IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 73-78
www.iosrjournals.org

Recent Trends and Rapid Development of Applications In Data Mining

Sadia Patka[1], M. S. Khatib[2], Kamlesh Kelwade[3]
[1,2]*Asst. Professor, Computer Science and Engineering, Anjuman College of Engineering and Technology, India*
[2]*H.O.D., Computer Science and Engineering, Anjuman College of Engineering and Technology, India*

ABSTRACT : *The development of Information Technology has generated enormous amount of databases and huge data in various areas. Loose coupling is adapted in Data Mining System, since it can fetch any portion or part of the data which is stored in database by more flexibility and in efficient manner. Therefore the Data mining system can be classified according to the kinds of databases and knowledge mined and also the techniques used or the application adapted. The traditional method is used to analyse data manually for patterns for the extraction of knowledge. In Banking, Health care, marketing, Science there will be a data analyst to work with data and scrutinizing the final role of decisions. This work is done by Data Mining. Data mining application can be generic or domain specific. It allows reusability of information in a feasible way and finally it makes possible to build large and scalable system. Applications of Data mining in computer security are designed to meet the needs of professionals such as researchers and practitioners in different fields. This paper gives the overview of Data mining system and few of its applications. Data mining is becoming a technology in activities as diverse as using large amount of historical data to predict the success of marketing.*
KEYWORDS: *Data Mining, Intrusion Detection, Predictive Data mining, clustering, e-commerce, web mining, Business intelligence*

I. INTRODUCTION

To generate information massive collection of data is required. The data can be simple like numerical data, figures and text documents, to more complex such as spatial data, multimedia data and hypertext documents. With large amount of data stored in databases, files, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge and patterns that could help in Decision making. Data mining is a set of activities or tool used to find new, hidden or unexpected patterns in data or unusual patterns in data. This paper study data mining related applications to draw the concepts and characters and then propose a selection model to meet the requirements to data mining categories.

II. RELATED WORK

Data mining (DM), also called Knowledge-Discovery [1] is one of the hot topics in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data. Mining can efficiently discover useful and interesting knowledge from large collection of data. It is a fairly recent topic in computer science [2] but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition as shown in Fig.1. Data mining is disciplines [3] works to finds the major relations between collections of data and enables to discover new and meaningful patterns.

As data mining matures new and increasingly innovative applications for it to emerge, in this paper the applications of data mining are categorized in the following way - Intrusion Detection, Health Care, Business, and Biological Data Analysis.

In the medical and health care areas, because of regulations and due to availability of computers, it is possible to make large amount of data available. Such a large amount of data cannot be analyzed and processed by humans in a short period of time to make diagnosis, prognosis and also to make treatment schedules. This problem is overcome with the help of

data mining so the applications of data mining in this field results timely and accurate decisions.

In business, bank and financial institutions offer a wide variety of services so the information collected is said to be complete, reliable and high quality which need data mining to provide security to help in fraud detection.

All the companies, organization's ultimate aim is to develop and expand their business to the extreme and to increase their turnover and goodwill in the society. To achieve this, it is mandatory to redesign their business process which suits to the current trends. In today's business world, there is an enormous amount of available data and a great need to make good use of it. At first, data must be organized by database tools and data warehouses, and then it needs an instrument for knowledge discovery. For the effective knowledge discovery, we implement or depend on Data Mining technology.

In the biomedical research ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome by discovering the large scale sequencing patterns and gene functions. So the data mining helps to do DNA analysis for the discovery of genetic causes for many diseases with the help of large databases.

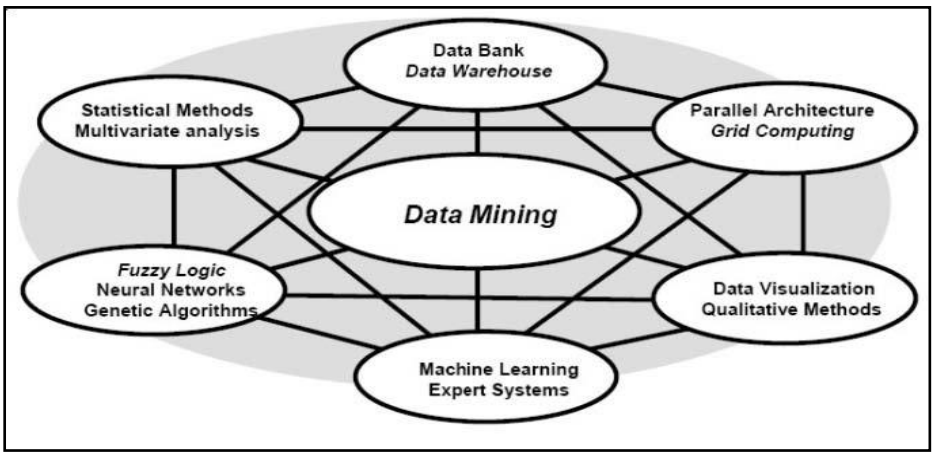

Fig.1 Data Mining in various fields

A provenance system can support a number of uses in derivation of history of a data product starting from its original sources. Thereby it helps to estimate the data quality, data reliability, based on the source data and transformations and also provide proof statements on data derivation. It reduces errors in data detection.

