voted on User identity access privilege and authorization management, single sign-on, user identity provisioning, Web-based access control, federated identity, role-based access management, password management, compliance and reporting.

In Intrusion detection and prevention category, readers network-based intrusion detection and prevention appliances, using signature-, behavior-, anomaly- and rate-based technologies to identify denial-of service, malware- and hacker-attack traffic patterns. In Mobile data security category, readers voted on smartphone and tablet data protection products including antimalware, mobile access, platform-specific security (Android, iOS, Windows and BlackBerry), mobile device management, mobile application management and mobile application security. In Network access control category, readers voted on appliance, software and infrastructure user and device network access policy creation, compliance, enforcement (802.1X, client-based, DHCP) and remediation products. In Policy and risk management category, readers voted on risk assessment and modeling, and policy creation, monitoring and reporting products and services, IT governance, risk and compliance products, and configuration management. In remote access category, IPsec VPN, SSL VPN (stand-alone and as part of application acceleration and delivery systems) and combined systems and products, as well as other remote access products and services.

In SIEM category, readers voted on security information and event management software, appliances and managed services for SMB and enterprise security monitoring, compliance and reporting. In Unified threat management category, readers voted on UTM appliances that integrate firewall, VPN, gateway antivirus, URL Web filtering, antispam. In Vulnerability management category, readers voted on network vulnerability assessment scanners, vulnerability risk management, reporting, remediation and compliance, patch management and vulnerability lifecycle management. In Web application firewalls category, readers voted on standalone Web application firewalls and WAFs that are part of application acceleration and delivery systems. In web security category, readers voted on Software and hardware products, hosted Web services for inbound and outbound content filtering for malware activity detection/prevention, static and dynamic URL filtering and application control (IM, P2P).

Table 2 Security Readers' Choice Awards 2013

Security Category	Gold	Silver	Bronze
Application security	QualysGuard WAS, Qualys Inc.	Juniper Networks AppSecure, Juniper Networks Inc.	API Gateways, Layer7 Technologies Inc.
Authentication	SecurID, RSA, the security division of EMC Corp.	Symantec Managed PKI for SSL, Symantec Corp.	
	Symantec User Authentication Solutions, Symantec Corp.		
Cloud security	Juniper Networks vGW Virtual Gateway, Juniper Networks Inc.	Symantec Email Security.cloud, Symantec Corp.	Symantec O3, Symantec Corp.
Data loss prevention	Symantec Data Loss Prevention, Symantec Corp.	Websense Data Security Suite, Websense, Inc.	McAfee Total Protection for Data, McAfee, Inc.
Email security	Messaging Gateway powered by Brightmail, Symantec Corp.	Cisco Email Security Appliance (formerly IronPort), Cisco Systems	Google Message Security, Google
Encryption	Dell Data Protection - Encryption, Dell Inc.	Check Point Full Disk Encryption, Check Point Software Technologies Ltd	SecureData Enterprise, Voltage Security, Inc.
Endpoint security	Symantec Endpoint Protection 12, Symantec Corp.	Kaspersky Endpoint Security for Business, Kaspersky Lab	AVG AntiVirus Business Edition, AVG Technologies
Enterprise firewalls	McAfee Firewall Enterprise, McAfee, Inc.	Juniper Networks SRX Series Services Gateways for the Data Center, Juniper Networks, Inc.	Juniper Networks ISG Series Integrated Security Gateways, Juniper Networks, Inc.

Table 2 (*continued*)

Identity and access management	Oracle Identity and Access Management Suite Plus, Oracle Corp.	RSA Identity Protection and Verification Suite, RSA, the security division of EMC	CA IdentityMinder, CA Technologies Inc.
Intrusion detection and prevention	Juniper Networks IDP Series Intrusion and Prevention Appliances, Juniper Networks Inc.	Fortinet FortiGate, Fortinet Inc.	Check Point IPS Software Blade, Check Point Software Technologies Ltd.
Mobile data security	McAfee Enterprise Mobility Management, McAfee Inc.	AirWatch MDM, AirWatch LLC	Check Point Mobile Access Software Blade, Check Point Software Technologies Ltd.
Network access control	Unified Access Control, Juniper Networks Inc.	McAfee Network Access Control, McAfee Inc.	Cisco NAC Appliance, Cisco Systems
Policy and risk management	IBM Tivoli Compliance Insight Manager, IBM Corp.	VMware vCenter Configuration Manager, VMware Inc.	McAfee ePolicy Orchestrator, McAfee Inc.
Remote access	Check Point Remote Access VPN Software Blade, Check Point Software Technologies LTD.	Juniper Networks SA Series SLL VPN Appliances, Juniper Networks	Netgear ProSafe VPN Firewall, Netgear
SIEM	Splunk Enterprise, Splunk Inc.	HP ArcSight Enterprise Security Manager (ESM), Hewlett-Packard Co.	McAfee Security Information and Event Manager, McAfee, Inc.
Unified threat management	Dell SonicWall, Dell Corp.	Check Point Unified Threat Management, Check Point Software Technologies LTD.	FortiGate, Fortinet, Inc.

Table 2 (*continued*)

Vulnerability management	Shavlik Protect, LANDesk Software	QualysGuard Vulnerability Management, Qualys, Inc.	Nessus Vulnerability Scanner, Tenable Network Security
Web application firewalls	Citrix NetScaler AppFirewall, Citrix Systems Inc.	FortiWeb-400C, Fortinet, Inc.	F5 Networks BIG-IP Application Security Manager, F5 Networks
Web security	Websense Web Security Gateway, Websense Inc.	Blue Coat Systems ProxySG appliances, Blue Coat Systems, Inc. 90-100 words	Symantec Web Security.cloud, Symantec Corp.

Source: http://searchsecurity.techtarget.com/essentialguide/Security-Readers-Choice-Awards-2013

4 Big Data Analytics

4.1 Overview

Big data analytics applies advanced analytics to very large data sets. According to a 2009 TDWI survey, 38% of organizations surveyed reported practicing advanced analytics, whereas 85% said they would be practicing it within three years. Forrester (2013) reports that 70% of IT decision makers consider big data a top priority now or in a year. In addition, a majority of companies that Forrester surveyed estimate that they are only analyzing 12% of the data they have. Big Data is changing the way of products, solutions, and services are being marketed. McKinsey & Company (2012) reported that big data and improved analytics can improve sales by $200 billion.

Big data analytics deal with complex data types. Querying such complex data types could be challenging and analysis consumes large amount of resources: storage, memory, CPU. Query performance could be affected. Building a solid infrastructure to support fast data through output and improving query performance are critical in big data analytics. Currently, analytics are used in reporting, dashboard, performance analytics, web analytics, and process, predictive, location analytics, advanced visualization, text analytics and streaming analytics.

4.2 Technologies

New technology – Hadoop holds the promise big data analytics. Hadoop and related products make data capturing, storage and analysis in a cost effective way. When investing in big data technologies, scalability is the key. Organizations must think ahead and be future-oriented. The volume of data is exploding and a variety of devices (mobile phones, sensors, websites) produce different data types (web data, image files, video and audio files). The new innovative devices coming to market will continue to change the future of the big data market. Thus, big data infrastructure must be able to accommodate future data growth needs.

Hadoop is a distributed file system that handles massive volumes of file-based unstructured data. It becomes the de facto standard for big data technologies. It is an open source software project administered by the Apache Software Foundation. Because Hadoop is essentially a distributed file system, it lacks some functionality of database management systems. Furthermore, a set of related software technologies (Pig, MapReduce, Hive, HBase) work together to become the Hadoop family of products. Both the Apache Software Foundation and several software vendors offer Hadoop family products. With the high hopes for Hadoop in dealing with big data, more and more vendors make an effort in integrating their products with Hadoop.

Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs and it can be used to create mapper and reducer to work on large data sets. Pig consists of two components: PigLatin and the runtime environment. MapReduce is a software framework for creating applications to handle vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware. Hive is the Apache data warehouse software performing querying and managing large datasets located in distributed storage. Hive is not a standard SQL but resemble SQL. HBase is a non-relational database and can host very large tables (billions of rows and millions of columns) for random, realtime read/write access. HBase was modeled after Google's Bigtable and has similar functions as Bigtable.

The commonly referred Hadoop essentially contains a set of related technologies in the big data environment and each of these technologies has their own unique advantages in handling and processing large data sets. Together, Hadoop, Pig, MapReduce, Hive, Hbase etc. leverage computing resources in order to efficiently process big data. Hadoop is designed for multi-structured data types whereas a data warehouse handles structured data.

According to TWDI survey (2012), most organizations are planning to integrate Hadoop into their existing architecture. TDWI also predicts that Hadoop technologies will complement the well-established products and practices for business intelligence (BI), data warehousing (DW), data integration (DI), and analytics. However, Hadoop is a new technology, its security and administrative tools need improvements. In addition it is hard to find Hadoop users and technical professionals. As time passes by, there will be more Hadoop users. 10% of organizations surveyed have a Hadoop implementation in production today (Russon 2013)

For example, an organization can store aggregated web log data in their relational database, while keeping the complete datasets at the most granular level in Hadoop. This allows them to run new queries against the full historical data at any time to find new insights, which can be a true game-changer as organizations aggressively look for new insights and offerings to differentiate from the competition. The popular Hadoop products include MapReduce, HDFS, Java, Hive, HBase, and Pig. Mahout, Zookeeper, and HCatalog are taking off.

Enabling big data analytics is the leading benefit of Hadoop, whereas a lack of Hadoop skills is the leading barrier. BI/DW aside, a few respondents also anticipate using Hadoop as a live archive (23%) or as a platform for content management (35%).

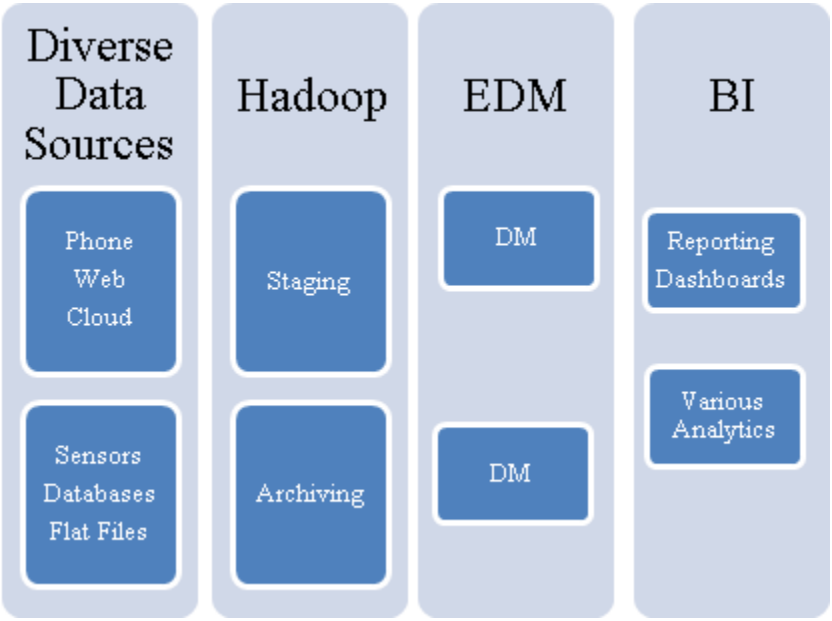

Fig. 1 Big Data Architecture here

4.3 Business Decision Making

Big data analytics operate on large data sets captured by diverse devices from sensors, devices, third parties, web applications, and social. Detailed and granular level data of business operations can be analyzed and new sights can be generated. Advanced analytics techniques such as: predictive analytics, data mining, statistics, and natural language processing can be applied to unstructured and structured data types consequently create fact-based decisions. Results derived from big data analytics may help in the areas of customer segmentation, fraud detection, risk analysis, and tracking evolving customer behaviors. Business decisions that use

deep and advanced data analytics provide benefits on operational, tactical and strategic levels. Furthermore, there are many advantages of implementing big data analytics, which result in optimization, performance improvement, costs reduction of the organization as a whole.

While long-term expectations are high, most of healthcare executives say they have yet to see substantial benefits from big data. The Society of Actuaries (SOA 2013) recent report found that 66% of leaders are enthusiastic about the potential of big data, while more than 87% said that data analytics will have an important impact on the business of healthcare in the future. Another half of payers saying there is substantial business benefit (53%) by implementing big data analytics.

Healthcare has collected enormous amount of data across many disparate systems. In order to reduce costs and improve care outcomes, analytics must be used. In the past, structured data have been used in the administrative and financial area of healthcare, with more and more text data captured. In terms of patient care, text data provide more valuable insights in understanding cause and outcome. Textual and predictive analytics tools can reveal hidden patterns.

5 Discussion

Though many organizations have and are eager to jump on the big data wagon, implementing big data solutions is not easy. It is quite common that large organizations have technologies, tools, and architectures in different generations with diverse, complex information compliance and security policies. Business Leaders are not as confident about data as the volume, variety and velocity of information increases. Data quality issue still persists. If business leaders do not think data they have is trustworthy, they will not make decisions based on analytics generated on data.

Data quality, privacy and security are the main issues that should be tackled by any organization. Processes and procedures are the root causes for data quality issues and these cannot be corrected by technologies. A disciplined methodology should be followed through in data quality management.

Data quality in the big data area is more complex. However proven data management methodologies still apply to the big data environment. To be effective in big data management, data quality tools, data integration tools, data stewardship, policies and procedures must be in place. In addition, top management support in big data quality initiative is also a necessity. Furthermore, to ensure success in big data implementations, it is important to focus on security intelligence solutions.

5.1 Market Demand for Big Data Talents

McKinsey Global Institute (2011) reports that there will be a shortage of 140,000 to 190,000 people with deep analytical skills and 1.5 million managers and analysts who have the analytic skills of big data for effective decision-making. In addition, most organizations are not prepared to address both the technical and management challenges posed by big data.

In healthcare, the shortage of big data skills is also a concern. The overall shortfalls of resources including staffing, budget, and infrastructure are the biggest barriers to the adoption of big data analytics in their organization. Payers and providers have difficulty obtaining the funding for recruiting staff skilled to gain the full benefits of big data. Most providers and payers have difficulty finding staff that can consolidate complex datasets and glean actionable information from them. In addition, it is hard for payers and providers to hire the right talent who can identify the business opportunities big data provides. It would be good idea for healthcare to find talents in other industry.

Furthermore, we need to educate professionals in the importance of big data security. It has become obvious that technology alone will not solve security issues. People are backbone in in protecting cyberspace. Though security education has been addressed in the past, threats and attacks have been increasingly insidious and harmful. In order to be more effective, we need to hold mandatory security training session for all employees on all aspects: workflow, process, security policies, software, antivirus software etc. For universities, we need to offer full-blown curriculum addressing security to better prepare students for cyber defense.

5.2 Big Data Solutions Implementations

When it comes to big data solution implementations, Forrester (2013) finds six challenges as 1) integrating big data solutions in a complex, heterogeneous data management environment; 2) technical implementation skills with employees; 3) meeting business demand for big data analytics; 4) understanding business values of big data solutions; 5) infrastructure budget constraints; 6) difficulty in finding and hiring people with needed skills. In addition, Forrester (2013) also recommends seven qualities for production-ready big data solutions: 1) manageability, 2) availability, 3) performance, 4) scalability, 5) adaptability, 6) security, 7) cost.

5.3 Analysis of Big Data Publications

We analyzed 220 white papers on big data published between 2010-2014 using SAS text miner. We examined the concept links of the top four terms (data, big, business, big data) based on frequency. Data is linked to analytics, big, base, design, big data, analyze, type, software, and analysis. Analytics links to unstructured, predictive, unstructured data, business intelligence, visualization, predictive analytics and sensor. Big data is linked to Hadoop, network, open, storage, industry, variety, amount, source, enterprise, support, access, information, cluster, and exist. Big, is linked to development, industry, big data, software, environment, manage, strategy and challenge. Business, performance, thing, decision, practice, result, opportunity, insight, strategy, and analytics.

Table 3 shows the frequencies of terms in descending order in the 220 white papers we selected. As shown TDWI publishes most papers related in big data. The combination of words tdwi, user, organization and analytic are commonly

mentioned in 10 documents. Next frequently occurred word group is pillar, four, governance, Hadoop and node and the combination of these words occur in 16 documents.

Table 3 Term/Topic Frequency

Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
0.554	0.038	tdwi,+tdwi,+user organization,+analytic,+bi	713	10
0.539	0.036	+pillar,four,governance,hadoop,+node	709	16
0.569	0.039	+integrator,+five,+data integration,+integration,heterogeneous	672	14
0.553	0.042	intelligence unit,+organisation,limited,+unit,rst	648	11
0.608	0.043	hr,hr,+bi,+customer,+workforce	643	26
0.637	0.045	hadoop,+hadoop,+home,+big,+big data	630	22
0.26	0.03	+mobile,digital,+agency,+big data,+sector	607	30
0.32	0.031	+quality,+review,+propose,+theory,+opinion	603	20
0.443	0.034	+analytical,+analytics,+bi,+predictive,+big	580	34
0.487	0.037	+patient,+healthcare,clinical,+readmissions,+health	575	12
0.496	0.038	,+myth,+quality,integrity,+data quality	566	13
0.429	0.036	+builder,iway,+quality,+information builders,ltd	558	10
0.397	0.035	governance,home,+home,+li,+data governance	557	22
0.498	0.035	+federation,hadoop,+lasr,hdfs,+cache	556	20
0.486	0.038	infosphere,+biginsights,+puredata,+puredata system,hadoop	527	26
0.449	0.034	+inmemory,+cube,cognos,+dynamic,+cache	520	17

Table 3 *(continued)*

0.494	0.037	hadoop,forrester,+forrester wave,edw,+wave	499	10
0.227	0.028	+sap,hp,hp,+tb,+in-memory	496	19
0.393	0.033	+chart,+visual,+visualization,+plot, +visualize	490	14
0.33	0.031	+backup,+encryption,+protection,+cloud,data protection	490	9
0.158	0.026	+chapter,+springer,+book,+quo,+style	488	13
0.468	0.041	+security,+threat,+attack,+malware, +breach	456	15
0.415	0.038	+title,+reference,+style,+pt,+head	404	9
0.431	0.037	s p o n s,o f,pa g e,pa,e d b y home	359	19
0.283	0.029	+bi,+rs,+benchmark,+dan,+ar	357	10

Table 4 shows terms appeared in documents in descending order. The word group (analytics, analytic, BI,big, predictive) appears in 34 documents; the word group (mobile,digital,+agency,+big data,+sector) occur in 30 documents; the word group (hr,hr,+bi,+customer,+workforce) appears in 26 documents. The occurrence of the words show the importance of the concepts.

Table 4 Document Frequencies

Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
0.443	0.034	+analytical,+analytics,+bi,+predictive, +big	580	34
0.26	0.03	+mobile,digital,+agency,+big data,+sector	607	30
0.608	0.043	hr,hr,+bi,+customer,+workforce	643	26
0.486	0.038	infosphere,+biginsights,+puredata,+puredata system,hadoop	527	26
0.637	0.045	hadoop,+hadoop,+home,+big,+big data	630	22
0.397	0.035	governance,home,+home,+li,+data governance	557	22
0.32	0.031	+quality,+review,+propose,+theory,+opinion	603	20

Table 4 *(continued)*

0.498	0.035	+federation,hadoop,+lasr,hdfs,+cache	556	20
0.227	0.028	+sap,hp,hp,+tb,+in-memory	496	19
0.431	0.037	s p o n s,o f,pa g e,pa,e d b y home	359	19
0.449	0.034	+in-memo-ry,+cube,cognos,+dynamic,+cache	520	17
0.539	0.036	+pillar,four,governance,hadoop,+node	709	16
0.468	0.041	+security,+threat,+attack,+malware,+breach	456	15
0.569	0.039	+integrator,+five,+data integra-tion,+integration,heterogeneous	672	14
0.393	0.033	+chart,+visual,+visualization,+plot,+visualize	490	14
0.496	0.038	,+myth,+quality,integrity,+data quality	566	13
0.158	0.026	+chapter,+springer,+book,+quo,+style	488	13
0.487	0.037	+patient,+healthcare,clinical,+readmissions,+health	575	12
0.553	0.042	intelligence unit,+organisation,limited,+unit,rst	648	11
0.554	0.038	tdwi,+tdwi,+user organiza-tion,+analytic,+bi	713	10
0.429	0.036	+builder,iway,+quality,+information builders,ltd	558	10
0.494	0.037	hadoop,forrester,+forrester wave,edw,+wave	499	10
0.283	0.029	+bi,+rs,+benchmark,+dan,+ar	357	10
0.33	0.031	+backup,+encryption,+protection,+cloud,data protection	490	9
0.415	0.038	+title,+reference,+style,+pt,+head	404	9

5.4 Big Data Security

Cyber security is of critical importance to any company and individual. Cyber-attacks have grown increasingly sophisticated, stealthy, and dangerous. According to the Verizon 2013 Data Breach Investigations Report, 66% of breaches are not

discovered for months. Security intelligence is a proactive strategy in advance to manage security. Though companies have many security systems, typically these security systems do not talk to each other. Integrating security data from many disparate systems is essential in security management. More security data, transaction data, unstructured data can help identify vulnerability and filter out noise. The main sources of vulnerability can be classified into two categories: internal and external. In general for internal vulnerability, human errors are the weakness links in the security model. Security experts should be able to analyze if vulnerability will be raised because of errors.

With the increasing level of sophistication of threat compounding the fact that disparate security systems cannot talk to each other, it is infeasible to obtain a consolidated view about security situation. Some attacks occur quietly and attack slowly. In reality, it's not a matter of if attacks come in; it is a matter of when attacks come in and how long it takes for the security team to figure out. Security teams should start with normal behavior, normal network traffic, and normal user activities. Using the normal behaviors as security baseline will help find anomaly.

It is critical that organizations build a model to address Advanced Persistent (APTs). APTs are hard to identify with disparate security systems. Insider threats come from entities who have the motivation and ability to harm enterprises. APT is a set of stealthy and continuous hacking processes which usually intend to harm organizations, nations for business or political motives. APT processes can be very long and slow. This type of malware hides itself on the system over a long period of time.

The Ponemon Institute reports that 67% of organizations state their current security activities are insufficient to stop a targeted attack. Targeted attacks are difficult to predict, diagnose and defend. A custom attack requires a custom defense. There are companies specializing security technologies that detect and analyze APTs and targeted attacks.

APTs are dangerous and its hackers are well organized with an intent to steal valuable intellectual property, such as confidential project descriptions, contracts, and patent information. Grimes (2012) suggest five signs to watch for APTs: 1) increasing in elevated log-ons late at night, 2) finding widespread backdoor Trojans, 3) unexpected information flows, 4) discovering unexpected data bundles, 5) detecting pass-the-hash hacking tools. In addition, Frank and Watson (2013) recommend the importance of detecting account abuse by insiders and APTs, pinpointing data exfiltration by APTs, alerting new program execution.

Privacy of patient information must be a key priority when implementing the new analytics systems. In addition, the information should be secure, especially when being transported via the web and/or the cloud from one place to another. HIPAA requires that this data be private and secure. As web-based systems and cloud computing gains ever increasing popularity, these types of systems bring their unique characteristics, which expose to new security breaches. Organizations should implement proper safety risk management before moving to web-based

systems and the cloud. Organizations need to continuously assess and reassess all systems and situations for possible data breaches. It is crucial that proper steps are taken from the very beginning to protect the organization against security and privacy breaches.

6 Conclusion

Big data has generated a lot of interest not only in industry but also in academia. In this book chapter we have discussed data quality, data integration, privacy, security and analytics. We believe that these topics continue to be important in the next decade. New technologies, techniques, policies are to be developed to keep up with big data development. In addition, ethical and legal issues regarding big data present numerous challenges and opportunities for researchers.

According to IDC, about 22% of digital information is suitable for analysis, a lot of data reside in data silos (Lev-ram 2014). Data integration efforts continue to sustain and the market for technological vendors are becoming more promising. Data centers will continue to grow. In the future big data will become bigger data (Lev-ram 2014). More and more companies will invest in their data centers and expect to derive values from these.

Data quality is of interest to both practitioners and academia. To ensure data quality, technologies, people, policies and procedures working together to produce desired effect. Data quality must be ensured into order for data analysis to be reliable. The commonly used dimensions of data quality will be cherished across organizations: data accuracy, relevancy, timeliness, trustworthy. As data is treated as an asset of an organization, the value of data quality should be emphasized across the different hierarchies.

Only 5% of digitized data is currently being analyzed (Lev-ram 2014). Given such a small percentage， big data analytics has many areas waiting to be explored. Some suggest technologies, new statistical techniques are useful for big data analysis and others state the future of big data analytics should focus on effect size and variance explained (George 2013) rather than p values. In addition, big data visualization is another key research directions in the future as it makes big data meaningful by transcribing massive amount of information into images that are easy for people to understand.

Finally, in the education arena, a few universities across the country started data science as a new major. The purpose of such a discipline is to train students in big data technologies and big data analytics to meet the emerging market demand. In fact, with the new job title coming to the market (e.g. chief financial technology officer), CFOs are required to understand technology and big data. Brand (2014) states that “The future of the accountancy profession lies at the intersection of finance, technology and information”. As an educator, we need to keep up with technological developments and collaborate with our colleagues to provide cross disciplinary skills for students.