Intrusion Detection System (IDS) [4] is used as a countermeasure to preserve data integrity and system availability from attacks. Intrusion Detection Systems is a combination of software and hardware that attempts to perform intrusion detection. Intrusion detection is a process of gathering knowledge related to attacks occurring in the process of monitoring the events and analyzing them for sign or intrusion. The system raises the alarm when a possible intrusion occurs in the system. The network data source of intrusion detection consists of enormous amount of textual information, which is difficult to comprehend and analyze. Many IDS can be described with three fundamental functional components – Information Source, Analysis, and Response. Different sources of information and events based on information are gathered to decide whether intrusion has taken place. This information is gathered at various levels such as system, host, application, etc.

III. DATA MINING APPLICATIONS AT A GLANCE

3.1 Intrusion Detection based on Data Mining

Data Mining is becoming one of the popular techniques for detecting intrusion or attacks in network. Intrusion detections can be classified [4,5] on the basis of their strategy of detection – Anomaly detection and Misuse detection. Data mining is one of the technologies that can be applied to intrusion detection to determine new pattern from the massive network data as well as to reduce the stains of the manual complications of the intrusion. Data Mining is helpful in detecting new vulnerabilities and intrusions, discover previous unknown patterns of attacker behaviours and provide decision support for intrusion management. Data Mining frequently used to analyze network data to gain intrusion related knowledge are Clustering, Classification, Outlier detection and Association rule.

In Clustering, grouping is accomplished by finding similarities between data according to characteristics found in the actual data. The largest cluster is found and other clusters are sorted and then the outliers are detected. Classification is a supervised learning concept. In Classification decision tree, Rule Based methods can be used to detect the attackers. Association rule mining is one of the most important and well researched techniques of data mining. It aims to extract the interesting correlations, frequent patterns that occur, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas like telecommunication networks, market and risk management, and inventory control. The main motivation resides behind data mining in intrusion detection is automation. Fig. 2 shows the Architecture for IDS. Pattern of the normal behaviour and pattern of the intrusion can be computed using data mining. Data mining is the modern technique for Intrusion Detection. Data mining approach can be contributed significantly in the attempt to generate an effective Intrusion Detection System.

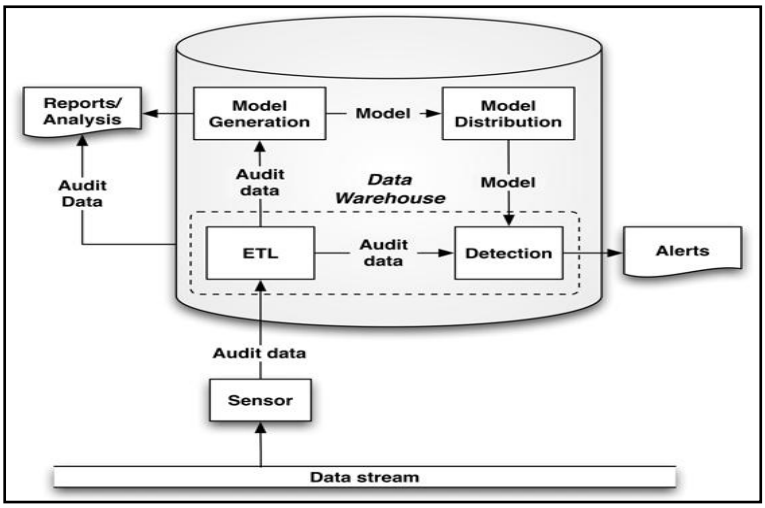

Fig.2 Architecture for Intrusion Detection System

3.2 Diabetes Prediction using DataMining

Data mining in health care management is unlikethe other fields owing to the fact that the data present are heterogeneous and that certain ethical, legal, and social constraints apply to Private medical information. Health care related data are voluminous in nature and they arrive from diverse sources all of them not entirely appropriate in structure or quality. These days, the exploitation of knowledge and

experience of numerous specialists and clinical screening Data of patients gathered in a database during the diagnosis procedure, has been widely recognized.

Diabetes is a major health problem and the most common endocrine disease across all population and age groups. This disease has become one of the major leading cause of death in developed countries and there is substantial evidence that it is reaching epidemic proportions in many developing and newly industrialized nations. Fig.3 depicts the Data Mining Architecture for Diabetes.