References

Birk, S.: Protecting patient medical data. Healthcare Executive 28(5), 20–28 (2013)

Brands, K.: Big Data and Business Intelligence for Management Accountants. Strategic Finance 96(6), 64–65 (2014)

Brand, H.: Big data: adapt or die (2014), https://www.accountancylive.com/big-data-adapt-or-die (accessed June 15, 2014)

Carr, D.F.: Hackers outsmart pacemakers, fitbits: worried yet? InformationWeek (2013), http://www.informationweek.com/healthcare/security-and-privacy/hackers-outsmart-pacemakers-fitbits-worried-yet/d/d-id/1113000?image_number=3 (accessed June 14, 2014)

English, L.P.: Information quality applied: best practices for improving business information, Processes and Systems. Wiley (2009)

Forrester, Is your big data solution production-ready? (2013), http://www.itworld.com/data-center/417766/your-big-data-solution-production-ready (accessed June 15, 2014)

Fulgoni, G.: Big data: friend or foe of digital advertising? Five ways marketers should use digital big data to their advantage. Journal of Advertising Research 53(4), 372–376 (2013)

Gartner report, Magic quadrant for data quality tools (2013), http://www.gartner.com/technology/reprints.do?id=1-1LE6U4H&ct=131008&st=sg (accessed February 26, 2014)

Kim, G.-H., Trimi, S., Chung, J.-H.: Big-data applications in the government sector. Communications of the ACM 57(3), 78–85 (2014)

George, G., Haas, M.R., Pentland, A.: Big data and management. Academy of Management Journal 57(2), 321–326 (2014)

Research Moz, Global data quality tools market is expected to reach a CAGR of 16.78% in 2016 (2013), http://www.prweb.com/releases/2013/11/prweb11352256.htm (accessed February 25, 2014)

Green, C.: Organizations will rapidly ramp up their data services in 2014 (2014), http://blogs.forrester.com/charles_green/14-02-06-organizations_will_rapidly_ramp_up_data_services_spend_in_2014 (accessed February 25, 2014)

Grimes, R.: 5 signs you've been hit with an advanced persistent threat (2012), http://www.infoworld.com/d/security/5-signs-youve-been-hit-advanced-persistent-threat-20494 (accessed March 24, 2014)

Hurst, S.: Top 10 security challenges for 2013. SC Magazine (2013)

IBM Corporation. Three guiding principles to improve data security and compliance: A holistic approach to data protection for a complex threat landscape (2012)

Kar, S.: Gartner report: big data will revolutionalize cybersecurity in the next two years. CloudTimes (2014)

Lawson, L.: Eight questions to ask before investing in data quality tools (2014), http://www.itbusinessedge.com/blogs/integration/eight-questions-to-ask-before-investing-in-data-quality-tools.html (accessed February 26, 2014)

McAffee, Needle in a datastack: the rise of big security data (2013), http://www.mcafee.com/us/about/news/2013/q2/20130617-01.aspx (accessed January 15, 2014)

Lev-ram, M.: What's the next big thing in big data? Bigger data. Fortune 169(8), 233–238 (2014)

Mcknight, W.: Seven sources of poor data quality. Information Management 19(2), 32–33 (2009)

McKinsey Global Institute, Big data: next frontier for innovation, competition, and productivity (2011)

McMillan, M., Cerrato, P.: Healthcare data breaches cost more than you think. InformationWeek Reports (2014)

Nunan, D., Di Domenico, M.: Market research and the ethics of big data. International Journal of Market Research 55(4), 2–13 (2013)

Petter, S., DeLone, W., McLean, E.R.: Information systems success: the quest for the independent variables. Journal of Management Information Systems 29(4), 7–62 (2013)

Russom, P.: Integrating hadoop into business intelligence and datawarehousing. TWDI Research (2013), http://www.cloudera.com/content/dam/cloudera/Resources/PDF/TDWI%20Best%20Practices%20report%20-%20Hadoop%20foro%20BI%20and%20DW%20-%20April%202013.pdf (accessed February 15, 2014)

Schwartz, M.J.: 7 Top Information security trends for 2013. InformationWeek (2012), http://www.darkreading.com/risk-management/7-top-information-security-trends-for-2013/d/d-id/1107955? (accessed January 23, 2014)

Setia, P., Venkatesh, V., Joglekar, S.: Leveraging digital technologies: how informa-tion quality leads to localized capabilities and customer service performance. MIS Quarterly 37(2), 565-A4 (2013)

Smith, R.F., Watson, B.: 3 Big data security analytics techniques you can apply now to catch advanced persistent threats. HP Enterprise Security (2013)

Society of Actuaries, Healthcare decision makers perspectives on big data (2013)

TDWI's Data Quality Report, http://tdwi.org/research/2002/02/tdwis-data-quality-report.aspx (accessed March 2, 2014)

Verizon. 2013 Data Breach Investigations Report, http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf (accessed March 2, 2014)

Woods, D.: Why data quality matters (2009), http://www.forbes.com/2009/08/31/software-engineers-enterprise-technology-cio-network-data.html (accessed February 25, 2014)

Search, Analysis and Visual Comparison of Massive and Heterogeneous Data

Application in the Medical Field

Ahmed Dridi, Salma Sassi, and Anis Tissaoui

Abstract. It is thanks to the continuous evolution of the hardware technology that enables information systems to store very large amounts of data, the latter that explode even more rapidly than the growth rate of computing power. This spectacular growth of data is at the origin of what is called the Big Data. As several fields are affected by digitization, the medical field has experienced in the past years, an important technological and digital revolution, which contributed to a large informational explosion of digital medical data. In addition to their massive quantity, these data are also characterized by the complexity, diversity and heterogeneity, and they are often contained in the so-called the Electronic Health Record (EHR). However, not having the right tools to explore the large amounts of data that have been collected because of their potential usefulness, the data becomes unnecessary and databases and their management systems become without advantage. In this context, we propose in this paper the Medical Multi-project ICOP system (M^2ICOP) which was an interactive system dedicated particularly to clinicians and researchers in the medical field to help them explore, visualize and analyze a set of medical data really large and heterogeneous. Practically, our system allows these users to visualize and interact with a large number of electronic health records, and the search of similar EHRs and the comparison between them to take advantage of best practices and shared experiences to improve the quality of treatment.

Ahmed Dridi · Salma Sassi
Faculty of Law, Economics and Management of Jendouba,
University of Jendouba, Avenue de l'U.M.A, 8189 Jendouba, Tunisia
e-mail: ahmed-dridi@outlook.com, sassisalma@yahoo.fr

Anis Tissaoui
High Institute of Management of Gabes, Rue Jilani Habib, Gabs 6002, Tunisia
e-mail: tissaouianis@yahoo.fr

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_15

1 Introduction

Since the 1980s until today, the digital and informational world has undergone significant and progressive growth of digital data amounts through the storage capacity that was almost doubled every forty months (Hilbert & López, 2011). According to (Hilbert & López, 2011), taking account of the average growth of the data that was estimated at 59% each year, this percentage will be probably much higher in a few years. In fact, from 2012, the world produces each day 2.5 exabytes (2.5 x 10^{18}) of data (Halper, 2012), and it is expected that the amount of data in 2020 touch the roof of 40 zettabytes after it was passed in 2012 by 2.8 zettabytes, according to an IDC study sponsored by EMC (Rometty, 2013).

This phenomenon is called *"Big Data"* and is identified as one of the biggest trends of Information Technology for the year 2012 (CeArley & Claunch, 2012). Many definitions can be found in the literature for this new term. Generally, it refers to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time (Snijders et al., 2012). In a research report in 2001, Doug Laney, an expert analyst at Gartner, said that the massive data are characterized by three "V" (Laney, 2001). The first 'V' indicates all that is around the volume, i.e. the ability to capture, store and access large volumes of data permanently. The second is associated with the velocity that represents both the frequency at which the data are generated, captured and shared. Finally, the last 'V' is relative to the variety that refers to the various types of data from different sources. In fact, traditionally, the companies and different organizations analyze data that are so-called structured data, such as that found in data bases and relatively simple spreadsheets, but they constitute only 20% of full data. 80% of the remaining data, according to IBM, are known as raw, semi-structured or unstructured data, which should be structured to use it (Pierre & Marc, 2013). To these three characteristics (Volume, Velocity, Variety), specialists agree to characterize the Big data (Stefan, 2012). Until our days, this model, collectively called the *3V model* is still widely used to describe this phenomenon (Beyer, 2011). In addition, another 'V' was added by IBM to indicate the Accuracy of data (Brian & Boris, 2011), which mainly refers to the integrity of data and the ability of an organization to trust the data and be able to confidently use to make crucial decisions.

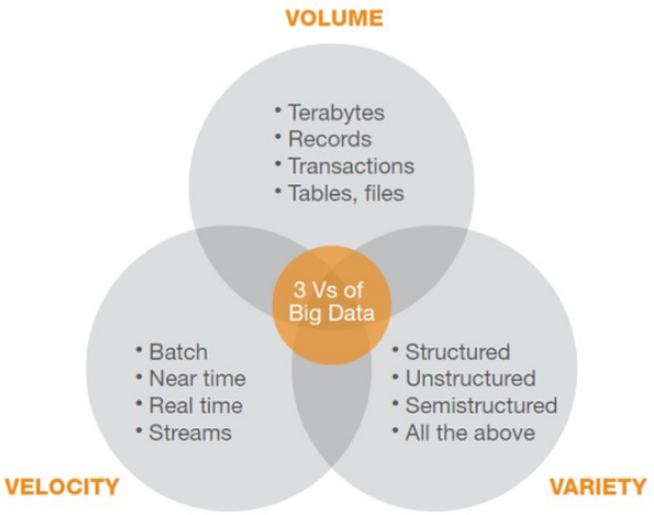

Fig. 1 The 3 V's of Big Data

The phenomenon of Big Data finds application in many fields such as the political field , commercial, cultural, social, scientific and technological fields, also the medical and bio-informatics ones. In this paper, we are particularly interested in the challenges of massive data in *health care*. In fact, in recent years, medical data have evolved exponentially and faster than health care organizations can consume (IBM, 2011). Therefore, the need to apply the Big Data and capabilities to capture, collect, view, and analyze the data is more topical than ever. These data are located in several places such as systems of analysis laboratories, radiology and imaging centers, doctors notes, claims systems, CRM systems, systems of organizations of medical insurance and social security, systems of different health care establishments and more particularly in the electronic health records of individuals (IBM, 2011).

In fact, the idea of keeping a numerical history of health for each individual and to overcome traditional medical records and their problems is the first step towards the emergence of what is called the *Electronic Health Record (EHR)*, which is defined as a systematic collection of electronic health information about individual patients or populations (Gunter & Terry, 2005). In the last few years and with the fast and growing development of EHRs, the designers of these systems are now in front of a big challenge to implement innovative and effective visual methods to support clinical decision-making and research. This will allow to meet the requirements and expectations of clinicians and researchers in the medical field.

In this context, several research efforts have been carried out whose derived results consist of a variety of systems and tools for visualization and management of EHRs. Some work is oriented towards the management of

a single record, and others are designed for the display of several projects. However, some of them do not support specifics users intentions like research and comparison.

This article is structured as follows: section 2 presents in the first part a State of the art on the data visualization and in the second part, we introduce an overview of existing work of EHR visualization system. Section 3 describes the system that we proposed by describing its architecture and its basic concepts. The fourth section presents a description of the validation prototype of our system. Section 5 is devoted to the evaluation of the system. And finally, in the last section the future work will conclude this article.

2 Related Work

2.1 Data Visualisation

Today more than ever, we have large amounts of data, without the right tools and adequate processing techniques, these data are worthless. To give meaning to these data they must be represented and displayed graphically (Cleveland, 1993). This is what is called the Data visualization. In the IT community, several definitions are already elaborated to define this notion. According to Michael Friendly, data visualization is the science of visual representation of data, defined as information which has been abstracted in some schematic form, including attributes or variables for the units of information (Friendly, 2008). The field data visualization has its origins in the early days of computer graphics, in the 1950s, when the first graphics and images were generated by computers (Owen, 1999). From the beginning, the main goal of data visualization is to communicate information clearly and effectively using graphical tools, in order to organize the complexity of the data, increase interaction with them and of course to amplify human cognition and help him to evaluate these data and extract new knowledge (Fernanda & Martin, 2011).

In this context, various types of data visualization techniques have been proposed, and for each type, many tools have been developed. In the following, we present an overview on the different types of data visualization and tools:

According to Solveig Vidal (vidal, 2006), there are four types of visualization techniques that have been developed from different perspectives:

Linear visualization: this type of visualization is related to linear data such as alphabetic lists, text documents, tables, the source code of the programs and chronologically ordered sets. There are three main approaches depending on whether we are dealing with chronologically ordered data or not:

- Large spreadsheets like *TableLens* (Rao & Card, 1994), *DEVise* (Livny et al., 1997), and *Seesoft* (Eick et al., 1992).
- Perspective walls as *Perspective Wall* (Mackinlay et al., 1991).

- The historical friezes or Timelines as *LifeStreams* (Freeman, 1997) and *LifeLines* (Plaisant et al., 1998).

Hierarchical visualization: that is employed in the case of hierarchical data organized hierarchically via tree structures with linkages between parent nodes and child nodes. It is most often information type catalog of libraries, file systems, hierarchical organizations such as business flowcharts, disk space management, genealogies, and classification systems. Because of the simplicity of understanding by the user, this type of visualization is the most represented in the world of visualization tools. Thus, there are greater diversity approaches than in other types of visualizations, among them, are presented:

- Conical trees like *Cat-a-cone* (Hearst & Karadi, 1997) and *LyberWorld* (Hemmje et al., 1994a).
- Hyperbolic trees such as *Star Tree*, *H3 3D Hyperbolic Browser* (Munzner, 1998) and *Walrus* (Hemmje et al., 1994b).

Multidimensional visualization: the information to be displayed in this type of visualization are usually information from relational databases where each element information item is represented according to its values in N-dimensional. Display systems will seek to reduce the number of dimensions to 2 or 3 for the purpose of facilitating interpretation and analysis by the end user. In this type of visualization, two approaches can be distinguished:

- The graphs of point cloud such as *Xgobi* (Swayne et al., 1998) *Envision* (Nowell et al., 1996), *SpotFire* (De Saussure, 1957) with the *FilmFinder* (Ahlberg & Shneiderman, 1994), *Miner3D* (Miner3D, 2014) and *LHN-FCSF* (Azar & Hassanien, 2014) .
- The Inselberg diagrams (or the system of polylines) as *Parallel Coordinates* (Inselberg & Dimsdale, 1991), *Attribute Explorer* (Tweedie et al., 1994).

Network Visualization: this sort of visualization concerns datasets such as networks of hypertext documents or people. These networks of objects are not necessarily in hierarchical order (co-quotations networks). For example, this allows navigation by hyperlinks from one document to a collection of other documents, or a part of document, or to a person. The visualization of networks should be able to visualize the most information possible in a single view to find the relationship between nodes and find interesting nodes.

In the following table, we present a comparative study of different data visualization techniques presented in this section:

2.2 EHR Visualisation System

A considerable number of research efforts have studied how end users interact with EHRs. The objective of these works is to help clinicians and researchers in the medical field in order to deepen the understanding of medical treatment and results. They are also interested to overcome complexity of the medical decision making based on the EHR. In this section, we present an overview of some of these works:

LifeLines (Plaisant et al., 1998) use a timeline visualization technique for representing personal histories, medical records and other types of biographical data. In LifeLines, horizontal bars are used to represent the time duration, also the location of the events that occur on a horizontal axis. Similar events are organized into facets, which can be developed and reduced to increase or decrease the level of detail. The color notations and line thickness are used to indicate the importance and the relations between events. To manage the regions with a high density of data. LifeLines provides a zoom feature that allows users to compress and to extend the time scale at any location. Additional content (e.g. multimedia) can be added.

Timelines (Harrison et al., 1994) is a temporal visualization system centered on the problem of patient records. The content of the EHR is integrated, reorganized and displayed in the User Interface (UI) along a timeline. Timelines is similar to Lifelines in the manner in which the various components of the EHR are grouped along the y-axis: Imaging, reports, lab tests, etc. However, unlike LifeLines (Plaisant et al., 1998), the TimeLines system uses a representation in the form of XML data to manage data from distributed and heterogeneous medical databases. The data elements that are displayed in the user interface are classified according to a knowledge base that guides the rules of inclusion and data visualization metaphors used to render the data.

LifeLine2 (Wang et al., 2008) is an extension of LifeLines, designed for simultaneous viewing of multiple projects. The system facilitates comparative visualization of dossiers by alignment means, filtering and sorting operations. By aligning patient records a common reference event (e.g., the first heart attack), a user scan identify concurrent and related events. In addition, filtering operations and classification complete the aligning by interactive reorganization into reducing overall records to respond to a change in orientation of the user. LifeLine2 used to view each EHR as a horizontal band on a time line. In each case, the events of the same type are placed in the same horizontal line. Event types are differentiated by color.

Similan (Wongsuphasawat & Shneiderman, 2009) is an interactive user interface that helps users to find similar records in a database that contains multiple files with the ability to customize the search parameters in order

Table 1 Summary table of the different types and techniques of data visualization

Type of visualization	Visualization technique	Advantages	Disadvantages
Linear visualization	Large spreadsheets	The user can have an overview, filter, detailing, rearrange into subgroups, sort the data and see the relationships between objects.	Cannot extract or export interest data once it identified.
	Perspective walls	There is time tracking events and you can get a better resolution on the periods and events of interest while keeping an overview.	Must already know where to look at the time area since it cannot filter except TimeWall that has integrated this interactive filtering feature.
Hierarchical visualization	The hyperbolic approach	This hyperbolic geometry allows better visualization nodes in the periphery relative to a radial diagram.	The untrained user must make significant efforts to reconstruct the cognitive context following the transformation "fisheye".
	The conical approach	3D view with transparency cones and each cone of interest is shown with the details of a node to keep the vision of the whole ("focus + context").	The cones being in front of each other, even with transparency, it is difficult to distinguish details.
Multidimensional visualization	The graphs of point cloud	It is possible to filter the information by dynamic queries.	The multiplicity of points on the map causing a difficulty in terms of the context.
	The Inselberg diagrams	The interest in this type of chart is immediate, because it allows to simultaneously visualizing large numbers of dimensions. It draws trends and correlations models.	This display mode is however not common and requires some familiarity. Must be played on all parameters to display relevant views.
Network Visualization	Social Networks	In general, it is very flexible in terms of interactivity except NetMap. The user can perform requests via concepts to find a person or a project or task to do.	In TheBrain, cannot go beyond two levels of hierarchy.
	Networks documents	You can view citing documents and cited documents to apprehend the importance of a scientific field and/or detect the main actors in this field.	Large co-quotations networks are quickly unreadable.

to deepen the understanding of the results. To search the similarity between folders, Simlian based on a measure called M & M (Match & Mismatch) which calculates the similarity scores. It adopts the concept of alignment LifeLines2 (Wang et al., 2008) and allows users to pretreat the data set by aligning the events in relation to a sentinel event. Similan displays all events for each record in a calendar. Each folder is stacked vertically on alternating background colors and identified by name on the left. The ranking scores appear to the left before the name. The events are displayed as colored squares on the timeline. By default, all records are presented by using the same time scale (with corresponding labels years or months displayed at the top) and the display is dimensioned so that the date range adapts on the screen.

LifeFlow (Wongsuphasawat et al., 2011) is a visual-interactive presentation of sequences of events. It will scale any number of records, summarizes all possible sequences and highlights the temporal spacing of the events within sequences. LifeFlow has a different way of viewing folders. In this system, all records are first grouped into a hierarchical structure called a sequence tree before being viewed. The raw data are represented by triangles colored in a horizontal line (using the traditional approach also used in LifeLines2). Each row represents a project. The records are grouped by sequence in a data structure called a sequence tree which is then converted into LifeFlow visualization. The conversion is to represent each node in the tree by an event bar color-coded corresponding to the color of the event type. The height of the bar is determined by the number of projects in this node in proportion to the total number of projects.

VISITORS (Klimov et al., 2010) is an intelligent tool for processing large amounts of data from multiple patients focused on the time from many sources for the purpose of analyzing the results of clinical trials and evaluation of treatment quality. It includes recovery tools, visualization, exploration and analysis of raw data based on time and abstract multiple patient records concepts. Thus, this system was based on an interactive exploration module based on ontology, which allows the user to visualize the raw data and abstract concepts for multiple patient records at various levels of temporal granularity, to explore these concepts and display the associations between the raw and abstract concepts. A function for delegate based on the knowledge is used to convert multiple data points in a single delegate value representing each time granule. To select the patient population to explore, the VISITORS system includes an ontology-based temporal-aggregation specification and an expression and graphical specification language Module. The expressions, applied by an external time mediator, retrieve a list of patients, a list of relevant time intervals and a list of datasets focused on patients over time, by using an expressive set of time constraints and value.

CLEF (Hallett, 2008) is an architecture of visualization to browse medical records, which includes visual navigation tools and automatically generated text summaries. The visual browser shows an overview of the patient's medical history by tracing events along three parallel lines of time, corresponding to diagnoses, treatments and investigations. The events of the medical history of a patient are represented as graphical objects visually differentiated by color and icons. In addition, zoom (front and rear) on the time scale that allows the user to obtain more or less detailed views on the events in this time, the browser provides interactive visualization of semantic relations between events (e.g. caused by, indicated by, etc..). It also allows the user to view numerical data (e.g. blood tests) by plotting the results of measurements on separate cards.

ICOP (Sassi et al., 2009) is a system to visualize information coming from different sources in a unified and iconic temporal representation. The main component of this system is the process of filtering based on access rights and user needs. The system also allows to visualize the various documents which have been created locally or not. ICOP offers a new visualization technique that is an iconic and timed visualization. For the visualization technique, ICOP is based on the creation of a model of graphic visualization with four main components: the icon, the context, timing and metadata. The goal of ICOP is to trace information relating to a particular field on a Iconic Information Card (I^2C). The latter represents a schematic history of the project of any subject on the basis of an iconic representation of events and a time axis where these events are placed (Sassi, 2009).

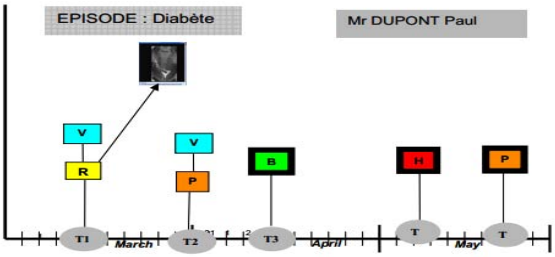

Fig. 2 Iconic Informations Card(I^2C)

On the basis of this of the literature review, we present a comparative study which is essentially based on three criteria of classification that we have deduced: the technique of visualization, display mode and the nature of temporality.

- *The visualization technique*: that distinguishes the work according to the visualization technique used to display the EHR. According to this criterion there are two visualization techniques:

  - Simple Graphical display: which consists of a graphical representation based on simple images and graphics (curves, histograms).
  - Iconic Visualization: that consists of a graphical representation based mainly on the notion of the icon.

- *The display mode*: this criterion is associated with number of projects displayed by the system. According to this criterion there are two classes:
  - The mono-project system: that manages and displays one project in the interface.
  - The multi-project system: which supports simultaneous visualization of multiple projects.

- *The temporality*: This criterion is associated with the nature of the time axis used in visualization projects. In fact, there are two types of temporality:
  - Continuous temporality: that annotates the various events of the project by their exact date, thus forming a historical project bounded by a start date and an ending date.
  - Non-continuous temporality: in this type of temporality, the time axis is divided into equal and regular time intervals. In some systems, these intervals are automatically calculated by the size of the application window and the full time data range. Thus, in each interval, the events are grouped and placed in the same order.

Table 2 Summary table of the HER visualization system

System	The Vis. technique		The display mode		The temporality	
	Simple	Iconic	Mono-project	Multi-projects	Continuous	Non-continuous
LifeLines	*		*			*
Timelines	*		*			*
LifesLine2	*			*	*	
Similan	*			*	*	
LifeFlow	*			*	*	
Visiteurs	*			*		*
CLEF		*	*		*	
ICOP		*	*			*

It is certain that visualization of records and analysis of treatment outcomes for decision-making is often the central objective that systems already implemented trying to achieve it. It is also clear these systems are characterized by the differences between them at some points and similarities in others. From the perspective of criticism, we identify the following findings:

- The majority of systems are designed to display only one project or several patients, but not both.
- Most systems use a continuous or non-continuous temporal axis, but not both
- Only systems that manage multiple folders of patients have good support for filter
- Some specific users intentions are rarely taken into account
- Most systems use a simple graphic language or iconic language, but not both
- No system allows to combine:
  - Treatment mono-project with the multi-project
  - Simple graphic visualization with iconic counterpart
  - The continuous temporality with the non-continuous temporality

Taking into account these points, we implement a system that combines both the simple graphic language with its iconic counterpart, mono-project visualization with multi-project visualization and the two type of temporality (continuous and non-continuous). To ensure our perception, we choose the ICOP system (Sassi et al., 2009) to be the starting point and the cornerstone of our work.

3 The Medical Multi-project System

Our system, *Medical Multi-project ICOP* (M^2ICOP) is an interactive health information system especially dedicated to professionals and researchers in the field of health. It consists of an extension of the ICOP system (Sassi et al., 2009), who plays the role of a semantic mediator which ensures communication and interoperability between different information systems to present these data and visualize them in a unified, graphic, iconic, chronological format. M^2ICOP is designed for simultaneous visualization of multiple projects (EHRs).

In fact, it helps users to find similar records in a database that contains multiple files with the ability to customize the search parameters. It allows the user to deepen the understanding of the results, compare two health records of two different patients on the same temporal axis and take advantage of best practices and shared experiences. In addition to its main function that is visualization multi-project, M^2ICOP facilitates the monitoring and monitoring of care by taking the management of chronic diseases.

3.1 Architecture of M²ICOP System

The architecture of our system is a four-tier architecture:

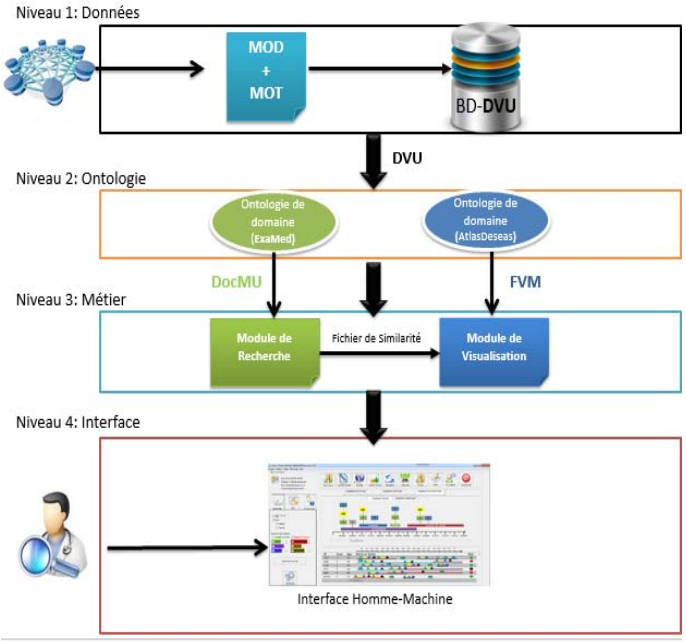

Fig. 3 Architecture of M²ICOP system

Level 1, **Data**: in which data from multiple sources of information are annotated, described, unified and contextualized according to two meta-ontologies (a Meta-Ontology for Domain "MOD" and Meta-Ontology for Task "MOT"). They solve the problem of heterogeneity and ensure semantic interoperability. A Unified Virtual Folder (UVF), for each patient, was built by the concepts already processed by the MOD and the MOT. All UVF created are stored in a relational database. In this database, each record has an identifier, and a set of metadata. In fact, these metadata consist of a summary of the knowledge extracted from the file itself. The main objective of this metadata is to facilitate research, especially in large amounts of data such as our case.

Level 2, **Ontology**: which includes two domain ontologies specifically designed to meet some features of our system. The first ontology is the ontology ExaMed which provides a classification of all medical examinations and tests

that a patient can do. The second ontology is the ontology AtlasDeseas which represents all diseases, classified by atlas of human body structure.

Level 3, **Art**: it is the heart of our system level. It consists of two main modules which are the research module and display module. The search module is devoted primarily to research similar records, while the visualization module is occupied by the objective of creating different types of visualization.

Level 4, **Interface**: this level is attached to the Human-Computer Interaction part of our system. It consists of all the graphical interfaces that allow the user to interact and communicate with the system.

More generally, it may deduct that the architecture of our system is characterized by three aspects:

1. The first aspect is the *representation of knowledge* that is associated with the first two levels; Data and ontology.
2. The second aspect is the *search for knowledge* that is associated with the Search module of the art-level.
3. And finally, the last aspect is the *visualization of knowledge* which is associated with the Display module of art-level and interface-level.

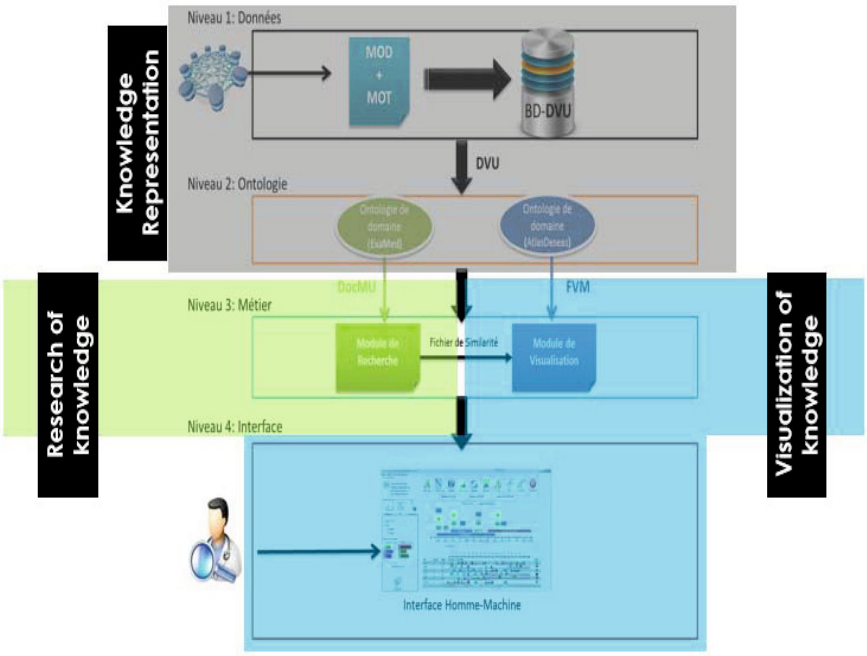

Fig. 4 The three aspects of M^{2}ICOP architecture

3.1.1 The Representation of Knowledge

For the representation of knowledge, M^{2}ICOP is thorough on a powerful semantic base formed by a Meta-ontology of Domain (MOD), a Meta-ontology of Task (MOT) and two domain ontologies (Examed and AtlasDeaseas):

The Meta-Ontology of Domain (MOD): defines the various concepts and semantic relations between concepts and was constructed from a set of existing data files. According to (Sassi, 2009), the MOD "provides a framework for reuse of meta-models in order to allow the specification of domain knowledge through concepts and relationships between these concepts. It allows the description and classification of concepts, properties, value domains and instances. "

The Meta-Ontologie of tasks (MOT): offers the user a framework description of the vocabulary concerning a generic task. In addition to the objective of defining the tasks of the field, it allows to contextualize the concepts derived from the MOD and this by specifying the environment of the task (its venue, the user and the time) " (Sassi, 2009). The MOT connects each profile of a user class that is well determined to a set of activities or tasks that are related to an ontological subset of domain.

These two Meta-ontology are designed to ensure semantic interoperability. In fact, data from heterogeneous data sources are annotated, described, unified and contextualized according to these two meta-ontologies to solve the problem of heterogeneity. For each patient, a Unified Virtual Record (UVR) is built by the concepts produced by the MOD and the MOT (Sassi, 2009). All UVR are stored in a database (BUVR).

ExaMed (Dridi, 2014) is a domain ontology that represents all concepts relating to tests and medical examinations. A medical test refers to a type of medical procedure that helps to detect and assess disease, deals with an individual's vulnerability and contributes to an individual as well as its care providers to determine the right treatment path (Al-Gwaiz & Babay, 2007). In fact, these medical examinations or tests can be classified according to several types. Generally, there are two major types: clinical and paraclinical (or supplementary) examination, this last type of examinations which in turn can be divided into different classes (biological examinations, examination of medical imaging, endoscopy, biopsy...). Thus, each medical test has parameters, upon which it is based. (Murthy & Halperin, 1995)

AtlasDeseas (Dridi, 2014) is an ontology of domain that covers all human diseases. It provides a classification of diseases according to the criterion of being a normal or chronic disease, and ensures the combination of Anatomy and pathology by associating each organ of the human body to its diseases. Scientifically, the human body represents the structure physical and material of human. It consists of six main parts consisting of the atlas of the body (head and neck; Skin, nail and hair; Thorax; Upper limbs; Abdomen and pelvis; Lower limbs), each part of atlas is composed of all limbs and organs (Bianconi et al., 2013). These members are vulnerable to a

variety of different diseases which can be classified into: normal and chronic diseases (Navarro-Alarcon & López-Martınez, 2000).

3.1.2 The Search for Knowledge

The Search Module (level 3) allows the search of the most possible relevant projects compared to a search query entered by the user by calculating the similarity between different projects. Practically, when the user of the system makes a request and launches the search operation, the system explores the basis of the UVR to find the UVR files relevant to the request. Great importance is also attributed to the metadata of the UVRs stored in the database due to their active roles to accelerate the research process. At this stage, a list of UVR files is selected. Thereafter, each UVR is transformed into a file Semantic Mediation (XMS) through an XMS generator and by taking into consideration the user profile. This file contains the metadata detailing the object context, the information on the source document associated with the

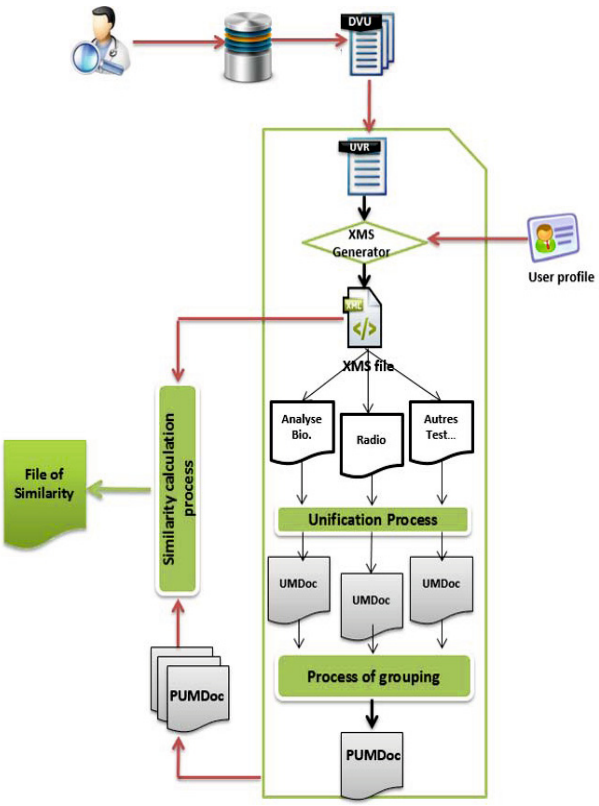

Fig. 5 Architecture of the Search Module

object and its url. Based on these Urls, this module will be able to retrieve records of medical examinations of each patient. Each recovered document will be at the entrance of a unification process that aims to reformulate the model of document. The result of this process is a set of files Unified Medical Document (UMDoc).

The result of this process is a set of Unified Medical Document (UMDoc) files associated with a patient, which represents the concepts of medical documents in a structured way and which are all with a unique model and a uniform structure. These UMDocs are subsequently passed to a second process called 'the grouping process' which in turn make together those in a new file called the Document Patient Unified Medical Document (PUMDoc) file. So this file is the concatenation of different UMDocs of the same patient and used for comparison with other PUMDocs other patients.

The resulting UMDocP with the corresponding XMS files are passed to the last process in this module, the process of calculating similarity which aims to calculate the similarity between the different projects. The result of calculation will be in the form of a file called File of Similarity (FoS). The similarity calculation process includes two steps of calculation:

- Basic calculation (default): where the role of comparator is to count the common medical episodes (diseases which present in the XMS files) between the two files to compare.
- Advanced calculation: the role of comparator in this level is not calculate diseases Commons, hands farther than that, he is interested in the tests and medical examinations and their parameters (values) that are in the file PUMDoc.

The result of this comparator is depicted in a file called *the file of similarity* in the form of a table, whose columns present medical episodes (diseases) relating to a patient given (in the case of basic calculation) and parameters medical tests in case of advanced calculation. And the lines present a list of patients (patient identifiers) that must be compared with the patient target.

3.1.3 The Visualization of Knowledge

The Visualization module (Level 3) of our architecture allows the visualization of data from heterogeneous sources and systems that are collected and described in the format of a UVR and also the resulting information of the Research Module. This new visual manner to serve the user in terms of understanding, cognition and interaction. This module included four sub-modules,

which each of them offers a particular mode of visualization. In fact, there are a Mono-project visualization mode, a multi-project mode, a bi-project mode and a mode for a metaphorical visualization.

The Mono-Project visualization: this is the basic display mode inherited from the ICOP system. It shows the simplest case, when the user want to view a medical record of a single patient. The sub-module on this method takes as input only one XMS file to generate an Iconic Information Card (I^2C) of a single patient. In this mode of visualization, the EHR is displayed in the form of an iconic information card which allows of represents a synoptic history of the patient from his birth until his death.

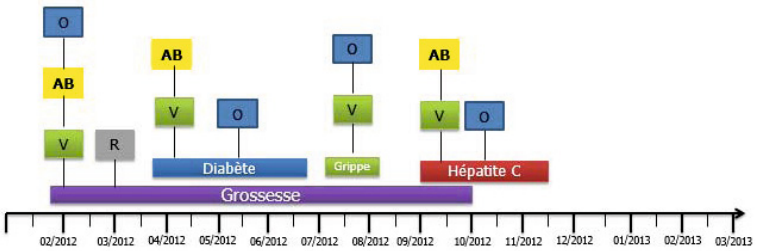

Fig. 6 The Mono-Project visualization

As the figure shows, the episodes are represented with rectangles with a start date and an ending date. Each episode is annotated by medical acts represented in the form of icons. Click on an icon displays the metadata of the medical act in the form of a brief description of the icon. If a document is assigned to the icon, a small diagram appears on the icon indicating the presence of a descriptive document of the medical act. The technique of visualization used in this mode is purely iconic, and the temporality used is of a continuous nature.

The Bi-Project visualization: allows the user to visualize two health records of two distinct patients, which thus gives him the opportunity to compare these two files. The sub-module of the display mode takes as input two XMS files and gives as a result a Card of Iconic Information with two projects.

In this mode of visualization, the two projects are displayed one above the other and are separated by a time axis which makes the task of comparison easier and more efficient.

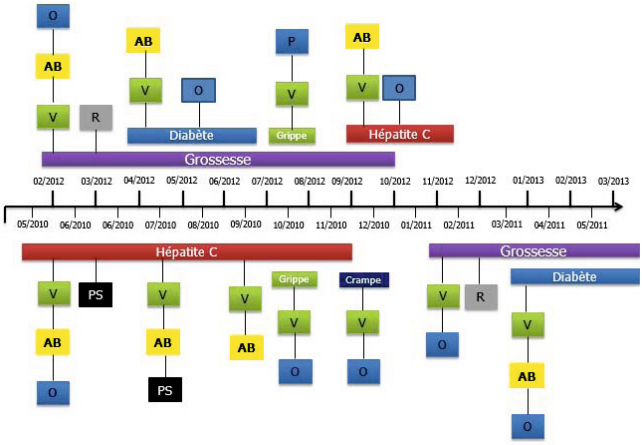

Fig. 7 The Bi-Projects visualization

The Multi-projects visualization: follow-up to an operation to search for similar projects, the visualization multi-project sub-module intervenes and display the result of research in the form of a list. This mode of visualization presents to the user a synthetic overview on similar to a health record given folders. In this mode, the projects are displayed in a table, or each line consists of a medical record. Each folder is stacked horizontally on background colors alternating and identified by its ID, the sex of the patient and his age in the left and his health situation (healed, still sick, dead) in the right. Here, the episodes are represented by horizontal lines arranged depending on their time. Thus, the various medical acts are represented by triangles, the "+" sign indicates that there is more of medical acts but are hidden. A passage from mouse on an episode or a medical procedure displays a brief description (the metadata of the object). In this mode, the technique of vi-

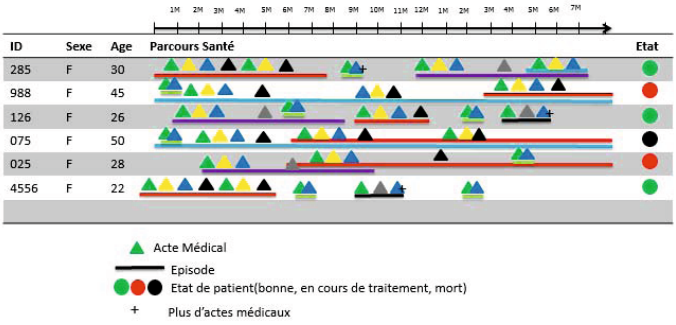

Fig. 8 The Multi-Projects visualization

sualization used is Simple Graphical Visualization, and the temporality used is non-continuous.

The Metaphorical visualization: this mode of visualization differs from those preceding in terms of operating principle, stakeholders components and the result obtained. It allows to display on a map of a human body different diseases of a patient, which allows the user to have an overview of the patient profile of more quickly way and more expressive. The patient's medical record in this mode is displayed on a map of the human body, or episodes are placed on the human atlas parts (the organs) infected, and represented by stars of two colors to make the differentiation of the type of diseases (red for chronic diseases and green for normal conditions). In addition to its graphical representation on the map of the human body, the episodes are represented by icons with their names. Clicking on one of these icons displays an iconic and chronological presentation of the episode. Other information about the patient can be presented in this mode visualizations such as length, weight and age.

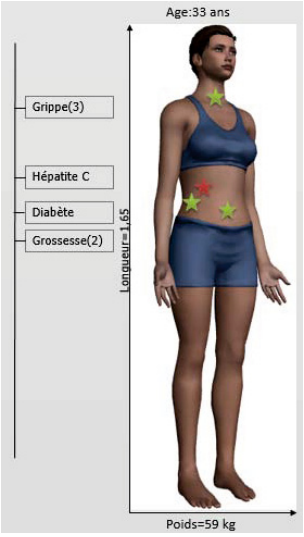

Fig. 9 The Metaphorical visualization

4 Experimentation

Practically, the prototype M^{2}ICOP is a virtual office for health professionals. When connecting, the doctor can access to various features offered by a virtual office as messaging, shared calendar, organizing contacts, backup shared documents and personal documents, notepad, tasks, favorites and

synchronous communication tools (chat and forum) and finally, the health record of the patient.

In fact, the interface of the M^{2}ICOP system is composed of four main panels:

1. The digital user badge: provides information about the person connected to the system.
2. Menu of features: consists of a set of tools that can be useful to the user (e-mail, calendar, sharing documents, Favorites, etc.).
3. The display of the projects Panel: present the space of which the Iconic Information Card displays.
4. Space management and research: this space is in direct relation with the display of the projects Panel. It consists of three tabs:
   - the General information tab: that presents the informations of patient(s)
   - the Filter tab: which offers the user the possibility to visualize the EHR of a patient according to different views of visualization:
     - View by episode: displays all medical procedures on a well-defined episode. In other words, it is a view that includes the objects annotating an episode.
     - View by the medical act: this view corresponds to a display by object type. It summarizes the medical procedures of same type annotating episode (s).
     - View by time interval: the view by time interval allows to visualize all the episodes and medical procedures in a given time interval.

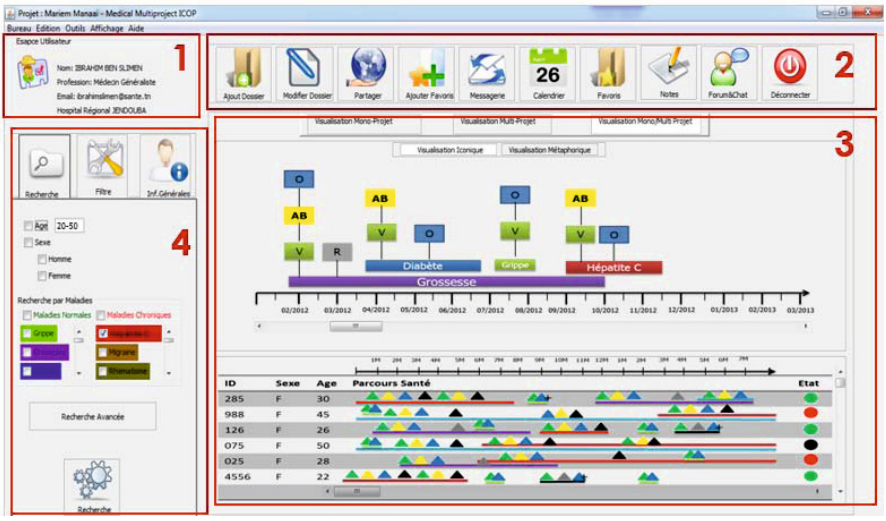

Fig. 10 The virtual desktop of M^{2}ICOP

- the Search tab: allows the user to search similar to his patient health records, thus giving it the ability to specify search parameters:
  - Default search: the search criteria are the patient's age, sex, and its diseases.
  - Advanced search: the search criteria are mainly the parameters of medical tests and examinations.

5 Evaluation of the Fonctionality and Usability of the M²ICOP System

The evaluation of health information systems is a challenge: the evaluation of a system in itself is complicated, but in the medical field, we must also focus on the technical and human aspects as well as the impact of these new systems on the practices of health professionals (Ammenwerth et al., 2003, Lemmetty & Häyrinen, 2005). Other authors, such as (Friedman et al., 1997, Brender, 1998), assert that assessments of medical systems are hardly generalizable because it must take into account the specificity of the tested object, its context of use still very specific in the health field and the context of the evaluations themselves. We designed and developed the system M²ICOP on the basis of the requirements of health care professionals. We conducted an evaluation of the system in order to answer the following three research questions:

- Overall functionality and usability: are interactive exploration of system operators they possible, functional and usable?
- Visualization functionality:
  - Confirm the usefulness of adopting an iconic representation and temporal visualization of the Iconic Card Information:
    - Is the adoption of the iconic representation a useful practice?
    - The adoption of the temporal representation via a timeline it is a useful practice?
  - Measure the usability of these Iconic features:
    - The functionality of different viewing modes (mono-project, bi-project, multi-project and metaphorical) are all used by health professionals?
    - Is the functionality of different visualization modes useful practice?
- Search functionality and comparison of similar files:
  - Confirm the utility and usability of the functionality of the search:
    - Are the results of research are relevant to users expectations?
    - Is the result display effective and useful research?
  - Measure the usability features comparison:

- Is the functionality comparison a useful practice?
- Do the comparison criteria you have defined seem useful?

5.1 Data Collection and Evaluation Process

In this work, we decided to confront the health professionals to this prototype to validate our assumptions about the utility and usability of its functionality for clinical practice and shoot the first recommendations on the usability functionality. To evaluate our system, we used a corpus that contains medical records of 37 patients from the regional hospital of Jendouba and private clinics. We conducted an evaluation with eight clinicians at different levels of medical training and different specialties. The participants were asked to answer fifteen clinical questions: five questions required the use of operators of general Exploration of M^{2}ICOP, five questions required to use functionalities of project comparison (selection criteria, classification, ...) and five questions required to use visualization features by mode (mono-project, bi-project, multi-project and metaphorical).

A preliminary step in the evaluation that allows to introduce participants to our system. We used the projection device on the big screen with a PowerPoint presentation. The objective of this presentation is to provide all participants with the knowledge necessary to make judgments of the usefulness of services offered by the system M^{2}ICOP, objectives and concrete nature of our research. This step also allows to familiarize participants with our system and with the concepts used in the evaluation, order to allow them to focus more on testing the usefulness of M^{2}ICOP. Following this demonstration, we asked each participant to answer three clinical questions in order to test their understanding and mastery of the various system functions. When the participant answers the questions correctly, so he can be considered as being ready to conduct the assessment.

5.2 Evaluation Results

This section summarizes the results of the evaluation in terms of the three research questions outlined in Section 4. The results of evaluations obtained are rather encouraging. Clinicians have found all the features useful and it's very easy to handle the interface. The main criticisms collected focused on the lack of the vocal aspect interfaces. Health care professionals would find it very relevant to integrate the notion of speech in interfaces to make them save time.

Although 70% of clinicians said they have not used this type of tools, they found the general use of the application rather very simple to 80%. Only 10% of clinicians said they had really search features (the others have found more easily) and only 20% of clinicians have confessed to handling errors. Manipulation, iconic visualization and comparison of projects as a whole seems to be

rather very easy to understand. What was most encouraging is that clinicians have all said that this prototype helps immensely to develop plans for the proper care and communicate more easily with other professionals of health through the functionality of M^{2}ICOP.

6 Conclusion

Arriving at the era of massive data, it is no question that this promising technology with its potentialities will make a leap forward in terms of capture, storage, visualization and analysis, and that it will really add value and importance to these data. In the field of health care, our application domain, Big Data has faced many challenges. It aims to strengthen and improve the inference of knowledge from sources of patients complex and heterogeneous, take advantage of the large volume of data and improve the quality of care and treatment, etc. In this context, we proposed a new interactive information system that we called the Medical Multi-project ICOP system. Our system allows clinicians and researchers in the medical field to manage and visualize massive medical data, which they contained in the electronic health records in four different display modes: single project, bi-project, multi-project and metaphorical. Thus, this system offers the user the opportunity to search for similar files, and make the comparison between two different projects. The originality of M^{2}ICOP mainly consists of the combination of mono-project and multi-project mode, simple graphical representation and purely iconic counterpart, and the use of the continuous temporality with non-continuous temporality.

Although the results are encouraging, several perspectives for future work may be considered. It may be interesting to incorporate new features in the GUI (e.g., graphic zoom). We also think about adapting our system in a mobile and distributed environment (e.g., the use of our system with a mobile device like PDA (Personal Digital Assistant)). Thus, it is interesting to develop and expand the application areas of our system, such as the adaptation of the latter in the legal field to view, manage and compare litigation.

References

Ahlberg, C., Shneiderman, B.: Visual information seeking using the filmfinder. In: Conference Companion on Human Factors in Computing Systems, CHI 1994, pp. 433–434. ACM, New York (1994)

Al-Gwaiz, L.A., Babay, H.H.: The diagnostic value of absolute neutrophil count, band count and morphological changes of neutrophils in predicting bacterial infections. Med. Princ. Pract. 16(5), 344–347 (2007)

Ammenwerth, E., Iller, C., Mansmann, U.: Can evaluation studies benefit from triangulation? a case study. International Journal of Medical Informatics 70(2), 237–248 (2003)

Azar, A.T., Hassanien, A.E.: Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Computing, 1–13 (2014)

Beyer, M.: Gartner says solving 'big data' challenge involves more than just managing volumes of data (2011), http://www.gartner.com/newsroom/id/1731916 (accessed: April 14, 2014)

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., et al.: An estimation of the number of cells in the human body. Annals of Human Biology 40(6), 463–471 (2013)

Brender, J.: Trends in assessment of it -based solutions in healthcare and recommendations for the future. International Journal of Medical Informatics 315(7109), 217–227 (1998)

Brian, H., Boris, E.: Expand your digital horizon with big data. Forrester Research Inc. (2011), http://www.asterdata.com/newsletter-images/30-04-2012/resources/Forrester_Expand_Your_Digital_Horiz.pdf

CeArley, D., Claunch, C.: Top 10 strategic technology trends for 2012 (2012), http://www.gartner.com/technology/research/top-10-technology-trends (accessed: April 14, 2014)

Cleveland, W.S.: Visualizing Data. Hobart Press, Summit (1993)

De Saussure, F.: Cours de linguistique générale (1908-1909). Cahiers Ferdinand de Saussure (15), 3–103 (1957)

Dridi, A.: Le systéme medical multi-project icop(m^2icop). Master's thesis, Faculty of Law, Economics and Management of Jendouba, University of Jendouba, Tunisia (2014)

Eick, S.G., Steffen, J.L., Sumner Jr., E.E.: Seesoft-a tool for visualizing line oriented software statistics. IEEE Trans. Softw. Eng. 18(11), 957–968 (1992)

Fernanda, V., Martin, W.: How to make data look sexy (2011), http://articles.cnn.com/2011-04-19/opinion/sexy.data_1_visualization-21st-century-engagement?_s=PM:OPINION (accessed: April 14, 2014)

Freeman, E.T.: The Lifestreams Software Architecture. PhD thesis, New Haven, CT, USA (1997) UMI Order No. GAX97-33943

Friedman, C.P., Wyatt, J.C., Faughnan, J.: Evaluation methods in medical informatics. BMJ-British Medical Journal-International Edition 52(7109), 689 (1997)

Friendly, M.: Milestones in the history of thematic cartography, statistical graphics, and data visualization. In: Proceedings of the 13th International Conference on Database and Expert Systems Applications (Dexa 2002), Aix En Provence, pp. 59–66. Press (2008)

Gunter, T.D., Terry, N.P.: The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions. Journal of Medical Internet Research 7(1), e3 (2005)

Hallett, C.: Multi-modal presentation of medical histories. In: Proceedings of the 13th International Conference on Intelligent User Interfaces, IUI 2008, pp. 80–89 (2008)

Halper, F.: Ibm what is big data? bringing big data to the enterprise (January 2012), http://www-01.ibm.com/software/in/data/bigdata/ (accessed: April 14, 2014)

Harrison, B.L., Owen, R., Baecker, R.M.: Timelines: an interactive system for the collection and visualization of temporal data. In: Graphics Interface 1994, pp. 141–141. Citeseer (1994)

Hearst, M.A., Karadi, C.: Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. ACM SIGIR Forum 31, 246–255 (1997)

Hemmje, M., Kunkel, C., Willett, A.: Lyberworld–a visualization user interface supporting fulltext retrieval. In: SIGIR 1994, pp. 249–259. Springer (1994a)

Hemmje, M., Kunkel, C., Willett, A.: Lyberworld–a visualization user interface supporting fulltext retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994, pp. 249–259. Springer-Verlag New York, Inc., New York (1994b)

Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. Science 332(6025), 60–65 (2011)

IBM. Harness your data resources in healthcare (2011), http://www-01.ibm.com/software/data/bigdata/industry-healthcare.html (accessed: April 14, 2014)

Inselberg, A., Dimsdale, B.: Parallel coordinates. In: Human-Machine Interactive Systems, pp. 199–233. Springer (1991)

Klimov, D., Shahar, Y., Taieb-Maimon, M.: Intelligent selection and retrieval of multiple time-oriented records. Journal of Intelligent Information Systems 35(2), 261–300 (2010)

Laney, D.: 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group (February 2001)

Lemmetty, K., Häyrinen, E.: Operation management system evaluation in the central finland health care district. In: Connecting Medical Informatics and Bio-Informatics, vol. 116, pp. 605–607. IOS Press (2005)

Livny, M., Ramakrishnan, R., Beyer, K., Chen, G., Donjerkovic, D., Lawande, S., Myllymaki, J., Wenger, K.: Devise: integrated querying and visual exploration of large datasets. ACM SIGMOD Record 26, 301–312 (1997)

Mackinlay, J.D., Robertson, G.G., Card, S.K.: The perspective wall: Detail and context smoothly integrated. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 173–176. ACM (1991)

Miner3D. Computer software (2014), http://www.miner3d.com/ (accessed: April 14, 2014)

Munzner, T.: Exploring large graphs in 3d hyperbolic space. IEEE Computer Graphics and Applications 18(4), 18–23 (1998)

Murthy, L., Halperin, W.: Medical screening and biological monitoring: A guide to the literature for physicians. Journal of Occupational and Environmental Medicine 37(2), 170–184 (1995)

Navarro-Alarcon, M., López-Martınez, M.C.: Essentiality of selenium in the human body: relationship with different diseases. Science of the Total Environment 249(1), 347–371 (2000)

Nowell, L.T., France, R.K., Hix, D., Heath, L.S., Fox, E.A.: Visualizing search results: some alternatives to query-document similarity. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 67–75. ACM (1996)

Owen, G.S.: History of visualization (1999), http://www.siggraph.org/education/materials/HyperVis/visgoals/visgoal3.htm (accessed: April 14, 2014)

Pierre, D., Marc, M.: Big data: du concept à la mise en œuvre. premiers bilans (2013), http://blog.dataraxy.com/public/TR4_Big_data.pdf (accessed: April 14, 2014)

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., Shneiderman, B.: Lifelines: using visualization to enhance navigation and analysis of patient records. In: Proceedings of the AMIA Symposium, p. 76. American Medical Informatics Association (1998)

Rao, R., Card, S.K.: The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1994, pp. 318–322. ACM, New York (1994)

Rometty, V.: Big data: la nouvelle révolution. La Tribune 42, 4 (2013)

Sassi, S.: Le système ICOP: représentation, visualisation et communication de l'information à partir d'une représentation iconique des données. Phd thesis, INSA de Lyon et ENSI de Manouba (2009)

Sassi, S., Verdier, C., Flory, A.: Collaborative tasks: reorganize the information representation and communication in the project management. Electronic Journal of Digital Enterprise 25, 10 (2009)

Snijders, C.C.P., Matzat, U., Reips, U.D.: "big data": Big gaps of knowledge in the field of internet science. International Journal of Internet Science 7, 1–5 (2012)

Stefan, S.: Les 3 v du big data: Volume, vitesse et varéité (2012), http://www.journaldunet.com/solutions/expert/51696/les-3-v-du-big-data---volume--vitesse-et-variete.shtml (accessed: April 14, 2014)

Swayne, D.F., Cook, D., Buja, A.: Xgobi: Interactive dynamic data visualization in the x window system. Journal of Computational and Graphical Statistics 7(1), 113–130 (1998)

Tweedie, L., Spence, B., Williams, D., Bhogal, R.: The attribute explorer. In: Conference Companion on Human Factors in Computing Systems, CHI 1994, pp. 435–436 (1994)

Vidal, S.: Visualisation de l'information - un panorama d'outils et de méthodes. Technical report, INIST-CNRS - Institute for scientific and technical information, Vandoeuvre-lès-Nancy, France (2006)

Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B.: Aligning temporal data by sentinel events: Discovering patterns in electronic health records. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 457–466. ACM, New York (2008)

Wongsuphasawat, K., Gómez, J.A.G., Plaisant, C., Wang, T., Taieb-Maimon, M., Shneiderman, B.: Lifeflow: Visualizing an overview of event sequences (video preview). In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2011, pp. 507–510 (2011)

Wongsuphasawat, K., Shneiderman, B.: Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In: IEEE Symposium on Visual Analytics Science and Technology, VAST 2009., pp. 27–34. IEEE (2009)

Modified Soft Rough Set Based ECG Signal Classification for Cardiac Arrhythmias

S. Senthil Kumar and H. Hannah Inbarani

Abstract. The objective of the present study is ECG signal classification for cardiac arrhythmias. Most of the pattern reorganization techniques involve significantly large amounts of computation and processing time for extracting the features and classification. Electrocardiogram (ECG) is the P, QRS, T wave demonstrating the electrical activity of the heart. Electrocardiogram is the most straightforwardly accessible bioelectric signal that provides the doctors with reasonably accurate data regarding the patient's heart disorder. Many of the cardiac problems are visible as distortions in the electrocardiogram (ECG). Different heart diseases are with different ECG wave shapes; in addition, there is large numbers of heart illnesses, so it is hard to accurately extract cardiology features from diverse ECG wave forms. Big Data is now rapidly expanding in all science and engineering domains, including physical, bio-medical and social sciences. It is used to build computational models directly from large ECG data sets. Rough set rule generation is specifically designed to extract human understandable decision rules from nominal data. Soft rough set theory is a new mathematical tool to deal uncertainty. Five types of rhythm including Normal Sinus Rhythm (NSR), Premature Ventricular Contraction (PVC), Left Bundle Branch Block (LBBB), Right Bundle Branch Block (RBBB) and Paced Rhythm (PR) are obtained from the MIT-BIH arrhythmia database. Five morphological features are extracted from each beat after the preprocessing of the selected records. In this chapter, the ECG signals were classified using Modified soft rough set technique. The empirical analysis shows that the proposed method shows better performance compared to the other six established techniques like Back Propagation Neural Network, Decision table, J48, JRip, Multilayer Perceptron and Naive Bayes. This chapter is focused on finding an easy but reliable features and best MSR structure to correctly classify five different cardiac conditions.

Keywords: ECG signal, Feature extraction, Soft rough set, Classification, Comparative analysis.

S. Senthil Kumar · H. Hannah Inbarani
Department of Computer Science, Periyar University, Salem-636011
e-mail: {pkssenthilmca,hhinba}@gmail.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_16

1 Introduction

The electrical activity of the cardiac muscle and its connection with the body surface potentials describes the electrocardiogram (ECG). In recent years, study has been shown towards the generation of imitation of ECG signals to simplify the testing of signal processing algorithms. Physicians understand the morphology of the ECG sine waves and decide whether the heartbeat goes to the normal sine waves or to the class of cardiac arrhythmias. The heart sound using ECG will be recorded simultaneously from patients (Ravindra et al. 2013). Signal processing is today found in virtually any system for ECG analysis, and has clearly demonstrated its importance for achieving improved diagnosis of a wide variety of cardiac arrhythmias. Signal processing is employed to deal with diverse issues in ECG analysis such as beat detection, noise reduction, signal separation, feature extraction and classification.

Health care professionals and patients are generating huge amounts of data from an array of devices such as ECG signal sequencing machines, ECG signal records that monitor patient health. Big Data does not arise out of a vacuum: it is recorded from some data producing source. For example, consider our ability to intellect and observe the world around us, from the heart rate (Azar et al. 2013; Inbarani et al. 2014b) of cardiac arrhythmias.

Health care monitoring systems are generating loosely structured data from different sensors that are connected to the patient over a period of time. And these are large complex systems requiring efficient algorithms to process these raw data and require huge computational power. Big data (Sam et al. 2012) refers to the data generated from different sensors which includes medical data (Azar 2014; Azar and Hassanien 2014; Inbarani et al. 2014b), traffic and social data.

For developing an infrastructure for big data analysis, it requires a mechanism on how to acquire the data, organize these data and process it to extract meaningful information (Jeffrey et al. 2008). This can be represented as data acquisition, and data pre-processing (Inbarani et al. 2013).

The signal is searched for the sequences of regular heart rhythms. The average length of the heart rhythm is calculated at first and then for all heart rhythms the nearest from left and right are found. The morphological features taken for classification of ECG signals are P, QRS Complex and T.

The extracted morphological features are classified into five different classes. The signal decision classes are normal sinus rhythm (NSR) and various cardiac arrhythmias including Left Bundle Branch Block (LBBB), premature ventricular contraction (PVC), Right Bundle Branch Block (RBBB), and Paced Rhythm (PR).

This research is mainly aimed to classify the extracted ECG signal through the modified soft rough set technique. The data will be used as an input to the classifier to identify the cardiac arrhythmias. Modified soft rough set is the most suitable method for cardiac arrhythmia classification. Rough set has strong ability in data processing and can extract useful rules from ECG signal dataset. A decision rule is a function which maps an observation to an appropriate action. Decision rules (Khoo et al., 1999) play an important role in the theory of Rough set. The lower approximation is a description of the domain objects which are known with

certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset. Deterministic rules correspond to the lower approximation and non-Deterministic rules correspond to the upper approximation. In these situations, MSR-sets can help us to find approximations of subsets. Our proposed work consists of two parts: Initially in the pre-processing stage, signal noise data are removed using low pass and high pass filters and features are extracted from ECG signal. In the second part, decision rules are generated from the ECG signal data set.

1.1 ECG Wave Form Description

Figure 1 shows the basic structure of ECG Signal. The electrocardiogram is a reading of the electrical signals of the heart. Each part of the waveform corresponds to a different action within the heart as it cycles through its stages to push blood efficiently. Every major point in the waveform is given a letter {P Q R S T}. The first increase in the wave's amplitude causes the atria to contract and is referred to as the P-wave. The QRS complex then monitors the P-wave after a small suspension, which causes the ventricles to bond, expelling the blood from the heart to the lungs and repose of the body. The T-wave is the last part of the cycle and its purpose is to return the ventricles to a peaceful state. Then there is a pause with very little activity and then the cycle repeats itself. The ECG wave is formed as a projection of summarized potential vector of the heart. ECG wave has several peaks and "formations", which is useful for its diagnosis.

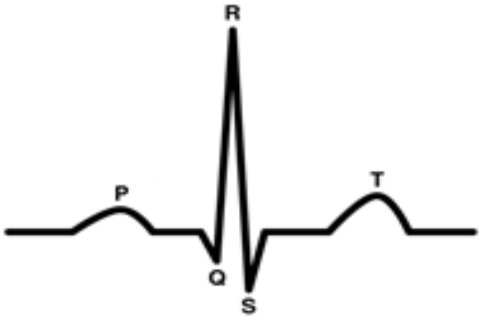

Fig. 1 Basic Structure of ECG Waveform

P wave - Represents the wave of depolarization that spreads from the SA node throughout the atria, and is usually 0.08 to 0.1 seconds (80-100 ms) length.

Q wave - Represents the normal left-to-right depolarization of the interventricular septum.

R wave - Represents early depolarization of the ventricles.

S wave - Represents late depolarization of the ventricles.

T wave - Represents ventricular repolarization and is longer in duration than depolarization.

The Pan-Tompkins algorithm (Pan et al., 1985) has been found to have a higher accuracy for various beat morphologies than other traditional real-time methods (Srinivasan et al. 2012). In this work, Modified soft rough set based classification is compared with four different classifier algorithms for ECG signals. The classification accuracy of Modified soft rough set based classification is determined for five different classes NSR, LBBB, PVC, RBBB and PR of Cardiac Arrhythmias.

1.2 ECG Interpretation

Application of the computer to the ECG for machine interpretation was one of the earliest uses of computers in the medicine (Jenkins 1981). Of primary interest in the computer-based systems was replacement of the human reader and elucidation of the standard waves and intervals. Originally this was performed by linking the ECG machine to a centralized computer via phone lines or computer network. The modern ECG diagnostic machine is completely integrated with an analog signal, a 12-to 16 bit analog to digital (A/D) converter, a central computational microprocessor, and dedicated input and output (I/O) processor. The above mentioned systems can compute a measurement matrix derived from the 12 lead signals and analyze this matrix with a set of rules to obtain the final set of interpretive statements (pryor et al. 1983).

There are hundreds of interpretive statements from which a specific diagnosis is made for each ECG, but there are only about five or six major classification groups for which the ECG is used. The first step is analyzing an ECG requirement determination of the rate and rhythm for the atria and ventricles. Included here would be any conduction disturbances either in the relationship between the various chambers or within the chambers themselves.

2 Related Work

The development of methods to analyze temporal data and more specifically ECG data has been the focus of several papers in a number of tasks such as prediction, description/summarization, classification and data visualization. Most of the classification problems are related to the extraction of relevant characteristics and increase in computational power; sophisticated algorithms have been proposed to improve the prediction accuracy of ECG waveform classification systems. This segment of the chapter discusses a variety of techniques proposed earlier in literature for feature extraction, classification of ECG signal. There are numerous studies in Feature extraction and Classification as reported in Table 1 focusing on the diagnosis of ECG signal.

Table 1 Related work for this study

Authors	Purpose	Technique
Serkan et al. 2011.	Classification	This work proposes Personalized long-term ECG classification, which helps professionals to quickly and accurately diagnose any latent heart disease by examining only the representative beats. The classification process produced results that were consistent with the manual labels with over 99% average accuracy. The proposed system over massive data (feature) collections in high dimensions.
Sung-Nien et al. 2009.	Classification	The effectiveness and efficiency of the proposed method and three other Independent Components arrangement strategies are studied in this work. Two kinds of classifiers, with probabilistic neural network and support vector machines, are used to evaluate the proposed method. The experiment results demonstrated that the proposed Independent Components arrangement strategy outperforms the other strategies.
Mohammadreza et al. 2013.	Classification	This work applies sensing approach and sensing matrix selection approach which illustrated 15% increase in Signal to Noise Ratio (SNR) and a good level of quality for the degree of incoherence between the random measurement and sparsity matrices.
Cheng et al. 2006.	Classification	The self-organizing cerebellar model articulation controller (SOCMAC) network is an unsupervised learning method proposed in this work. This method achieves a classification accuracy of 98.21%, which is comparable to the existing results.

Table 1 (*continued*)

Authors	Purpose	Technique
Adam et al. 2010.	Feature Extraction	The new Feature extraction method is applied to create the distinctive normal heartbeat samples from ECG real time signal.
Pan et al. 1985.	Feature Extraction	Real-time algorithm for detection of the QRS complexes of ECG signals is proposed in this work. It reliably recognizes QRS complexes based upon digital considers of slope, amplitude, and width. A special digital band pass filter reduces false detections caused by the various types of interference present in ECG signals.
Natalia et al. 2008.	Feature Extraction	This proposed Hamilton-Tompkins approach is applied for feature extraction.
Khoo et al. 1999.	Classification	In this work, novel approach for the classification and rule induction of inconsistent information systems is proposed.
Senthilkumar et al. 2014.	Classification	Modified soft rough set based classification is applied in this work for medical data.
Udhayakumar et al. 2013 & 2014.	Classification	In this work, bijective soft set based classification is applied for medical data.
Hari et al. 2013.	Classification	The proposed system deals with ECG signal analysis based on Artificial Neural Network based (discrete wavelet transform and morphology) features. We proposed a technique to truthfully classify ECG signal data into two classes (abnormal and normal class) using various neural classifiers.
YogendraNarain et al. 2011.	Classification	This paper proposed new techniques to delineate P and T waves efficiently from heartbeats and classification results are compared.

3 Proposed Work

The methodology adopted in this work is presented in Figure 2. In the first step, Filtering technique is applied for de-noising ECG signals. In the second step, Morphological features are extracted from ECG signal using Pan-Tompkins algorithm (Pan et al. 1985). In the third step, modified soft rough set based classification approach is applied for generating rules from the trained ECG signals and rule matching is applied for test data to compute the decision class based on reliability analysis (Senthilkumar et al. 2014).

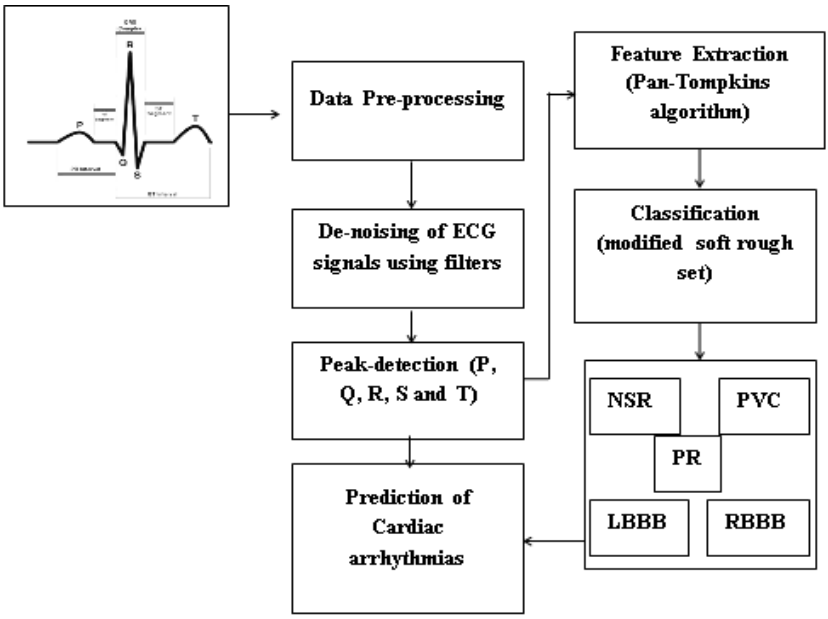

Fig. 2 Proposed Methodology

3.1 Signal Acquisition

This is first stage of signal processing; database collection is one of the most significant tasks of signal processing. The data used in this research is the ECG signals from the MIT–BIH (Massachusetts Institute of Technology –Boston's Beth Israel Hospital) Arrhythmia available on physionet website (Mark et al. & http://www.physionet.org/physiobank/database). This database contains 48 files divided into two portions first one is of 23 files (Records number 100 to 124 inclusive with some of missing records) selected at random from this set, and another one contains 25 files (numbered from 200 to 234 inclusive with some numbers of lacking). Each of the 48 records is slightly over 30 min long.

The database includes approximately 109,000 beat labels. The ECG signals from MIT–BIH database are illustrated by – a text header file (.hea), a binary file

(.dat) and a binary annotation file (.atr). The header files describe the detailed information such as number of samples, sampling rate, format of ECG signal data, type of ECG signal leads and number of ECG signal leads, patient's history and the detailed clinical information.

3.2 Preprocessing

It is to be expected that any ECG gratitude system will have to operate in a noisy hospital atmosphere. The ECG signal is normally corrupted with different types of noise. Often the information cannot be freely extracted from the raw signal, which must be handled first for a useful result.

The ECG signals were preprocessed to remove noise due to power line interference, respiration, muscles, tremors and spikes etc. The signal samples affect the segment selected and care must be taken to pick at least one cardiac cycle so that the signal can be accurately demonstrated and can be useful in diagnosis. In order to moderate high and low rate components, the ECG signal was filtered using a low pass and high pass filter. Figure 3 shows the original (noisy) ECG signal (Ravindra et al. 2013).

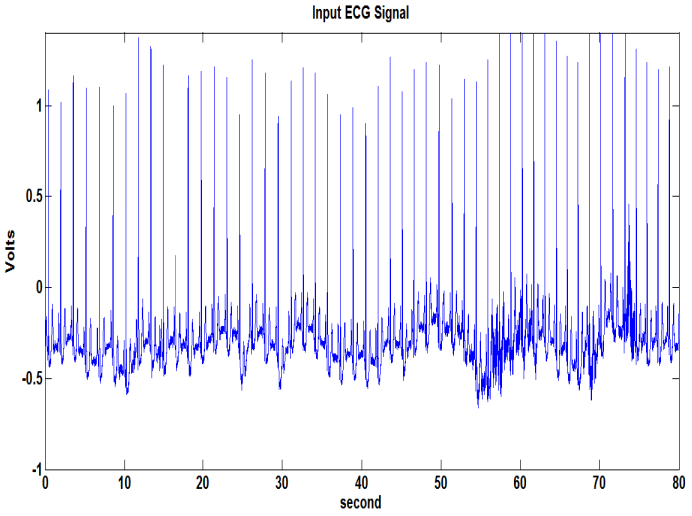

Fig. 3 Input ECG Signal

3.2.1 Filtering

A **low-pass filter** is a filter that passes low-frequency signals and attenuates (reduces the amplitude of) signals with frequencies higher than the cutoff rate. The actual amount of attenuation for each frequency varies depending on specific filter design. It is occasionally called a high-cut

filter, or three times cut filter in audio applications. A low-pass filter is the reverse of a high-pass filter. Figure 4 illustrates ECG signal after applying low-pass filter (Sameni et al. 2006; Saurabh et al. 2012).

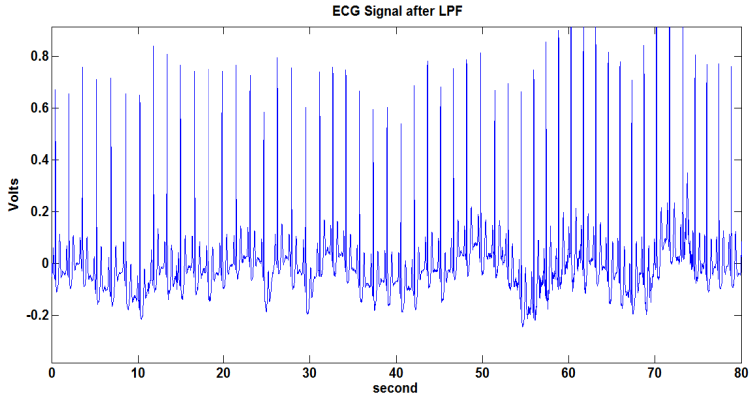

Fig. 4 ECG signal after applying Low pass filter

A **high-pass filter** (HPF) is an electronic filter that passes high-frequency signals but mitigates (reduces the amplitude of) signals with rates lower than the cut-off rate. The actual amount of attenuation for each frequency varies from filter to filter. A high pass filter is usually showed off as a linear time-invariant system. It is occasionally called a low-cut filter or bass-cut filter (Sameni et al. 2006; Saurabh et al. 2012). ECG signal after applying high pass filter is shown in Figures 5.

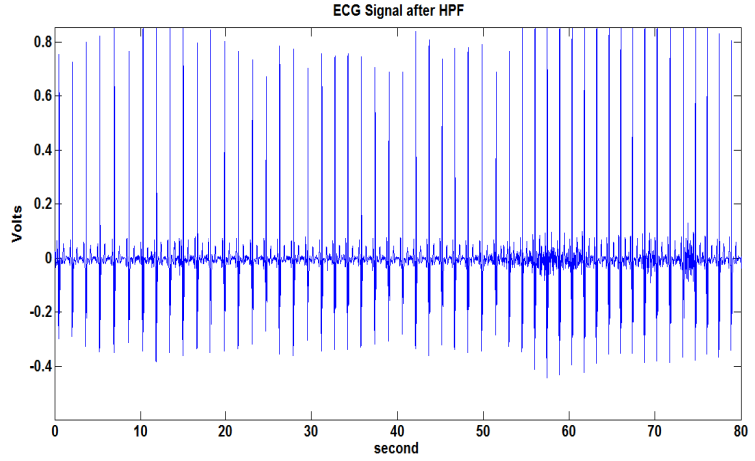

Fig. 5 ECG signal after applying high pass filter

Since a physician classifies arrhythmia with the information of rhythm and morphology, an input direction can involve of features that illustrate the rhythm and morphology features as polarity of each ECG signal. The positive, biphasic or negative polarity was considered with values 1, 0, or −1 respectively. Existence or absence of notching in signal was shown with 1 or 0.

3.3 Feature Extraction

The goal of the feature extraction stage is to find the smallest set of features that enables acceptable classification rates to be reached. In general developer cannot estimate the performance of a set of features without training and testing the classification system. Therefore a feature selection is an step by step process that involves trailing different feature sets until acceptable classification performance is achieved (Devashreejoshi et al. 2013).

This work applies Pan - Tompkins algorithm to extract the morphological features from ECG signal. This algorithm detects P, R and T peaks from ECG signals. In many cases of cardiac diseases, some parts of signal particularly P wave is by-passed and is covered by T-wave from previous heartbeat. Some parts of the heart could also inversely polarize, leading to inverted parts of signal (Adam et al. 2010). This is done by searching for the nearest maximum for P, R and T points and nearby minimum for Q and S points. The extracted morphological features are shown in Figure 6.

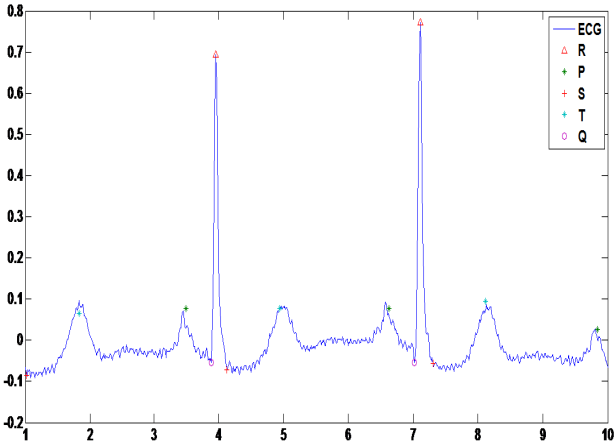

Fig. 6 Extracted P, QRS and T features

This chapter aims to develop a simple, robust and enhanced technique that would (a) derive digital time series ECG waveform from the scanned ECG image and (b) extract the ECG clinical information like amplitude and time interval of ECG parameters (characteristics wave peaks and time duration). It eliminates the

noise and pre-processing stage and morphological interpretation leading to reduction in the computational time and improvement in the accuracy. This would help in automatic diseases diagnosis (Dhar et al. 2008).

4 Preliminaries

4.1 Rough Sets

Rough set theory was initiated by Pawlak (Pawlak et al. 1982 & 2007) for dealing with vagueness and granularity in information systems. This theory handles the approximation of an arbitrary subset of a universe by two definable or observable subsets called lower and upper approximations. It has been successfully applied to machine learning, intelligent systems, inductive reasoning, pattern recognition, mere logy, image processing, signal analysis, knowledge discovery, decision analysis, expert systems, and many other fields.

Definition 1: Let R be an equivalence relation on U. The pair (U, R) is called a Pawlak approximation space. The equivalence relation R is often called an indiscernibility relation. Using the indiscernibility relation R, one can define the following two rough approximations:

$$R_*(x) = \{x \epsilon U : [x]_R \subseteq \mathrm{X}\}$$
$$R^*(x) = \{x \epsilon U : [x]_R X = \theta\}$$

$R_*(x)$ and $R^*(x)$ are called the Pawlak lower approximation and the Pawlak upper approximation of X, respectively. In general, we refer to $R_*(x)$ and $R^*(x)$ as Pawlak rough approximation operators and $R_*(x)$ and $R^*(x)$ as Pawlak rough approximations of X. The Pawlak boundary region of X is defined by the difference between these Pawlak rough approximations; that is, $\mathrm{Bnd}(X) = R_*(x) - R^*(x)$ It can easily be seen that $R_*(x) \subseteq X \subseteq R^*(x)$. A set is Pawlak rough if its boundary region is not empty; otherwise, the set is crisp. Thus, X is Pawlak rough if $R_*(x) \neq R^*(x)$.

4.2 Soft Set Theory

The soft set theory offers a general mathematical tool for dealing with uncertainty not clearly defined objects. In this section, we describe the basic notions of soft sets. Let U be initial universe of objects and E be a set of parameters in relation to objects in U. Parameters are often attributes, characteristics or properties of objects.

Definition 2: A pair (F, A) is called a soft set over U, where F is a mapping given by F: A→ P (U). In other words, a soft set over U is a parameterized family of subsets of the universe U. For ε ∈ A, F (ε) may be considered as

the set of ε -elements of the soft set (F, A) or as the set of ε -approximate elements of the soft set (Molodstov 1999).

Example 1:

Let U is the set of houses under consideration. *E* is the set of parameters. Each parameter is a word or a sentence. *E* = {expensive; beautiful; wooden; cheap; in the green surroundings; modern; in good repair; in bad repair}. In this case, to define a soft set means to point out expensive houses, beautiful houses, and so on. The soft set *(F, E)* describes the "attractiveness of the houses" which Mr. X (say) is going to buy.

Suppose that there are six houses in the universe U given by

U = {hl, h2, h3, h4, h5, h6) and E= {el, e2, e3, e4, e5):

Where E,

el stands for the parameter 'expensive',
e2 stands for the parameter 'beautiful',
e3 stands for the parameter 'wooden',
e4 stands for the parameter 'cheap',
e5 stands for the parameter 'in the green surroundings'.

Suppose that

F(e1) = {h2, h4},
F(e2) = {h1, h3),
F(e3) = {h3, h4, h5},
F(e4) = {h1, h3, h5},
F(e5) = {h1}

The soft set (F, E) is a parameterized family $\{F(e_i), i = 1,2,3,. ,8\}$ of subsets of the set U and gives us a collection of approximate descriptions of an object. Consider the mapping F which is "houses (.)" where dot (.) is to be filled up by a parameter E. Therefore, F(el) means "houses (expensive)" whose functional-value is the set { h2, h4}.Thus, we can view the soft set (F, E) as a collection of approximations as below: (F, E) ={expensive houses = { h2, h4}, beautiful houses = { hl , h3}, wooden houses = { h3, h4, h5}, cheap houses = {hl, h3, h5}, in the green surroundings = {h1}}, where each approximation has two parts:

(i) A predicate p; and
(ii) An approximate value-set u (or simply to be called value-set u).

For example, for the approximation "expensive houses = {h2, h4}", *we* have the following:

(i) The predicate name is expensive houses; and
(ii) The approximate value set or value set is {h2, h4}.

4.3 Soft Rough Sets

Soft rough sets, which can be seen as a generalized rough set model based on soft sets. The standard soft set model is used to form the granulation structure of the universe, namely the soft approximation space. Based on this granulation structure, we then define soft rough approximations, soft rough sets and some related notions. As the hybrid model combining rough sets with soft sets, soft rough sets could be exploited to extend many practical applications based on rough sets or soft sets.

Definition 3: Let S = (F, A) be a soft set over U. Then the pair P = (U, S) is called a soft approximation space. Based on P, following two operations are defined (Feng et al. 2011; Zhaowen et al. 2007):

$$\underline{apr}\ (x) = \{u \epsilon U :]a \epsilon A[u \epsilon F(a) \subseteq \mathrm{X}]\}$$
$$\overline{apr}(x) = \{u \epsilon U :]a \epsilon A\ [u \epsilon F(a), \mathrm{F(a)X} \neq \phi\ \subseteq \mathrm{X}]\}$$

for any subset X of U. Two subsets $\underline{apr}(x)$ and $\overline{apr}(x)$ called the lower and upper soft rough approximations of X in P, respectively are obtained. Moreover,

$$\mathrm{Pos(x)} = \underline{aprP}(x);$$
$$\mathrm{Neg(x)} = \mathrm{U} - \overline{aprP}(x);$$
$$\mathrm{BND(x)} = \overline{apr}P(x) - \underline{aprP}(x);$$

are called soft positive, soft negative and soft boundary regions of X, respectively. If $\overline{aprP}(x) = \underline{aprP}(x)$; X is said to be soft definable; otherwise X is called a soft rough set. Perhaps this definition does not meet the exact criteria of rough sets presented by Pawlak.

Example 2:

Table 2 Table for soft set S = (F, A)

	U1	**U2**	**U3**	**U4**	**U5**	**U6**
E1	1	0	0	0	0	1
E2	0	0	1	0	0	0
E3	0	0	0	0	0	0
E4	1	1	0	0	1	0

Let U = {u1, u2, u3, u4, u5, u6}, A = {e1, e2, e3, e4} ⊆ E. The soft set over U is S = (F, A), which is given by Table 2 and the soft approximation space is P = (U, S). From Table 2, it is clear that [u1] = {u1}, [u2] = [u5] = {u2, u5}, [u3] = {u3}, [u4] = {u4}, [u6] = {u6}. For X = {u3, u4, u5}, we have $\underline{apr}$ (X) = {u3, u4} and $\overline{apr}(X)$ = {u2, u3, u4, u5}.

5 Classification

Classification (Azar and Hassanien 2014; Inbarani et al. 2014a; Azar 2014) is a data mining technique used to predict group membership for data instances. In this final phase of classification, a classifier is constructed from the rules generated in the previous phase, and the accuracy and coverage of the classifier is tested by means of input testing data with predefined label of class attributes. A classifier can serve as a predictive model to predict the class label of unknown records if its ability of prediction is satisfied. This paper aims at proposing a classification approach, which is applied for ECG classification with relative high classification accuracy (Abdelhamid et al. 2012; Abawajy et al. 2013; Zumray et al. 2001; Zidelmal et al. 2013). The model is used to classify new ECG signals which are used as test data. In this work, classification accuracy of the proposed approach is compared with three different classifiers Back Propagation Network (BPN), Naive Bayes, JRip, J48, Multilayer Perceptron (MLP) and decision table using classification accuracy measures Precision, Recall and F-measure.

5.1 Naive Bayes

The Naive Bayes classifier assumes that all other variables are conditionally independent of each other given the classification variable. Several studies have shown Naive Bayes to be competitive with more sophisticated classifiers (Gurkan et al. 2012). Considering the task of arrhythmia classification in this study, we assume that the ECG data was generated by a parametric model and use the training data to calculate Bayes optimal estimates of the model parameters. Then, equipped with these estimates, the classifier classifies new signals using Bayes' rule to calculate the posterior probability that a class would have generated the signals. This classifier computes the conditional probabilities of the different classes given the values of attributes and then selects the class with the highest conditional probability. Supervised discretization is used to convert numeric attributes to nominal ones.

5.2 Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) is a neural network that is usually used to solve classification problems. A common type of MLP is the Feed forward Back propagation Neural Network (FFBNN), whose name is given because the input is forward propagated and the errors are back propagated to correct the neurons weights. The input layer propagates the input values to the hidden layer, multiplying the value of each neuron with the respective weight. The final values on the hidden layer neurons are obtained applying the activation function on the summation of all weighted values. The value of the output layer neurons are calculated with the same procedure, however, this value is compared with the desired value from the training set, and the difference (the error) is back propagated to correct the neuron weights for all layers. This method could be able to classify each

rhythm in a small time. Due the nature of the Multi-layer perceptron classifier, that just wants to propagate the inputs through the network to find the target class. The MLP classifier applied for ECG Signal data (Biju et al. 1997; Mehmet et al. 2010).

5.3 Back Propagation Network (BPN)

Back propagation is a form of supervised learning for multilayer nets, also known as the universal delta rule. Error data at the output layer is "backpropagated" to previous ones, allowing entering weights to these layers to be updated. It is maximum often used as training algorithm in current neural network applications. The process starts by applying the first input pattern and the corresponding target output. The input causes a response to the neurons of the first layer, which in turn cause a response to the neurons of the following layer, and so on, until a response is achieved at the output layer. That response is then compared with the target response; and the difference (the error signal) is calculated. From the error difference at the output neurons, the algorithm computes the rate at which the error changes as the activity level of the neuron changes. So far, the calculations were computed forward (i.e., from the input layer to the output layer). Now, the algorithms step back one layer before that output layer and recalculate the weights of the output layer (the weights between the last hidden layer and the neurons of the output layer) so that the output error is minimized. The algorithm next calculates the error output at the last hidden layer and computes new values for its weights (the weights between the last and next-to-last hidden layers). The algorithm continues calculating the error and computing new weight values, moving layer by layer backward, toward the input. When the input is stretched and the weights do not change, (i.e., when they have reached a stable state), then the algorithm picks the next pair of input-target patterns and repeats the process. Although responses move in a forward direction, weights are calculated by touching backward, therefore the name back propagation (Deepak et al. 2013;Gupta et al. 2012).

5.4 J48

It builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class.

5.5 JRip

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

5.5.1 Building Stage

Repeat 1.1 and 1.2 until the description length (DL) of the rule set and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

5.5.1.1 Grow Phase: Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

***5.5.1.2 Prune Phase*:** Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is (p-n)/(p+n) – but it's actually 2p/(p+n) -1, so we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

5.5.2 Optimization Stage

After generating the initial rule set {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 5.5.1 and 5.5.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the rule set. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

5.5.3 Delete the rules from the rule set that would increase the DL of the whole rule set if it were in it. Add resultant rule set to RS.

ENDDO

Note that there seem to be 2 bugs in the original ripper program that would affect the rule set size and accuracy slightly.

5.6 Decision Table

Given a training set of labeled instances, an induction algorithm builds a classifier. We describe two variants of decision table classifiers based conceptually on a simple lookup table. The first classifier, called DTMaj (Decision Table Majority) returns the majority of the training set if the decision table cell matching the new instance is empty, i.e., it does not contain any training instances. The second classier, called DTLoc (Decision Table Local), is a new variant that searches for a decision table entry with fewer matching attributes (larger cells) if the matching cell is empty. This variant therefore returns an answer from the local neighbourhood, which we hypothesized, will generalize better for real datasets that tend to

be smooth, i.e., small changes in a relevant attribute do not result in changes to the label.

A decision table has two components:

- ✓ A schema, which is a list of attributes
- ✓ A body, which is a multi-set of labeled instances. Each instance

consists of a value for each of the attributes in the schema and a value for the label. The set of instances with the same values for the schema attributes is called a cell. Given an unlabeled instance, x, the label assigned to the instance by a decision table classier is computed as follows. Let I be the set of labeled instances in the cell that exactly matches the given instance x, where only the attributes in the schema are required to match and all other attributes are ignored. If I ≠ 0 return the majority class in I, breaking ties arbitrarily. Otherwise (I = 0), the behavior depends on the type of decision table used:

1. A DTMaj returns the majority class in the decision table.
2. A DTLoc removes attributes from the end of the list in the schema and tries to and matches based on fewer attributes until one or more matches are found and their majority label is returned. This increases the cell coverage until training instances match x. Unknown values are treated as distinct values in the matching process. Given a dataset and a list of attributes for the schema, a decision table is well defined functionally.

5.7 Modified Soft Rough Set (MSR)

Soft set theory deals with uncertainty and vagueness on the one hand, while on the other hand, it has enough parameterization tools. (Feng et al., 2011) introduced soft rough sets and provided a comparative analysis of rough sets and soft sets. In order to strengthen the concept of soft rough sets, a new approach called MSR was proposed (Muhammad et al. 2013). The definitions for lower and upper soft rough approximations are given below:

Definition 4: Let (F, A) be a soft set over U, where F is a map F: A→ P (U). Let u: U→P (A) be another map defined as u(x) = {a: x ϵ F(a)}. Then the pair (U, u) is called MSR-approximation space and for any X ⊆U, lower MSR-approximation is defined as

$$\underline{X}\phi = \{X \epsilon U : \phi(x) \neq \phi(y), for\ all\ y \epsilon x^c$$

Where $x^c = U - x$ and its upper MSR-approximation is defined as

$$\bar{X}\phi = \{\, X \epsilon U : \phi(x) = \phi(y), for\ some\ y \epsilon x$$

If $\underline{X}\phi = \overline{X}\phi$, then X is said to be MSR-set.

However in above definition, parameter set A of the soft set (F, A) plays its role in defining approximations of a subset X of U. Modified Soft rough set based classification (Senthilkumar et al. 2014) is described in Figure 7.

Example 3:
Let U = {s1, s2, s3, s4, s5, s6} be the set of six stores (universe set) and A = {e1, e2, e3} ⊆ E, where

e1 represents empowerment of sales personnel,
e2 represents perceived quality of merchandise,
e3 represents high traffic location

The soft set (F,A) is representing this data in Table 3.

Table 3 Table for Soft set (F, A)

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
e1	1	0	0	1	0	1
e2	1	1	1	0	0	0
e3	0	0	0	0	1	1

Based on information presented by soft set (F,A), suppose someone may think, the set of stores X = {s1, s3, s6} represents the stores which are doing good business. Since s3 ϵ X and s2 ϵ X^c, so both of these two cannot be in $\underline{X}\phi$. Hence, according to this mapping lower and upper MSR approximations of X are

$$X=\{s1, s3, s6\}, X^c=\{s2, s4, s5\}$$

$$\underline{X}\phi= \{s1, s6\}, \overline{X}\phi=\{s1, s2, s3, s6\}$$

Algorithm: MSR based Classification

Input: Given ECG signal Dataset with conditional attributes P, Q, R, S and T and the Decision attributes NSR, LBBB, PVC, RBBB, PR

Output: Generated Decision Rules

Step1: Construct MSR approximation space for the given cardiac arrhythmia dataset

Step 2: Apply AND operation for all conditional attributes.

Step 3: Generate deterministic rules using

$$\underline{X}\phi = \{X \epsilon U: \phi(x) \neq \phi(y), for\ all\ y \epsilon x^c$$

Step 4: Generate non-deterministic rules by using

$$\overline{X}\phi = \{X \epsilon U: \phi(x) = \phi(y), for\ some\ y \epsilon x$$

Step 5: Compute the support value for each non-deterministic rule

$$support = \frac{support(A \wedge B)}{support(A)}$$

Fig. 7 MSR Based Classification

6 Experimental Analysis and Results

Data set taken for experiment is obtained from PhysioBank (http://physionet.org/physiobank/database/mitdb/). The ECG signal data sets were taken for experimental analysis of the modified soft rough set based classification approach. This is the basis of the cardiac rhythm analysis, which is usually applied to 24 ECG recordings to identify the cardiac diseases. ECG signal data sets were very popularly used in cardiac arrhythmia research literature. The proposed algorithm is applied to training data and the generated classification rules are matched with test data to determine exact class. The attributes in these data sets are all numerical. 80% of the data is chosen as the training set and 20% as testing data. The performance of the proposed approach is compared with other classification approaches like Back Propagation Neural Network, Decision table, J48, JRip, Multilayer Perceptron and Naive Bayes. The classes of ECG signals, their records and annotations are shown in Table 4.

Table 4 ECG signal Database Classes, Records and Annotations

Classes	Records	Annotations
1. Normal Sinus Rhythm (NSR)	101, 112, 113, 115, 117, 121, 122, 209, 230, 234	97, 119, 80, 87, 69, 83, 121, 130, 108, 128
2. Left Bundle Branch Block (LBBB)	109	124
3. Premature Ventricular Contraction (PVC)	114,203, 205, 219, 228	76, 146, 124, 105, 110
4. Right Bundle Branch Block (RBBB)	105, 106, 124, 202,212, 232	115, 92, 69, 73, 125, 86
5. Paced Rhythm (PR)	102, 217	102, 101

6.1 Evaluation Measures

Evaluation is the key to making advances in data mining, and it is especially important when the area is still at the early stage of its development. Demanding dataset community has criticized the use of non-class independent evaluation measures, such as the overall accuracy, for reporting experimental results on well-adjusted datasets. The non-class independent evaluation fails because the results only reflect the learning performance of the majority class, and the more skewed the class distribution is the worse the effect will be. Therefore, when we evaluate the performance on demanding datasets, we want to focus on individual classes.

There are many evaluation measures in data mining, some are: precision, recall, F-measure (Cheng et al. 2009).

Precision is the average probability of relevant retrieval. Recall is the average probability of complete retrieval. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure. A Precision, Recall and F-measure evaluation measures collects and reports a suite of descriptive statistics for binary classification tasks. The basis of a precision recall evaluation is a matrix of counts of actual and predicted classifications. This matrix is known as confusion matrix for two class problem (Cheng et al. 2009) as shown in Table 5.

Table 5 Confusion matrix

	Predicted Class	
Actual Class	True Positive	False Negative
	False Positive	True Negative

- ✓ Correctly detected (True positives)
- ✓ Falsely detected (False positives)
- ✓ Undetected (False negatives)

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

6.2 Performance Evaluation

Table 6 and Figure 8 illustrate the performance of proposed as well as compared classification approaches. As illustrated in the figure, MSR-based classification outperforms all the other classification approaches for Cardiac arrhythmia data set. Precision of MSR is higher than the J48, decision table, Naïve bayes, MLP, BPN, J48 and JRip. Based on F-Measure, MSR based classification approach outperforms all the other approaches. Recall of MSR is higher than J48, Decision table, Naïve bayes, MLP, BPN, J48 and JRip.

Table 6 Performance analysis of the classification algorithms for ECG signals

Algorithms	Confusion Matrix						Precision	Recall	F-Measure
	Actual Class	Predicted Class							
		NSR	LBBB	PVC	RBBB	PR			
Decision Table	NSR	183	1	4	0	1	0.67	0.968	0.792
	LBBB	7	2	0	16	1	0.977	0.697	0.813
	PVC	31	3	76	0	1	0.927	0.685	0.788
	RBBB	35	0	2	0	85	0.833	0.638	0.723
	PR	17	30	0	0	0	1	0.64	0.78
	Over all Measures						**0.836**	**0.789**	**0.789**
JRip	NSR	171	0	15	2	1	0.803	0.905	0.851
	LBBB	3	0	0	22	0	0.846	0.88	0.863
	PVC	19	3	85	0	4	0.833	0.766	0.798
	RBBB	10	0	2	1	109	0.948	0.893	0.92
	PR	10	35	0	1	1	0.921	0.745	0.824
	Over all Measures						**0.859**	**0.854**	**0.854**
Naïve Bayes	NSR	106	3	73	4	3	0.475	0.561	0.515
	LBBB	0	0	2	23	0	0.821	0.92	0.868
	PVC	40	0	70	0	1	0.407	0.631	0.495
	RBBB	52	0	21	0	49	0.891	0.402	0.554
	PR	25	13	6	1	2	0.813	0.277	0.413
	Over all Measures						**0.612**	**0.528**	**0.528**
MLP	NSR	178	5	4	0	2	0.809	0.942	0.87
	LBBB	0	0	0	25	0	0.781	1	0.877
	PVC	11	0	78	0	22	0.729	0.703	0.716
	RBBB	21	1	15	0	85	0.759	0.697	0.726
	PR	10	17	10	7	3	0.739	0.362	0.486
	Over all Measures						**0.771**	**0.775**	**0.764**
J48	NSR	172	0	14	1	2	0.896	0.91	0.903
	LBBB	1	0	0	24	0	0.889	0.96	0.923
	PVC	9	0	97	0	5	0.843	0.874	0.858
	RBBB	6	0	3	0	113	0.934	0.926	0.93
	PR	4	39	1	2	1	1	0.83	0.907
	Over all Measures						**0.903**	**0.901**	**0.901**
BPN	NSR	167	0	17	4	1	0.795	0.901	0.831
	LBBB	3	0	1	21	0	0.845	0.86	0.860
	PVC	20	3	81	0	6	0.833	0.766	0.798
	RBBB	12	0	4	1	105	0.938	0.893	0.921
	PR	10	35	0	1	1	0.911	0.735	0.834
	Over all Measures						**0.85**	**0.854**	**0.854**
MSR-Sets	NSR	1022	0	0	0	0	1	1	1
	LBBB	0	124	0	0	0	1	1	1
	PVC	0	0	559	0	0	1	1	1
	RBBB	0	0	0	556	0	1	1	1
	PR	3	0	0	2	201	1	1	1
	Over all Measures						**1**	**1**	**1**

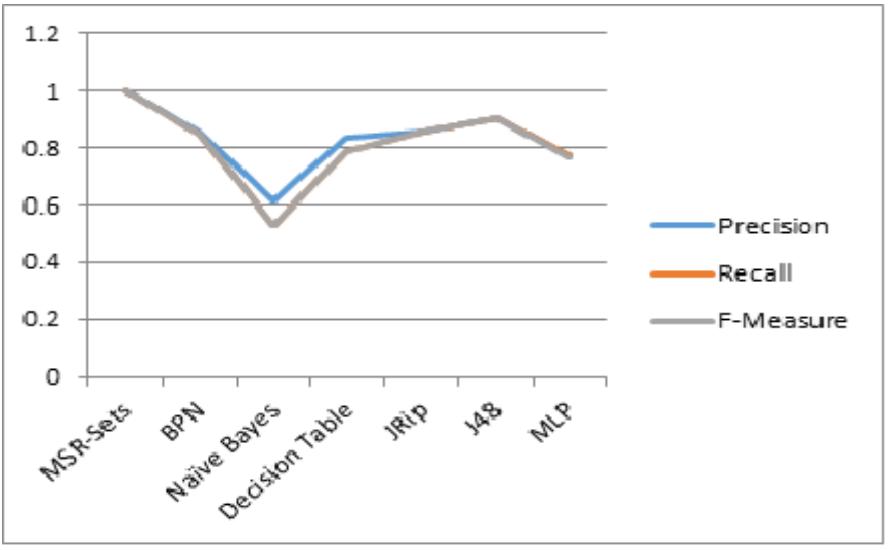

Fig. 8 Comparative analysis of classification algorithms for ECG signal

6.3 Discussion

Table 6 and Figure 8 demonstrate the precision, recall and F-measure values. It is interesting to note that an increase in classification accuracy is recorded for the proposed method. The MSR based classification accuracy is higher than Back Propagation Neural Network, Decision Table, JRip, J48, Multilayer perceptron and Naive Bayes approach.

From the above results, we can easily determine that the MSR-algorithm is the best method for ECG signal diagnosis since it achieves 100% accuracy. The proposed classification method is applied to extracted data set thus reducing the number of rules while leading to significantly better classification accuracy.

Based on the results of the MSR-Set, it can be said that the modified soft rough set classification that is designed in this chapter obtained excellent results using the proposed signal processing and feature extraction methods on the ECG data obtained from MIT-BIH database. Most of the signal classification methods and QRS detection algorithms are standard and the feature extraction algorithm proposed in this chapter proved to be very successful for this job. Rather than taking whole ECG beat samples and feeding them into network as features, extracting only P, Q, R, S, T waves which approximately intervals represented each different classes of ECG wave effectively and reduced complexity of the classification system based MSR Classification, making it faster and more accurate (Dingfei et al. 2002). The framework can provide the full functionality that a big data processing platform requires and use in various medical domains. In this chapter gives a new paradigm for integrating various ECG signal aspects and thus

provides a complete solution for handling various clinical issues such as examination, prediction and handling larger volume of data in real time.

However current very high results of the classification system are very dependent on its current design. If we had tried to classify five cardiac conditions (diseases) only have used noisy data and detecting the QRS complexes. To avoid that, more efficient signal classification and filtered can be applied to the ECG data and since with increasing number of classes, the samples between the intervals of the ECG waves will become more and more similar and it will become harder for the classifier to distinguish the differences between them, so another way of feature extraction based on the combination of statistical, morphological features can be considered. But for the conditions of our chapter, it can be easily repeated that our ECG pattern recognition and classification system did what it should do and classified five different cardiac conditions very accurately, effectively and efficiently.

7 Conclusion

MSR based classification of ECG signals has been proposed in this paper for classification of ECG signals. The important morphological features were extracted from ECG signals. MSR-set approach provides higher accuracy than Back Propagation Neural Network, Decision Table, JRip, J48, Multilayer perceptron and Naive Bayes approaches and it generates more compact rules. It can be concluded that the time domain based filters performed well and QRS detection based on Pan-Tompkins algorithm proved to be very suitable for our case. The method proposed for feature extraction was an effective method in classifying five different cardiac conditions. The advantage of the MSR based classification using the proposed system is its simplicity and ease of implementation. The proposed system achieved 100% accuracy.

Acknowledgement. The second author would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-41-650/2012 (SR).

References

Abawajy, J.H., Kelarev, A.V., Chowdhury, M.: Multistage approach for clustering and classification of ECG data. Computer Methods and Programs in Biomedicine 112(3), 720–730 (2013)

Abdelhamid, D., Latifa, H., Naif, A., Farid, M.: A wavelet optimization approach for ECG Signal Classification. Biomedical Signal Processing and Control 7(4), 342–349 (2012)

Szczepański, A., Saeed, K., Ferscha, A.: A New Method for ECG Signal Feature Extraction. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2010, Part II. LNCS, vol. 6375, pp. 334–341. Springer, Heidelberg (2010)

Azar, A.T.: Neuro-fuzzy feature selection approach based on linguistic hedges for medical diagnosis. International Journal of Modelling, Identification and Control (IJMIC) 22(3) (forthcoming, 2014)

Azar, A.T., Hassanien, A.E.: Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier. Soft Computing (2014), doi:10.1007/s00500-014-1327-4

Azar, A.T., Banu, P.K.N., Inbarani, H.H.: PSORR - An Unsupervised Feature Selection Technique for Fetal Heart Rate. In: 5th International Conference on Modeling, Identification and Control (ICMIC 2013), Egypt, August, 31, September 1-2 (2013)

Biju, P., Simon, E.C.: An ECG Classifier Designed Using Modified Decision Based Neural Networks. Computers and Biomedical Research 30(4), 257–272 (1997)

Cheng, G., Weng, J.P.: A New Evaluation Measure for Imbalanced Datasets. In: Proceedings of the 7th Australasian Data Mining Conference, vol. 87, pp. 27–32 (2006)

Cheng, W., Teng-chiao, L., Kung-chiung, C., Chih-hung, H.: Classification of ECG complex using self-organizing CMAC. Measurement 42(3), 395–407 (2009)

Deepak, D., Avinash, W.: Study of Hybrid Genetic Algorithm Using Artificial Neural Network in Data Mining for the Diagnosis of Stroke Disease. International Journal of Computational Engineering Research 3(4), 95–100 (2013)

Dingfei, G., Narayanan, S., Shankar, M.K.: Cardiac arrhythmia classification using autoregressive modelling. Bio-Medical Engineering On Line 1(5), 1–12 (2002)

Devashree, J., Rajesh, G.: Performance analysis of feature extraction schemes for ECG signal classification. International Journal of Electrical, Electronics and Data Communication 1 (2013) 2320-2084

Dhar, P.K., Hee-Sung, J., Jong-Myon, K.: Design and implementation of digital filters for audio signal processing. In: The Third International Forum on Strategic Technologies, pp. 332–335 (2008)

Feng, F., Xiaoyan, L., Leoreanu-Fotea, V., Young, B.J.: Soft sets and soft rough sets. Information Sciences 181(6), 1125–1137 (2011)

Jenkins, J.M.: Computerized electrocardiography. Critical Review Bio-Engineering: CRC 6(4), 307–357 (1981)

Jeffrey, D., Sanjay, G.: Map Reduce: simplified data processing on large clusters. Communications of the ACM - 50th Anniversary 51(1), 107–113 (2008)

Gupta, K.O., Chatur, P.N.: ECG Signal Analysis and Classification using Data Mining and Artificial Neural Networks. International Journal of Emerging Technology and Advanced Engineering 2(1), 2250–2459 (2012)

Gurkan, H.: Compression of ECG signals using variable-length classified vector sets and wavelet transforms. EURASIP Journal on Advances in Signal Processing 119(1), 1–17 (2012)

Hari, M.R., Anuragm, T., Shailja, S.: ECG signal processing for abnormalities detection using multi-resolution wavelet transform and Artificial Neural Network classifier. Science Direct 46(9), 3238–3246 (2013)

Inbarani, H.H., Jothi, G., Azar, A.T.: Hybrid Tolerance-PSO Based Supervised Feature Selection For Digital Mammogram Images. International Journal of Fuzzy System Applications (IJFSA) 3(4), 15–30 (2013)

Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Computer Methods and Programs in Biomedicine 113(1), 175–185 (2014a)

Inbarani, H.H., Banu, P.K.N., Azar, A.T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Computing and Applications (2014b), doi:10.1007/s00521-014-1552-x

Mark, R., Moody, G.: MIT–BIH arrhythmia database directory, http://ecg.mit.edu/dbinfo.html

Mehmet, K., Berat, D.: ECG beat classification using particle swarm optimization and radial basis function neural network. Expert Systems with Applications 37(12), 7563–7569 (2010)

Muhammad, S., Muhammad, I.A., Tanzeela, S.: Another approach to soft rough sets. Knowledge-Based Systems 40, 72–80 (2013)

Mohammadreza, B., Kaamran, R., Sridhar, K.: Robust Ultra-Low-Power Algorithm for Normal and Abnormal ECG Signals Based on Compressed Sensing Theory. In: The 4th International Conference on Ambient Systems, Networks and Technologies, vol. 19, pp. 206–213 (2013)

Molodtsov: Soft set theory-Rough first results. Computational Mathematics Application 37(4-5), 19–31 (1999)

Arzeno, N.M., Zhi-De, D., Chi-Sang, P.: Analysis of First-Derivative Based QRS Detection Algorithms. IEEE Transactions on Biomedical Engineering 55(2), 478–484 (2008)

Khoo, L.P., Tor, S.B., Zhai, Y.L.: A Rough-Set-Based Approach for Classification and Rule Induction. International Journal Advanced Manufacturing Technology 15(6), 438–444 (1999)

Pan, J., Tompkins, W.: A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering 32(3), 230–236 (1985)

Pawlak, Z.: Rough sets. International Journal of Computer Information Science 11(5), 341–356 (1982)

Pawlak, Z., Skowron, A.: Rough sets: some extensions. Information Science 177(1), 28–40 (2007)

Pawlak, Z., Skowron, A.: Rough sets and Boolean reasoning. Information Science 177(1), 41–73 (2007)

Pryor, T.A., Gardner, R.M., Clayton, P.D., Warner, H.R.: The Help system. Journal of Medical Systems 7(2), 87–102 (1983)

Ravindra, P.N., Seema, V., Singhal, P.K.: Reduction of Noise from ECG Signal Using Fir Low Pass Filter With Various Window Techniques. Current Research in Engineering, Science and Technology (CREST) Journals 1(5), 117–122 (2013)

Sam, Madden: From Databases to Big Data. IEEE Internet Computing 16(3), 4–6 (2012)

Sameni, R., Vrins, F., Parmentier, F., Hérail, C., Vigneron, V., Verleysen, M., Jutten, C., Shamsollahi, M.B.: Electrode Selection for Noninvasive Fetal Electrocardiogram Extraction using Mutual Information Criteria. In: 26th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, CNRS Paris (France), vol. 872, pp. 97–104 (2006)

Saurabh, S.R., Bhadauria, S.S.: Implementation of FIR Filter Using Efficient Window Function and Its Application In Filtering a Speech Signal. International Journal Electrical, Electronic and Mechanical Controls 1(1), 1–12 (2012)

Senthilkumar, S., Inbarani, H.H., Udhayakumar, S.: Modified Soft Rough set for Multiclass Classification. In: Krishnan, G.S.S., Anitha, R., Lekshmi, R.S., Senthil Kumar, M., Bonato, A., Graña, M. (eds.) Computational Intelligence, Cyber Security and Computational Models. AISC, vol. 246, pp. 379–384. Springer, Heidelberg (2014)

Srinivasan, J., Vani, D.: An Improved Method for ECG Morphological Features Extraction from Scanned ECG Records. In: 4th International Conference on Bioinformatics and Biomedical Technology, vol. 29, pp. 64–68 (2012)

Sung-Nien, Y., Kuan-To, C.: Selection of significant Independent Components for ECG beat Classification. Expert Systems with Applications 36(2), 2088–2096 (2009)

Serkan, K., Turker, I., Jenni, P., Moncef, G.: Personalized long-term ECG Classification: A systematic approach. Expert Systems and with Applications 38(4), 3220–3226 (2011)

Udhayakumar, S., Inbarani, H.H., Senthilkumar, S.: Bijective soft set based classification of Medical data. In: International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), pp. 517–521 (2013)

Udhayakumar, S., Inbarani, H.H., Senthilkumar, S.: Improved Bijective-Soft-Set-Based Classification for Gene Expression Data. In: Krishnan, G.S.S., Anitha, R., Lekshmi, R.S., Senthil Kumar, M., Bonato, A., Graña, M. (eds.) Computational Intelligence, Cyber Security and Computational Models. AISC, vol. 246, pp. 127–132. Springer, Heidelberg (2014)

YogendraNarain, S., Phalguni, G.: Correlation-based classification of heartbeats for individual Identification. Soft Computing 15(3), 449–460 (2011)

Zhaowen, L., Bin, Q., Zhangyong, C.: Soft Rough Approximation Operators and Related Results. Journal of Applied Mathematics,15 pages (2007, January 2013)

Zidelmal, Z., Amirou, A., Ould-Abdeslam, D., Merckle, J.: ECG Beat Classification using a cost sensitive Classifier. Computer Methods amd Programs in Biomedicine 111(3), 570–577 (2013)

Zumray, D., Tamer, O.: ECG beat Classification by a novel hybrid neural network. Computer Methods and Programs in Biomedicine 66(2-3), 167–181 (2001)

Towards a New Architecture for the Description and Manipulation of Large Distributed Data

Fadoua Hassen and Amel Grissa Touzi

Abstract. The exponential growth of generated information volume, the loss of structure meaning due to data and sources variety along with a highly exhausting applications and end-users led to centralized databases distribution. One of the common approaches to satisfy performance need and preserve relational integrity, is a correctly designed and implemented decentralized database. IT systems migration from centralized to distributed database may imply heavy costs including review of existing systems core and interfaces. Also, an incongruous design may be fatal in Big Data processing systems such as data loss due to completeness rule break. A simple field replication may be acceptable in "normal size" databases, but will result to significant storage space waste. Indeed, marketed DDBMS are currently very far from automated support for the large distributed data. This heavy task is always done without any GUI or a friendly assistance that insures distribution rules (completeness, disjointness and reconstruction). Moreover, database transparency still not automatically ensured even with reported distribution script. Data treatment stored procedures and functions must take in consideration the distributed context. This context switch may result to re-writing the complete algorithm of data treatment. The aim of this paper is to propose a new architecture for the description and manipulation of Large Distributed Data. The result of this approach is a distribution context aware tool that respects database distribution rules and helps designers to easily create reliable DDB scripts. To avoid the core application and interfaces review, an automated translator from centralized format queries to distribution context aware queries. Even after the migration, end-users and application will see the distributed database as it was before splitting. This level of transparency is guaranteed by the queries translator.

Fadoua Hassen
Université Tunis El Manar, LIPAH, FST, Tunis, Tunisia
e-mail: hassen.fadoua@gmail.com

Amel Grissa Touzi
Université Tunis El Manar, ENIT, LIPAH, FST
Bp. 37, Le Belvédère 1002 Tunis, Tunisia
e-mail: amel.touzi @enit.rnu.tn

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_17

1 Introduction

Nowadays, with the development of information and storage of large volumes of distributed and heterogeneous data, Distributed Database Systems Management (DDBMS) have become an indisputable necessity for the majority of information systems. Relational databases are not the leading platforms for real-time processing performance in distributed contexts as reported in many benchmarks (Dede et al., 2014) but still hold the unique solution for the data integrity considerations in critical systems. This criterion is not always discussed in Big Data treatment solutions as it is always a knowledge extraction aim from non-structured data regardless distributed integrity constraints check (Patterson, 2013). Some works on data integrity in distributed context discussed a relaxed ACID approach to overcome temporary DDB anomalies. However, the relaxed ACID is not the perfect solution for critical systems as consistency criteria is tolerable in this approach. The distributed databases based processing of Big Data (Manyika et al., 2011) has many advantages especially related to distributed storage and parallel abilities (Wu et al., 2013). The MapReduce (Krishnanal et al., 2010) algorithm is one of the most efficient algorithms in this context (Sakr and Liu, 2014) (Polo, 2013). It is used to perform aggregated queries on a DDB. In the Big Data community, MapReduce has been seen as one of the key enabling approaches for meeting continuously increasing demands on computing resources imposed by massive data sets. The reason for this is the high scalability of the MapReduce paradigm which allows for massively parallel and distributed execution over a large number of computing nodes (Lindblad et al., 2014). The query request must be executed separately in each node (Map) and then aggregated back to the client (Reduce). The returned result is a distinct union of collected results among nodes. The client gets the aggregated query as if it was a single query. This algorithm allows parallel processing, and thus drastically improves queries response time. When combined with an efficient database distribution design, MapReduce fetch queries are faster (Aji et al., 2013) (Giese et al., 2013). An efficient database distribution is the scheme that prevents any data loss, minimizes the storage space needed and maintains initial data integrity. These specifications are always set on distributed database creation among different nodes of the distribution.

This paper focuses on Big Data processing enhancement by ensuring the best possible distribution strategy. Distributed databases design must be compliant with correctness rules (Silberschatz, 2002) to avoid data loss and unneeded storage consumption. These rules are completeness, disjointness and reconstruction. The first rule is essential for preserving original data while disjointness avoids data duplication and thus extra updates and storage space. The reconstruction aspect of data distribution preserves the relational integrity. In a small database model, it is always possible to check the three rules compliance with the distribution policy. But, when moving to Big Data context, the database schema becomes more complex and manual checking is not always possible. Thus, an automated

validation tool is required to assist designers in their distributed schema implementation. Moreover, transforming centralized CRUD (Seung et al., 2014) procedures, functions and triggers to handle distributed integrity tests is relative to database object's count and relationships. The fragment's allocation is another important consideration to invest in when creating a distributed database scheme. The communications costs between sites for data retrieval are one of the most critical criteria. Thus data allocation algorithms try to minimize this cost by assigning fragments at or near the sites they may be needed. Data Allocation Problem (DAP) is known to be NP-Hard and this makes heuristic algorithms desirable for solving this problem. The QAP is a well-known problem that has been applied to different problems successfully (Tosun et al., 2013).

One more important process in distributed databases design is the replicas synchronization. There are two main approaches for data synchronization: Either dedicated this task to the relational character of DDBMS or implement another fault tolerant mechanism based on database basic tools such as triggers and updatable views (materialized). Of course, there are external tools to do so via hardware and OS replication mechanisms (Etemad and Küpçü 2013), (Soares et al., 2013). These clustering and synchronization mechanism are always very expensive and their efficiency is related to many factors. Thus introducing new components is not always the good choice as it will add additional levels of complexity.

This paper suggests an enhancement for distributed databases querying by keeping a distribution dictionary on each site acting as a soft-switch route queries to concerned fragments.

This paper is organized in 5 main chapters. The first chapter introduces distributed databases contribution in Big Data context. The second chapter shows a sample design and implementation of an Oracle DDB in order to illustrate the task quandary on a sample limited scheme. Chapter three describes the new approach's ground and basics. Chapter four describes the implemented solution and explains how this is useful for the "perfect" distributed database's creation. Chapter five explains work limits and what needs to be done in further works.

2 Distributed Databases and Big Data

A distributed database is defined as a collection of logically interconnected data physically stretched over a network (QIAN et al., 2010). The software that manages a distributed database and ensures its transparency towards users is the DDBMS. This chapter explains how DDBMS can help overcoming Big Data challenges.

2.1 Big Data Shortcomings on Centralized Architecture

The first need expression behind big data is always described as "I have too much data coming in in too fast to handle with any RDBMS (Goswami and Kundu, 2013)". The market leader of CPU (Pukdesree et al., 2006) and electronic computer parts, Intel (Stinson and Chandramouly, 2014), seems to be the unique entity that believes that this need may be satisfied in a centralized entity nowadays. This is not a just a theory but has its pros in real life experiences. Most of companies invested huge amounts on building and securing their datacenters and may not be yet ready for the change. The storage spaces are always a cumulated deposit of various data format even describing the same documents and objects. This may be the result of normalization changes or used applications upgrades. The loss of documents and information structures then is the first step toward Big Data enforce. Moreover, the digitalization of every single object in real life and the market growth for successful companies infers large amounts of data to be stored and processed correctly. This is where data consistency seems to be the most important thing to consider, and the centralized databases architectures are better prepared for such a constraint; Maintaining integrity constraints across distributed database is not as easy as in a centralized RDBMS where a single query and transaction engine masters all the relationships' constraints. In the centralized use case, designers must keep in mind that processing strategy is a serial algorithm as RDBMS refers to a centralized CPU. This is main drawback of a centralized architecture: The serial processing. Furthermore, hardware inputs (network cards) are always outdated even with redundancy and regardless their perfection levels compared to data exchange evolution. Works on parallel stations has shown a limit on heavy processing even with the huge progress they made in the last three decades. Parallel processing operating systems then bet on networks progress to manage distributed nodes and fully profit from grid computing techniques.

Until now, the third Big Data phenomenon basic was not yet mentioned: The velocity (Embley and Liddle, 2013). When coming to performance field, parallel processing are more efficient than serial treatments. This criterion is crucial in many systems' contexts (Example: Purse mining, military defense ...). Thus, distributing storage and management is a fatality for performance enhancement. Running parallel queries among standalone processing units reduces the response time considerably. This confirmation assumes that data distribution are well split and located in order to avoid fetching repeated records where there is no need to replicate them and accurate packet sizes when exchanging results over a network. Such an assumption may not be taken for granted when dealing with huge relational models because it depends closely on applications business intelligence and queries weights over the full system history.

2.2 *Performance Consideration on Distributed Databases Context*

The second need always expressed by big data candidate companies as "I have a lot of server stretched around the world and I need to read and write in near real-time. The contribution of distributed databases on querying performance comes from the native parallelism on such architectures.

Based on this specification, the collaborative nodes, managed by the DDBMS, work on the same query but in smaller fragments. Consider an initial dataset of Facebook (Lewisa et al., 2008) contacts for example described by their IDs, names and addresses. A quick comparison between a horizontally fragmented table against the initial table highlights the processing time gain.

The sample query processing described in figure 1 contains three transactions processing in a centralized DBMS compared to a DDBMS. Each transaction is treated and then committed after all nodes response. Querying data over only four nodes already highlights the considerable response time gain.

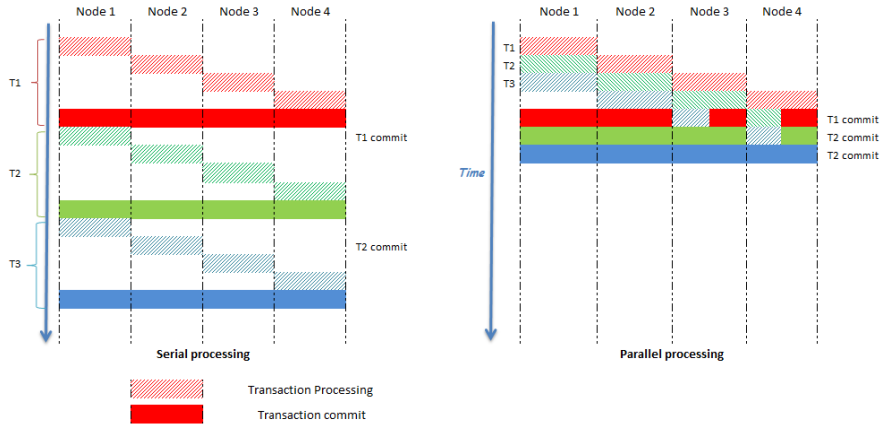

Fig. 1 Parallel processing vs serial processing

2.3 *Transparency Considerations*

DDB implementation must comply with a series of transparencies. This set of transparencies' main aim is to let end-user see a DDB as a centralized database. Those levels can be described as:

- Distribution transparency ensures that end-user can ignore data replication or splitting (Wang, 2014). As a direct consequence, on data update, the system takes care of updating all the copies of the altered row.
- Transaction transparency grants overall database transparency on concurrent users' access to data and on system faults.

- Performance transparency which provides a reliable query management even when referencing data on multiple sites.
- Query transparency on mixed content call from different sites in the same call.
- DBMS transparency which ensures access to different DBMS without user being aware of it.

3 Overview of the Existing DDBMS

3.1 DDBMS Specifications

By definition, a distributed database management system is software that manages a distributed database and ensures its transparency towards users. The most important DDBMS on the market are Oracle (Chen et al., 2014), 2, MySql (Krishnan et al., 2010) INGRES (Stonebraker, 1979), CASSANDRA (Hewitt, 2010) and F1 (Shute et al., 2012)

In 1987, DATE (Silberschatz, 2002) established twelve rules to define the perfect DDBMS, based on a fundamental principle: User must see DDBMS as a non-distributed database management system. The twelve rules are: Local autonomy, no dependency to a central site, continuous availability, Location independence, Fragmentation independency, Replication independence, Distributed query processing, Distributed transaction processing (Özsu and Valduriez, 2011) Hardware abstraction, Operating systems independence, Network independence and database independence.

3.2 Example of DDB Implementation under Oracle

A distributed database appears to a user as a single database but is, in fact, a set of databases stored on multiple computers. The data on several computers can be simultaneously accessed and modified over a network access. Each database server in the distributed database is controlled by its local DBMS, and each one cooperates to maintain the consistency of the global database. Oracle supports two types of distributed databases. In a system based on homogeneous distributed data, all databases are Oracle databases. In a system of distributed heterogeneous database, at least one of the databases is not an Oracle database.

In this section, a part of the distributed database implementation is described via a real Oracle 10g script to highlight the designer task complexity.

Example
Three institutions of the University of Tunis El Manar: National Engineering School of Tunis (ENIT), Faculty of Mathematical, Physical and Natural Sciences of Tunis (FST) and Faculty of Economic Sciences and Management of Tunis (FESMT) have decided to pool their libraries and service loans, to enable all students to borrow books in all the libraries of the participating institutions.

Joint management of libraries and borrowing is done by a database distributed over 3 sites (Site1 = ENIT, Site2 = FST and Site3 = FESMT), the global schema is as follows:

Table 1 Sample centralized database schema

EMPLOYEE (**SSN**, fname, lname, address, status, assignment)
STUDENT (**NCE**, stud_fname, stu_lname, address, institution, class, nb_borrow)
BOOK (**Id_book**, Title, editor, Year, Area, Stock, website)
AUTHOR (**Id_book**, au_lname, au_fname)
LOANS (**Id_book**, NCE, **date_borrows**, return_date)

The management of this application is based on the following assumptions:

- An employee is assigned to a single site. Each site is responsible of managing its employee.
- A student is enrolled in a single institution, but can borrow from all libraries.
- A book borrowed from a library is rendered to the same library.
- The nb_borrow field of STUDENT relation is used to limit the number of books borrowed by a student simultaneously on all libraries. It is updated at each loan and each return, regardless of the lending library.
- Each institution manages its own students.
- Each library manages its staff and works it holds.

The first conclusion a DBA can deduct is that overall relations must be split and distributed over different sites. The second inference of functional hypothesis is to split STUDENT table into two vertical fragments and duplicate the vertical fragment created on every site. The table fragment to duplicate is the sub-relation containing student id (NCE) and the numbers of active borrows (nb_borrow).

The resulting local schema for site FST for example would be:

Table 2 Local schema description

D language script: Local Schema for site FST
EMPLOYEE_FST(**SSN**, fname, lname, address, status, assignment)
STUDENT_FST (NCE, stud_fname, stu_lname, Address, Institution, Class)
STUDENT_BIBLIO (NCE, nb_borrow)
BOOKS_FST (**Id_book,** Title, editor, Year, Area, Stock, website)
AUTHORS_FST (**Id_book**, au_lname, au_fname)
LOANS_FST (**Id_book**, NCE, **date_borrows**, return_date)

The main advantages of the described distribution policy are:

- Reducing exchanged data size among sites when requesting students' data between sites.

- Lowering network activity by local checks on active borrows (nb_borrows) as it will be available locally thanks to STUDENT_BIBLIO fragment replication.

This distribution policy isn't the perfect strategy because of the more expensive updates on STUDENT_BIBLIO.

To illustrate the necessary very long task for implementing a DDB, the listing (Table 3) describes a part of the necessary script for implementing the relation STUDENT.

Table 3 Local schema creation script

PL/SQL Script: An excerpt from site creation script

```
--------------------------------------------------------------------------------
-- DDL for DB Link BB1
--------------------------------------------------------------------------------
 CREATE DATABASE LINK "ENIT_dblink"
  CONNECT TO "ROOT" IDENTIFIED BY VALUES 'root'
  USING '(DESCRIPTION= (ADDRESS= (PROTOCOL=TCP)
(HOST=127.0.0.1) (PORT=1521)) (CONNECT_DATA=
(SERVICE_NAME=ENIT)))';
-----------------------------------------------------
-- DDL for Table STUDENT_FST
-----------------------------------------------------
 CREATE TABLE STUDENT_FST ( NCE NUMBER, ST_FNAME
VARCHAR2(200),
ST_FLNAME VARCHAR2(200), ADRESS VARCHAR2(200 CHAR),
CLASS NUMBER, CURSUS VARCHAR2(200 CHAR),Constraint PK11
primary key (NCE) ) ;
--------------------------------------------------------------------------------
-- DDL for Synonym ENIT
--------------------------------------------------------------------------------
CREATE SYNONYM STUDENT_ENIT FOR STUDENT_ENIT@ENIT;
--------------------------------------------------------------------------------
-- DDL for materialized view STUDENT
--------------------------------------------------------------------------------
create materialized view STUDENT
refresh complete start with sysdate next sysdate + 7
 as
(SELECT * from FOR STUDENT_ENIT@ENIT) UNION
(SELECT * from FOR STUDENT_ENIT@FST) UNION
(SELECT * from FOR STUDENT_ENIT@FESMT)
--------------------------------------------------------------------------------
-- DDL for trigger insetSudent
--------------------------------------------------------------------------------
```

Table 3 (*continued*)

```
CREATE OR REPLACE trigger insertStudent
before insert on Student
for each rowdeclare
excep exception ;
nbTuples number ;
begin
nbTuples :=0;
select count (*) into nbTuples from Etudiant where
NCE=:new. NCE ;
if ( nbTuples != 1) then raise excep ;
else
IF : new .Institution := "ENIT" THEN
INSERT INTO Etudiant_ENIT (NCE,ST_FNAME, ST_LNAME, ADRESS,
Institution,"CLASS", CURSUS) VALUE
( : new . NCE , : new. ST_FNAME,: new. ST_LNAME, : new . adress ,
 new . Institution,: new . "CLASS" ,: new . cursus )
INSERT INTO Student_lib( NCE , Nb_borrow,ST_FNAME, ST_LNAME )
VALUE( : new . NCE , : new. Nb_borrow, : new. ST_FNAME,: new.
ST_LNAME )
ELSIF : new .Institution :="FST" THEN
INSERT INTO Etudiant_FST
(NCE,ST_FNAME, ST_LNAME, ADRESS, Institution
            "CLASS", CURSUS) VALUE
( : new . NCE , : new. ST_FNAME,: new. ST_LNAME, : new . adress ,
 new . Institution,: new . "CLASS" ,: new . cursus )
INSERT INTO Student_lib
( NCE , Nb_borrow,ST_FNAME, ST_LNAME )
VALUE( : new . NCE , : new. Nb_borrow, : new. ST_FNAME,: new.
ST_LNAME )
ELSIF : new .Institution :="FESMT" THEN
INSERT INTO Etudiant_FESMT(NCE,ST_FNAME, ST_LNAME, ADRESS,
Institution "CLASS", CURSUS) VALUE
( : new . NCE , : new. ST_FNAME,: new. ST_LNAME, : new . adress ,
 new . Institution,: new . "CLASS" ,: new . cursus )
INSERT INTO Student_lib( NCE , Nb_borrow,ST_FNAME, ST_LNAME )
VALUE( : new . NCE , : new. Nb_borrow, : new. ST_FNAME,: new.
ST_LNAME )
ELSE
RETURN 'The university name is invalid'
END IF
END IF
when excep then
raise_application_error (−20009 , 'constraint violation' )
    END;
```

To ensure tables update, a reusable procedure may be implemented to handle different table synonyms and attributes.

Table 4 A reusable Oracle update procedure

Oracle generic procedure

CREATE OR REPLACE Procedure update_table (
tbl VARCHAR2,
column VARCHAR2,
value VARCHAR2,
id__column_name VARCHAR2,
id VARCHAR2)
IS
stmt_str VARCHAR2(500);
BEGIN
stmt_str := 'UPDATE ' ‖ tble ‖ '
set ' ‖ column ‖ ' = ' ‖
valeur ‖ ' WHERE ' ‖
id_name ‖ ' = ' ‖ id ;
EXECUTE IMMEDIATE stmt_str ;
END;

This update procedure creates and executes any update on any table where column to update type is VARCHAR2. More specific tuning is to add if the procedure is intended to any column type.

4 Motivation

As described previously, DDB designers are still facing the following issues:

1. DDB design is not an easy task. Multiple criteria must be considered on this sensitive operation: Sites number, exhausting user needs and frequent queries. The designer must establish a compromise between data duplication and performance's cost of update and select queries. He must find out relationship to fragment or duplicate and the update type to consider on each synchronous or asynchronous relationship.
2. DDB implementation is still a heavy task nowadays especially with huge databases called from numerous sites. Existing DDBMS (Bassil, 2012) have several do not have an integrated component which ensures the automatic distribution of the initial centralized database. Actually this implementation, executed manually by designer, have to make a DDB look like a centralize BD, by ensuring a transparency list. This is not yet affordable automatically on current DDBMS.

3. Querying the distributed data must comply with distribution transparency and performance transparency. The querying language and format must be almost the same as for a centralized database. Any syntax review will cause failure for interacting systems with the target database. Thus, a DDB must come with a standard query language that won't need any (or only a few) changes in the client applications side.

The complexity level of the design of a DDB is strongly coupled to the start schema size. In production contexts, it is very common to face very complex schema with hundreds of tables. If the distribution is processed manually through Oracle command line tools, a considerable error probability is expected.

In the following, this work suggests a revision of Oracle DDBMS. The idea is based on extending it by an assistance layer that provides: 1) creation of different types of fragmentation through a GUI for defining different sites geographically dispersed 2) allocation and replication of different databases. The system must automatically generate SQL (Pribyl and Feuerstein, 2001) scripts for each site of the original configuration.

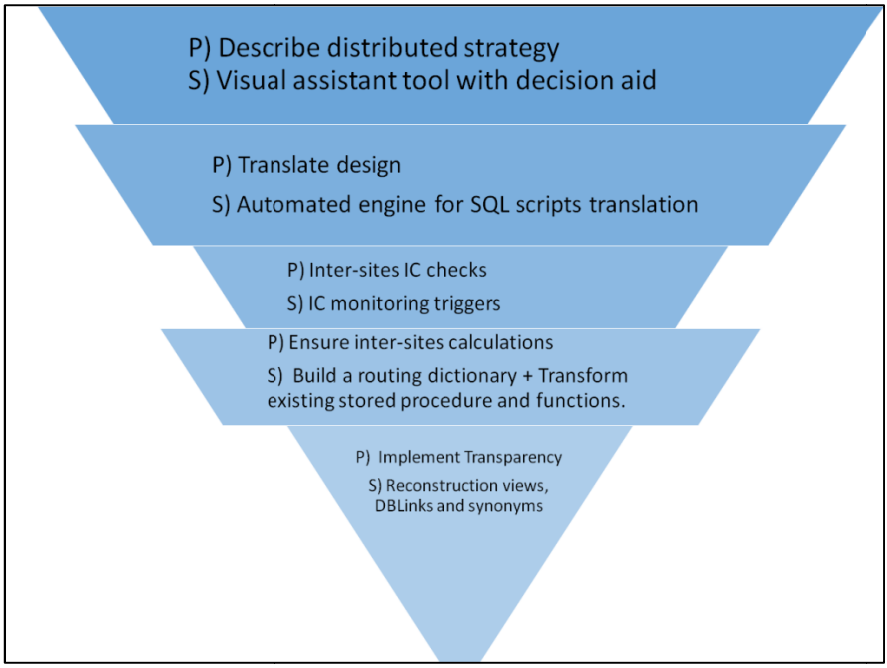

Fig. 2 Main steps to follow to reach the perfect DDB

5 New Architecture for the Description and Manipulation of Large Distributed Data

5.1 Objectives of the Approach

Ideally, the new layer must satisfy the following objectives:

1. Design help for distributed schema: The layer must provide the designer a friendly and productive interface that allows him to represent the draft of the design as a comprehensive and accessible schema to review and collaboration. Fields, tables, sites suggestion lists and work tools (fragmentation and replication) must be provided to designer to ease schema graphical description and avoid additional task complication.
2. Automated implementation of design schema: Once distribution schema is established and validated by the designer along with the wizard assistance, the component "Script generator" must afford the ability to translate accurately described distribution policy to valid SQL scripts. Generated scripts can be directly executed on sites from the layer if access are already prepared, or give deliverable files to transmit to each site administrator.
3. Ensure the integrity constraints and inter-site calculations: The fragmentation horizontal and vertical fragmentation (Nishimura et al., 2013) can corrupt relations starting point for the majority of DBMS (including Oracle) do not handle the constraints of integrities between two distinct sites. It is proposed to substitute the integrity check of the local DBMS triggers that maintain the uniqueness of the primary key between sites, and the validity of the foreign keys between two or more remote fragments
4. Ensure transparency toward the end-user: Creating views that combine data from remote fragments automatically. This reconstruction is an essential constituent of the transparency of the distribution
5. Affording a custom query builder for Large Distributed Databases based on distribution strategy and optimized for migration aims. "No change" in the client side is the perfect strategy. A middleware may be integrated transparently on top of the DDB to translate the centralized target's queries to the distributed context.

5.2 Suggested Layer Architecture

The architecture of the Intelligent DDB, illustrated in Figure 2 depicts the architecture of the implemented layer. The assistance feature is available on the following steps of the distribution process:

1. Access to centralized database to distribute
2. DB link creation
3. Horizontal, vertical and nested fragmentation

4. Fragmentation result validation
5. Data replication

At the end of the process, two options are suggested to execute scripts, depending on afforded preconditions:

1. Automatically: If design environment has a valid access to remote sites, the layer executes scripts on each remote site.
2. Manually: User transfer files using an external tool and takes in charge then execution on remote sites.

The "Universal Connector" is an abstract layer to ensure DBMS abstraction. This is the first step to heterogeneous architecture support. Through this component, the tool can access any database given relative connection string and valid credentials.

The design translation component is responsible of validating and re-writing visual description given through the GUI to a formal SQL scripts. It holds the "Validation Wizard" that processes correctness rules on each step and reports possible failures. "Script Generator" function is dedicated to different SQL variant transcription. As a first release, only Oracle SQL variant (PLSQL) is supported. Further releases are scheduled to support T-SQL and PGPSQL variants. Inside the "Script Generator" component, "MViews Manager" ensures transparency layer by rewriting views to an inter-site transitive format. The integrity check Manager (IC Manager) generates additional procedures to ensure integrity constraints in the

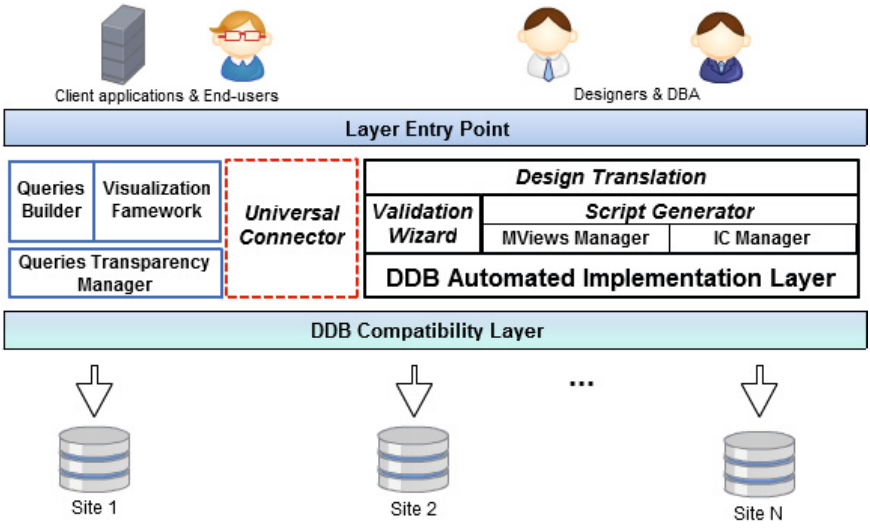

Fig. 3 Layer Architecture

distributed context. In the given sample schema, the IC manager will create a specific procedure to ensure NCE unique constraint by checking the highest attributed primary key across all sites before a student subscription on the database.

The operating principle is as follows:

Table 5 Distribution processing algorithm

```
Algorithm 1 : Distribution Processing
BEGIN
1.      Select database fragment
2.      Enter the list of distribution sites
3.      Retrieve the list of relationships based on initial schema
4.      While " Complete Fragmentation " is false
a.      Select the table fragment
b.      Choose the type of fragmentation(A)
   i.   If Horizontal Fragmentation
        1.      Select column fragmentation
        2.      Affect the value of fragmentation for each site
  ii.   ELSE IF Vertical Fragmentation
        1.      Name the fragment
        2.      Select the columns of the fragment
Select the host site of the vertical fragment
 iii.   If Hybrid Fragmentation
        1.      Treat the hybrid fragment
c.      Validate the resulting fragmentation
d.      Show validation report
e.      If validation is negative,return in (4.b)
f.      If the resulting fragment has foreign keys
            iv.     Perform the derivative fragmentation
4.      END WHILE
5.      For each site
a.      Generate scripts for creating database links from other sites
b.      Generate Scripts fragments of this site
c.      Generate CRUD procedures
d.      Generate materialized views
e.      Write the script in a file name of the site
END FOR
END
```

6 Intelligent-Large Distributed Data

This section is dedicated for the resulting work description and the evaluation of the implemented tool's overall contribution and limits.

6.1 Work Result

Distribution wizard "Intelligent-DDB" is intended to help users graphically distribute a centralized DB, supports the creation of DB links, horizontal, vertical, hybrid and derived fragmentation and replication. The final result is a set of SQL scripts to run on each site.

The figure 4 shows conduct of the assistant.

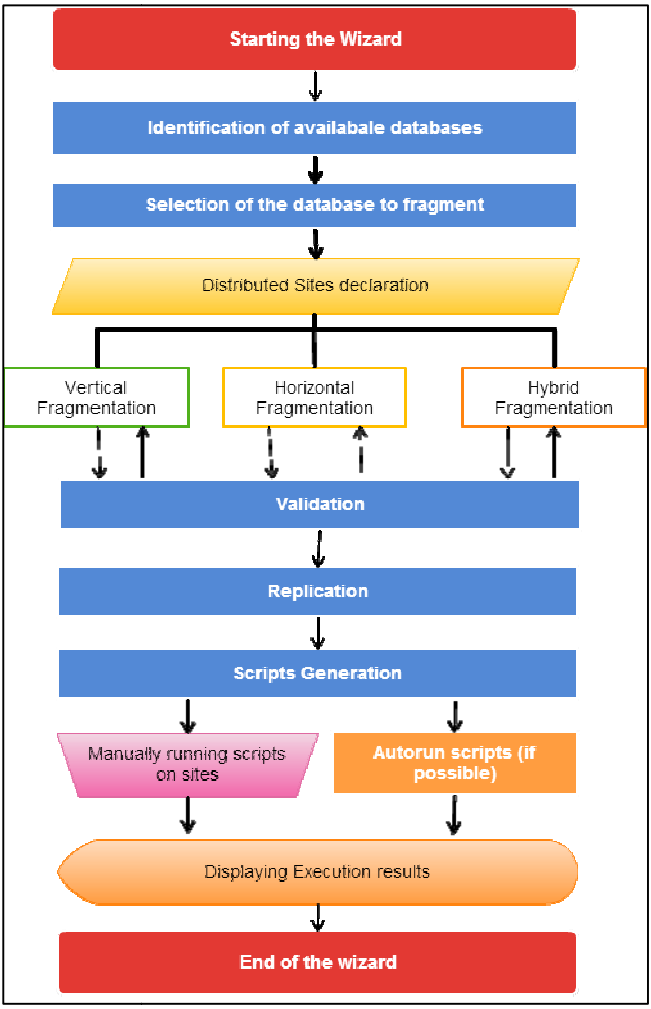

Fig. 4 DDB creation steps

To implement the tool, used operating system is Microsoft Windows Seven. Simulation nodes in network, was made by installing two virtual machines (Oracle Virtual Box) on the chosen host. The apprehended development environment is

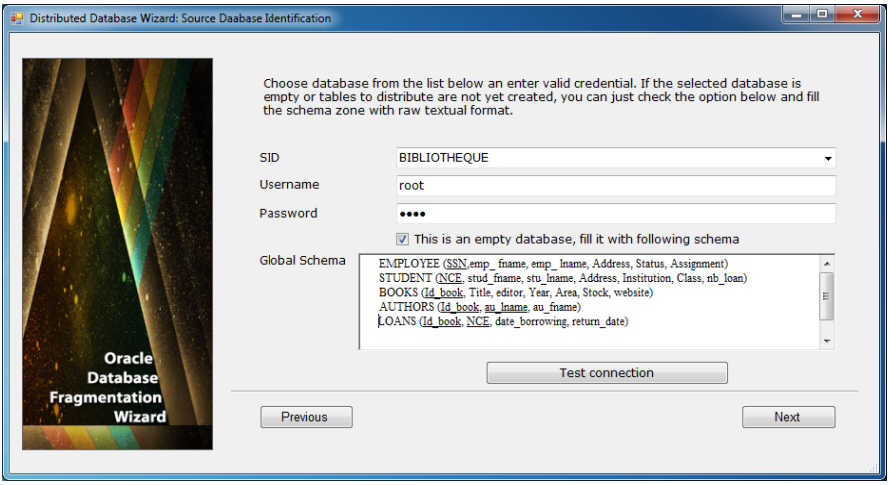

Fig. 5 Authentication screen

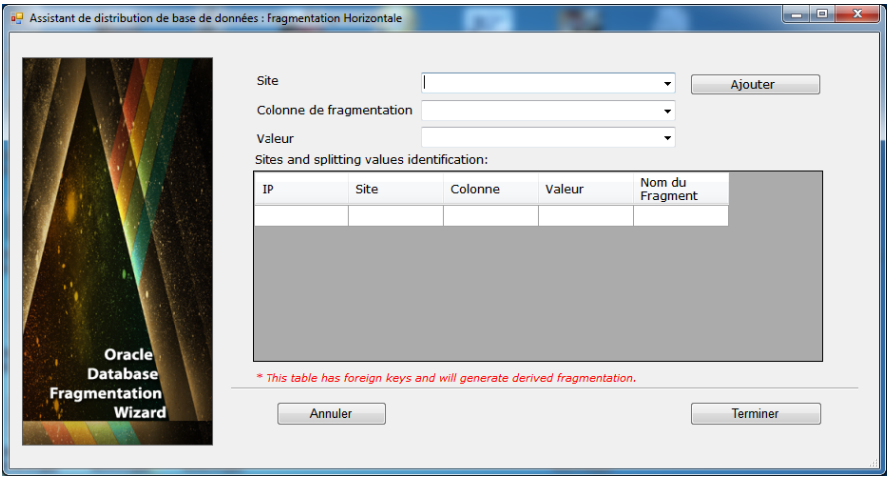

Fig. 6 Sample Horizontal fragmentation screen with implicit derived notification

DotNet framework 4.5 (Rajshekhar, 2013) (CSharp). The figure 5 shows conduct of the assistant. Intelligent-DDB provides designers with multiple screens.

The next section illustrates some examples on the following:

After welcome screen and the tool introduction and interactive help access, user accesses the connection panel to identify target centralized database (Fig.3). If the designer does not want to access existing database or does not yet have the necessary access to it, a standard SDL may be written in the rich text box on the connection panel to describe the target database. The given description must

be compliant with D specification as described in “The First Manifesto”. The given sample in section 4 is a sample of C Date relational description language format.

On successful connection test, next screen is just a popup asking for the number of sites on the distribution. Then, a visual map is displayed with raw nodes. Designer must identify each site with network address (either a name or an ip), a logical name and the DB link name (Fig.4).

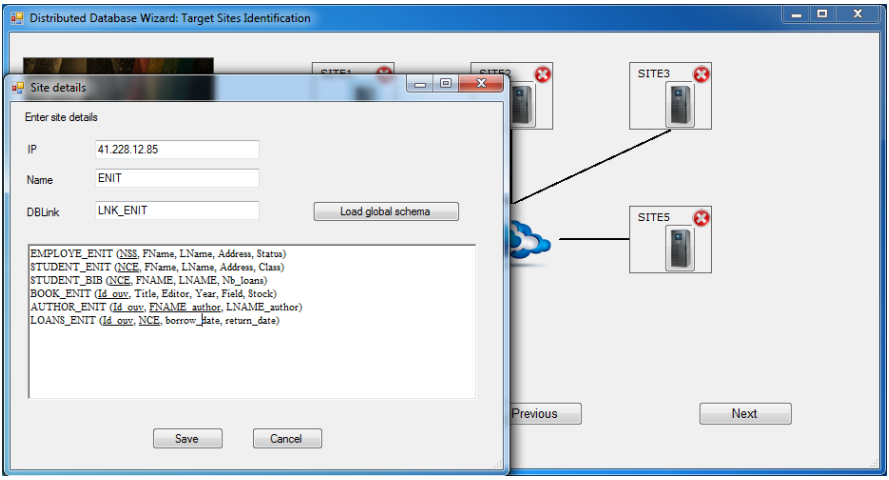

Fig. 7 Sample site details screen

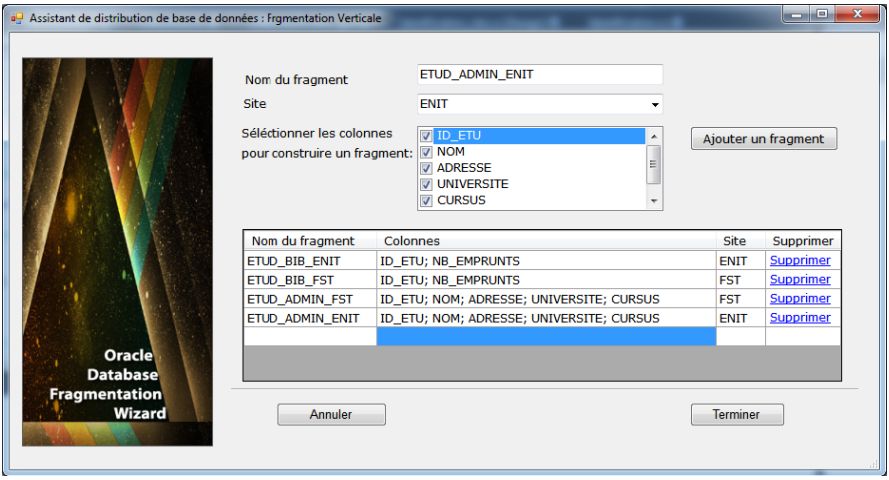

Fig. 8 Vertical fragmentation screen sample

Next step after sites definition is the fragmentation screen. List of accessible tables for the previously defined user is added as an auto complete on the first suggesti list. The second suggestion list suggests fragmentation types (horizontal,

vertical and nested). Derived fragmentation is transparent to designer and is only mentioned for information.

As example, horizontal fragmentation interface provides user with the list of columns of chosen table (Fig.5). User enters fragment name and chooses hosting site and then checks columns related to this fragment (Fig.6). By default the tool keeps the last selection of columns so that the designer can affect the same fragment to multiple sites without redefining the fragment columns. If the designer needs to flush selection, a shortcut on F5 key is linked and functional. In order to split a table horizontally, the designer can follow the dedicated procedure:

- Choose host site from the nodes list
- Choose fragmentation attribute from target table columns
- Choose or manually enter discrimination value
- Click finish to process validation

Dealing with vertical fragments is as easy as horizontal fragmentation because the tool suggests available values necessary for vertical splitting. To create a vertical fragment, the designer can use the following steps:

- Choose host site from the nodes list
- Enter fragment name. This field is pre-filled with a significant fragment name made by parent relation name, chosen host size and a numeric distinct value
- Select the related columns by checking included fields from the parent table columns.
- Click add fragment button

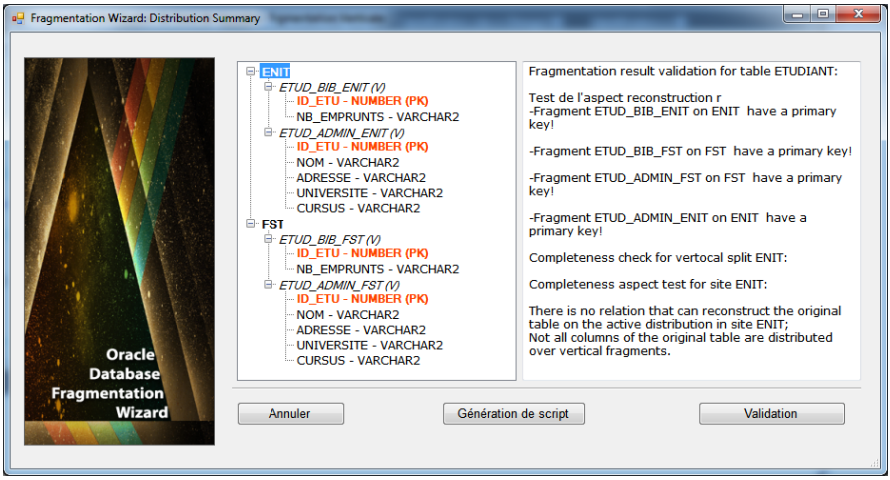

Fig. 9 Sample validation report

To ease designer's task, the tool's selection is persisted after adding the last created object to the fragments summary list. The designer can reuse this fragment

by changing host site and eventually renaming it. Reconstruction and disjointness rules can be validated on the fly as it is only based on primary key and chosen field intersection. The tools alerts about these rules on fragment creation. Completeness aspect cannot be validated on the fly as the complete list of fragments is mandatory to check whether all the components of the start relation where distributed among fragments or a certain data loss is generated by the proposed distribution schema.

Once finished the fragmentation for a table, the wizard starts an automated validation for the described configuration. Adding a fragment without a primary key is already controlled while creating the fragment (on "Add Fragment" button click). Validation screen is displayed then: The left canvas holds a fragment tree with first level nodes as sites, second level nodes as fragment names and leaves are the columns. Primary key is highlighted (orange color). In the right container, the validation report is displayed for the three validation criteria: Reconstruction, completeness then disjointness.

In the end of the whole process, if the policy is validated by the wizard and designer, the tool takes in charge the transcription of visual design into SQL scripts to run on remote sites. The only necessary parameter for this operation is scripts location. Script files naming convention is as follows: [SITE_NAME]_DDB_SCRIPT.sql.

The generation process goes through all sites and generates the script to create symbolic links, then transforms into a standard fragments and commented SQL script

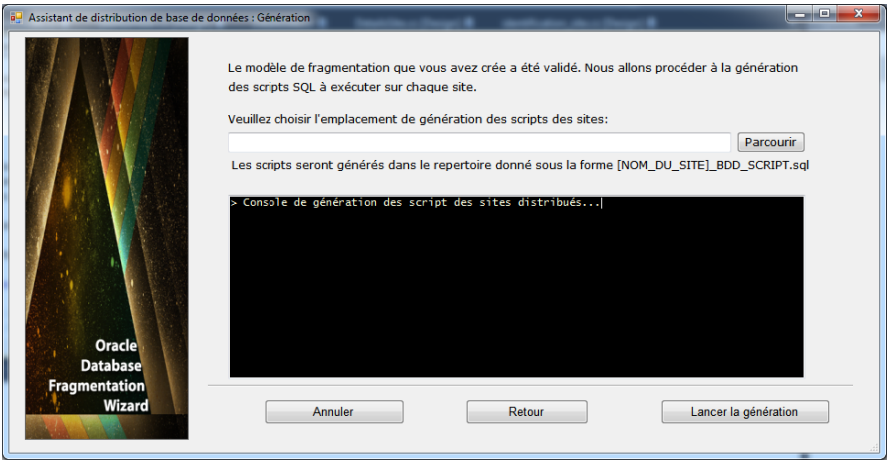

Fig. 10 Scripts generation interface for different sites implementation

The field names and types are consistent with the starting table (same name and same type). Procedures, views, triggers, and the various components are then written accordingly.

Algorithm for generating trigger cross-site verification of IC

The following algorithm describes the method of creating an example of triggers for the integrity constraints control between sites. The implementation of this algorithm in CSharp details are available in the code of the tool created. The algorithm is specifically to the procedures insertion between the horizontal entities fragments. It generates a trigger that performs the verification of the uniqueness of the primary key of a record at all sites before inserting. If the primary key is unique, it proceeds to insert in the next column and the criterion of horizontal fragmentation suitable site.

BEGIN
1. Write trigger header insertion entity
2. For each site:
a. Write the query SELECT ID of the entity (given as parameters) in a number_id variable.
b. Write if the condition is greater than 0 number_id
c. Write RAISE EXCEPTION statement to indicate the insertion error
3. For all sites:
a.Write IF condition VALEUR_COLONNE_FRAGMENTATION_DU_SITE_X = VALEUR_ELEMENT_A_INSERER
b. If the script is being generated for the site X
i. Write the INSERT statement in the table in question to the website base X otherwise
ii. Write the INSERT statement in the synonym table in question
4. Write the return code and the exception message of uniqueness
5. Close the trigger
END.

The main aim of this algorithm is to perform a correct and optimized fragmentation. The client part of the works intended to enhance the transparencies series. Building queries to operate on distributed databases and afford an optimal time response is a very sensitive part of a distributed database creation. In Large Databases, fetching queries can cancel all the previous work on data distribution if built without distributed context consideration.

```
CREATE OR REPLACE PROCEDURE "LOAN_BOOK"
( FNameST IN VARCHAR2,
LNameST IN VARCHAR2
, book IN VARCHAR2
) IS
ID_st NUMBER;
Cou NUMBER;
i VARCHAR2(100);
NB_loan NUMBER;
ADRESS VARCHAR2(100);
ID_book NUMBER;
TITLE_Book VARCHAR2(100);
NB_stock BOOK.STOCK%TYPE;
BEGIN
  Cou :=0;
    select NCE into ID_st from STUDENT  where ST_FNAME= FNameST and
ST_LNAME=LNameST ;
  select NB_loan into NB_laon from STUDENT_bib  where NCE = ID_st;
  DBMS_OUTPUT.PUT_LINE( 'NB_EMPRUNTS -> ' || NB_PRETS ) ;
  commit;
  DBMS_OUTPUT.PUT_LINE( 'BOOK  -> ' || book ) ;
  select ID_book into ID_BOOK from BOOK  where title=book;
  DBMS_OUTPUT.PUT_LINE( 'ID_book -> ' || ID_book || ' -  IN STOK ' ||
nb_stock) ;
  commit;
  if( nb_stock >0) THEN
    insert into loans (ID_BOOK, NCE, LOAN_DATE, RETURN_DATE)
values (ID_book,ID_st,sysdate,'',SYSTIMESTAMP);
    update student_biblio  set nb_loan= nb_loan+ 1 where NCE= ID_std
    update book  set stock = nb_stock - 1 where id_book=ID_book;
  else
    dbms_output.put_line('NO COPY N IS AVAILABLE AT THIS TIME.');
  end if;
  commit;
EXCEPTION
   WHEN NO_DATA_FOUND THEN
   BEGIN
      /* la cle n'existe pas -> INSERTION */
      DBMS_OUTPUT.PUT_LINE( 'EX ID_book -> ' || ID_book);
   END;
END;
```

The idea of the algorithm for generating stored procedures is quite simple. When creating sites scripts for websites, the tool generates the materialized views that reconstruct the tables out. Oracle materialized views can be updated instead of conventional views based on this pillar of transparency to make selections and

updates. However, the detection of these cases requires the storage of such information in a temporary table (accessible during generation): The X fragment Site S is from the MX table. Stored procedures start site are adapted to the context distributed via materialized views in Oracle. In the script example, the materialized view "etudiant_prets_global_view ' replaces the student table base.

The generation process generates implicitly a configuration table that holds the distribution strategy information. This table is built on the generation processing algorithm of the tool and describes each fragment by the following attributes:

1. Fragment ID: The unique key of each fragment generated by a serial sequence
2. Fragment Name: The fragment name entered on the fragmentation step.
3. Fragment Site: The database link name pointing the location site.
4. Fragment Parent: The parent table of the current fragment on the centralized database.
5. Fragment Type: This flag indicates whether this is a horizontal fragment or a vertical one.
6. Fragment Columns: This field holds a comma separated values of column names contained on the current fragment.
7. Fragment Split Column: If the current table is a horizontal fragment this attribute contains the column on which splitting was executed.
8. Fragment Split Condition: If the current table is a horizontal fragment this attribute contains the condition on which splitting was executed.
9. Fragment Mirrors: This field holds a comma separated list of the current fragment synonyms over the distribution sites.

The proposed structure of the fragments dictionary is very useful on procedures and triggers generation as well as queries routing issues. The routing based on horizontal each split criteria improves querying time as it forward queries directly to data location and avoid extra queries to identify the site holding the information.

The mirrors attribute on the dictionary describes the list of duplicate fragments. Based on this attribute, if the current fragment is updated, and the mirrors attribute holds a non empty list, updates will occur on each copy of this fragment. This behavior enhances real-time data synchronization and improves data consistency among a DDB.

The client part of the work is built as a rich web client to make querying accessible for all users with a visual preview of the distribution structure. The idea behind this interface is to take the control of the query transformation in order to protect the performance indicator of the distribution and ensure a certain level of transparency to the end-user.

The query building feature included on the DDBMS Smart Query interface. It displays the list of tables, views and materialized views as if they belong to same centralized database. On target object's selection, a multi-state panel is shown on the query builder space. The default state of the panel is an overview of the object

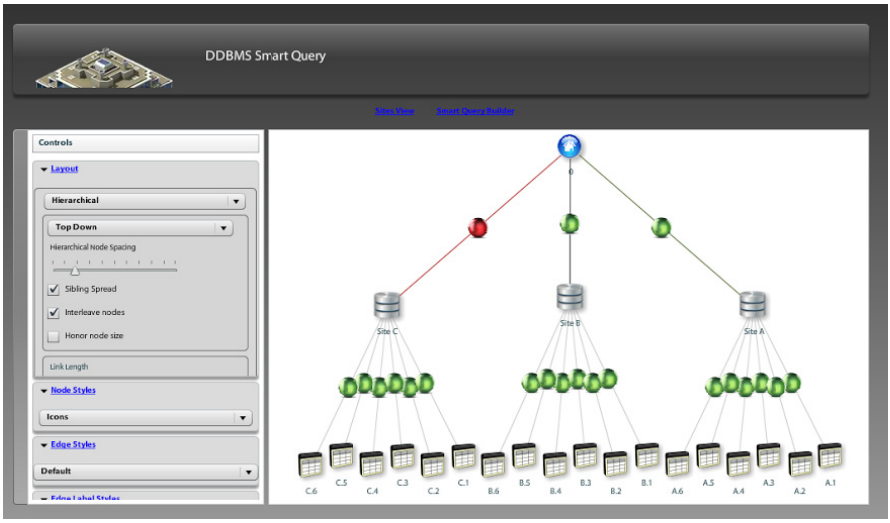

Fig. 11 Visual map of the distribution on end-users screen

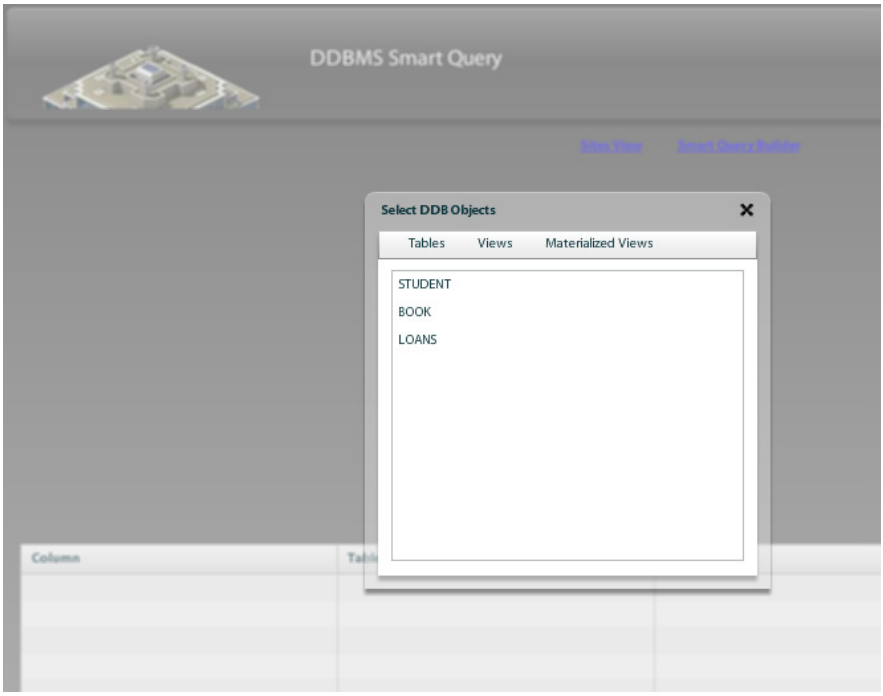

Fig. 12 Visual query assistant

structure (field or column names). The second state describes more details about field's types and constraints. The third state displays a sample of data held in this object. This may be useful for end-users to check the field to include in the result preview if he is not sure about the column content (given just its name). To create a query, the end user may choose one or more objects to query. The user chooses then fields to display in the query's result and restriction criterion. The criteria field is a free text field. If the condition matches a fragment split condition (C1) and all required fields (C2) are included in this fragment then the query is directly sent to the target site. This implicit routing is done by auto-generated procedures in the DDB implementation process. The condition of C2 is mandatory because in a nested fragment, the split condition is kept for the horizontal splitting step, but the same fragment may be re-fragmented vertically and then one or more columns may be missing from the user requested fields.

6.2 Result Evaluation

The suggested tool is a serious candidate to be appended to current Oracle partitioning solution. Until latest release, Oracle 12c, the vendor seems to ignore the importance of a graphical tool to directly execute a distribution schema. Only few console commands are given to designers in order to execute their distribution policy draft. The tool offers a set of suggestions and pre-loaded lists along with live result's validation. The ability to enter an object schema based on standard schema description language (SDL) enhances abstraction level and contributes to the tool's evolution towards heterogeneous environments. The validation report gives the designer a sharp reference to fix the distribution strategy in order to avoid data loss and useless duplicate information.

The distribution dictionary embedded on each site improves the response time for queries holding split criteria on their where clause and addresses the query directly to the site where the target data are located. For example, fetching a student subscribed on institution FST will be redirected from any site to site FST based on the fragments dictionary as it keeps in memory the split criterion which is INSTITUTION='FST'. In the same way, a vertical fragment are invoked directly if a query requests and includes only a subset of columns held in one fragment by comparing the given list of columns to the columns flat list in the fragment's dictionary.

The work result main lacks are composite keys handling, many-to-many relations implicit treatment process. The integration of a suggestion expert system based on start schema and centralized database statics is so far a good feature to enhance the tool's assistance feature. Rim's (Moussa, 2011) work may be integrated in the lifecycle of the created tool in order to complete the assistant intelligent aspect. Distribution result simulator is a very helpful feature to integrate in order to give designers more decision criteria to choose the best distribution policy.

7 Conclusion

The big data phenomenon has already changed the IT development trends. Despite the great investment that market main actors are doing, many issues remain unsatisfied. Mining accuracy in critical systems is always compromised by the NoSql trend. Thus, the data integrity aspect in relational databases kept RDBMS as a considerate alternative in big data processing. While centralized RDBMS has shown many shortcomings especially in performance considerations, the DDBMS seems to solve the performance issue thanks to its native parallelism. Meanwhile, DDBMS still suffer from many lacks such as design and implementation tools in the creation step and a comprehensive standard querying language and APIs to avoid client applications rewriting when migrating a centralized database to the distributed context.

The discussed issue in this paper is the funding of a correct way to remedy to such a lack in DDB design and implementation with a full compliance to distribution laws and rules. The distribution correctness is a fatal factor in the migration process as a wrong strategy implementation may result to performance and data loss issues. Thus, an assistance layer to help designers create DDB is a necessity. The so presented approach was designed and implemented as a middleware with an entry point and a standard connector to facilitate its integration inside existing applications. The implemented visual assistant made possible for designers to quickly create the different sites' scripts and ensure their design consistency against correctness rules. Client applications and end-users also can profit from this layer by integrating the query translator instead of rewriting any existing application code. The query translation component uses the same queries as a centralized database, and transfers them to the distributed context. This component is based on the model dictionary created in the first step by the "DDB automated implementation layer". The processing acceleration provided by this component is a first step to solve performance consideration issues in Big Data world with distributed relational databases.

The implemented layer has shown an acceptable performance improvement in distributed database processing life-cycle. A few reviews such as handling duplicate fragments synchronization and smart data indexing may be implemented on Big Data benchmarks.

References

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop GIS: a high performance spatial data warehousing system over Mapreduce. Proceedings of the VLDB Endowment 6(11), 1009–1020 (2013)

Bassil, Y.: A Comparative Study on the Performance of the Top DBMS Systems. Journal of Computer Science & Research 1(6), 20–31 (2012)

Chen, M., Mao, S., Zhang, Y., Leung, V.C.: Big data applications. In: Chen, M., Mao, S., Zhang, Y., Leung, V.C.M. (eds.) Big Data, pp. 59–79. Springer (2014)

Dede, E., Fadika, Z., Govindaraju, M., Ramakrishnan, L.: Benchmarking Mapreduce implementations under different application scenarios. In: 2011 12th IEEE/ACM International Conference on Grid Computing (GRID), Lyon, September 21-23, pp. 28–31 (2014), doi:10.1109/Grid.2011.21

Embley, D.W., Liddle, S.W.: Big Data—Conceptual Modeling to the Rescue. In: Ng, W., Storey, V.C., Trujillo, J.C. (eds.) ER 2013. LNCS, vol. 8217, pp. 1–8. Springer, Heidelberg (2013)

Etemad, M., Küpçü, A.: Transparent, distributed, and replicated dynamic provable data possession. In: Jacobson, M., Locasto, M., Mohassel, P., Safavi-Naini, R. (eds.) ACNS 2013. LNCS, vol. 7954, pp. 1–18. Springer, Heidelberg (2013)

Giese, M., et al.: Scalable end-user access to big data. In: Akerkar, R. (ed.) Big Data Computing, pp. 205–245. CRC Press, New York (2013)

Goswami, S., Kundu, C.: Xml based advanced distributed database: implemented on library system. International Journal of Information Management 33(1), 389–399 (2013)

Hewitt, E.: Cassandra: The Definitive Guide. Distributed Data at Web Scale. O'Reilly Media (2010), http://it-ebooks.info/book/623/

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N.: Tastes, ties, and time: A new social network dataset using facebook.com. Social Networks 30(4), 330–342 (2008)

Manyika, J., et al.: Big data: The next frontier for innovation, competition, and productivity, pp. 34–67. McKinsey Global Institute (2011)

Moussa, R.: DDB Expert: A Recommender for Distributed Databases Design. In: Database and Expert Systems Applications (DEXA), Toulouse, August 29-September 02, pp. 534–538 (2011), doi:10.1109/DEXA.2011.25

Nishimura, S., Das, S., Agrawal, D., El Abbadi, A.: MD-hbase: Design and implementation of an elastic data infrastructure for cloud-scale location services. Distributed and Parallel Databases 31(2), 289–319 (2013)

Özsu, M.T., Valduriez, P.: Principles of distributed database systems, 3rd edn. Springer, New York (2011)

Patterson, T.: Information integrity in the age of big data and complex information analytics systems. EDPACS 48(6), 1–10 (2013)

Polo, J., Carrera, D., Becerra, Y., Torres, J., Ayguadé, E., Steinder, M., Whalley, I.: Performance-driven task co-scheduling for Mapreduce environments. In: Network Operations and Management Symposium (NOMS), San Diego, April 19-23, pp. 373–380. IEEE (2010), doi:10.1109/ICPP.2010.73

Pribyl, B., Feuerstein, S. (eds.): Learning Oracle PL/SQL. Oracle Development Languages, 6th edn., pp. 128–180. O'Reilly Media (2001)

Pukdesree, S., Sukstrienwong, A., Lacharoj, V.: Performance Evaluation of Distributed Database on PC Cluster Computers. WSEAS Transactions on Information Science and Applications 10(1), 21–30 (2006)

Rajshekhar, A.: .Net Framework 4.5 Expert Programming Cookbook, 3rd edn. Packt Publishing Ltd. (2013)

Sakr, S., Liu, A.: The family of map-reduce. In: Gkoulalas-Divanis, A., Labbi, A. (eds.) Large-Scale Data Analytics, pp. 1–39. Springer, New York (2014)

Shute, J., Oancea, M., Ellner, S., Handy, B., Rollins, E., Samwel, B., Vingralek, R., Whipkey, C., Chen, X., Jegerlehner, B., Littleeld, K., Tong, P.: F1: the fault-tolerant distributed rdbms supporting google's ad business. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, New York, May 19-22, pp. 777–778 (2012), doi:10.1145/2213836.2213954

Silberschatz, A., Korth, H.F., Sudarshan, S.: Database system concepts, vol. 4. McGraw-Hill, New York (2002)

Soares, J., Lourenço, J., Preguiça, N.: MacroDB: Scaling Database Engines on Multicores. In: Wolf, F., Mohr, B., an Mey, D. (eds.) Euro-Par 2013. LNCS, vol. 8097, pp. 607–619. Springer, Heidelberg (2013)

Krishnan, S., Baru, C., Crosby, C.: Evaluation of MapReduce for Gridding LIDAR Data. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), Indianapolis, November 30-December 3, pp. 33–40 (2010), doi:10.1109/CloudCom.2010.34

Stinson, A., Chandramouly, K.: Enabling big data solutions with centralized data management. IT@Intel White Paper, 7 pages (2013)

Stonebraker, M.: Concurrency control and consistency of multiple copies of data in distributed Ingres. Berkeley Workshop 5(3), 235–258 (1979)

Tosun, U., Dokeroglu, T., Cosar, A.: Heuristic algorithms for fragment allocation in a distributed database system. In: Gelenbe, E., Lent, R. (eds.) Computer and Information Sciences III, pp. 401–408. Springer, London (2013)

Wang, X.: Research of data replication on cluster heterogenous database. In: Liu, X., Ye, Y. (eds.) Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, pp. 249–260. Springer, Berlin (2014)

Wu, L., Barker, R.J., Kim, M.A., Ross, K.A.: Navigating Big Data with high-throughput, Energy-efficient data partitioning. SIGARCH Computer Architecture News 41(3), 249–260 (2013)

Author Index