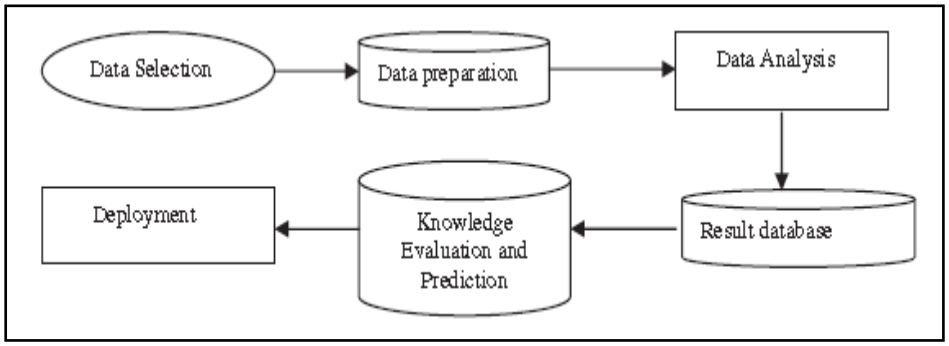

Fig.3 Data Mining Architecture for Diabetes

Data Mining aims at discovering knowledge out of data and presenting it in a form that can be easily understandable to humans. Advanced data mining techniques are used to discover useful knowledge in database and for medical research, particularly in Heart disease prediction. The analysed prediction systems for Heart disease uses more number of input attributes. It uses medical terms such as sex, blood pressure, cholesterol, obesity, smoking and etc. to predict the likelihood of patient getting a Heart disease. The data mining classification techniques such as Decision Trees, Naive Bayes, and Neural Networks are analysed on Heart disease database.

3.3 Use of Data Mining in E-Commerce and Business Intelligence (BI)

E-commerce has changed the face of most business functions in competitive enterprises. Internet technologies have seamlessly automated interface processes between customers and retailers, retailers and distributors, distributors and factories, and factories and their myriad suppliers. In general, e-commerce and e-business have enabled on-line transactions. Also, generating large-scale real-time data has never been easier. With data pertaining to various views of business transactions being readily available, it is only apposite to seek the services of data mining to make (business) sense out of these data sets.

Data mining is about finding useful patterns in data. The patterns that are discovered by data mining are useful because they extend existing business knowledge in useful ways. But new business knowledge is not created "in a vacuum"; but it builds on existing business knowledge, and this existing knowledge is in the mind of the business expert. The business expert therefore plays a critical and vital role in data mining, both as an essential source of input (business knowledge) and as the consumer of the results of data mining, The business expert not only, uses the results of data mining but also evaluates them, and this evaluation should, be a continual source of guidance for the data mining process. Data mining reveals meaningful patterns in data, but only the business expert can judge their usefulness. Fig.4 shows the Architecture for an effective use of DM in BI.

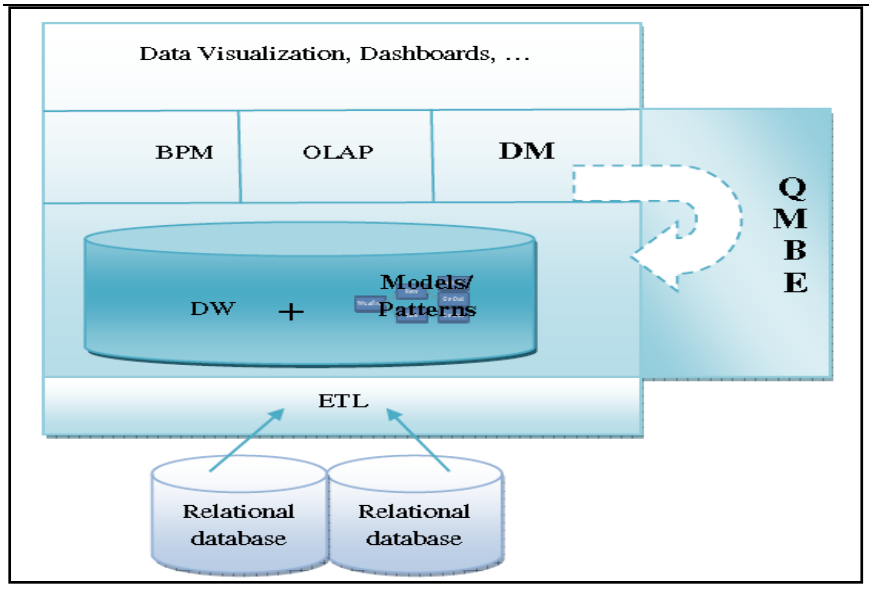

Fig.4 Architecture for an effective use of DM in BI

It is very important to know that the data is not the business, but only a dim reflection of it. The Patterns found by mining data may fail to be useful for many different reasons. The patterns may reflect properties of the data, which do not represent reality at all, for example when an artifact of data collection, such as the time a snapshot is taken, distorts its reflection of the business. Alternatively, the patterns found in the data may be true reflections of the business, but they merely describe the problem that data mining was intended to solve - for example arriving at the conclusion that "purchasers of this product have high incomes" in a project to market the product to a broader range of income groups. If the business knowledge is insufficiently informed, Data Mining may produce useless results for reasons like the above.

3.4 Data Mining in Bioinformatics

The entire human genome and the complete set of genetic information within each human cell has now been determined. Understanding these genetic instructions allow the scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological processes such as growth and ageing and to clearly track our evolution and its relationship with other species. The key obstacle lying between investigators and the knowledge they seek is the sheer volume of data available. This is evident from the following figure which shows the rapid increase in the number of base pairs and DNA sequences in the repository of GenBank.

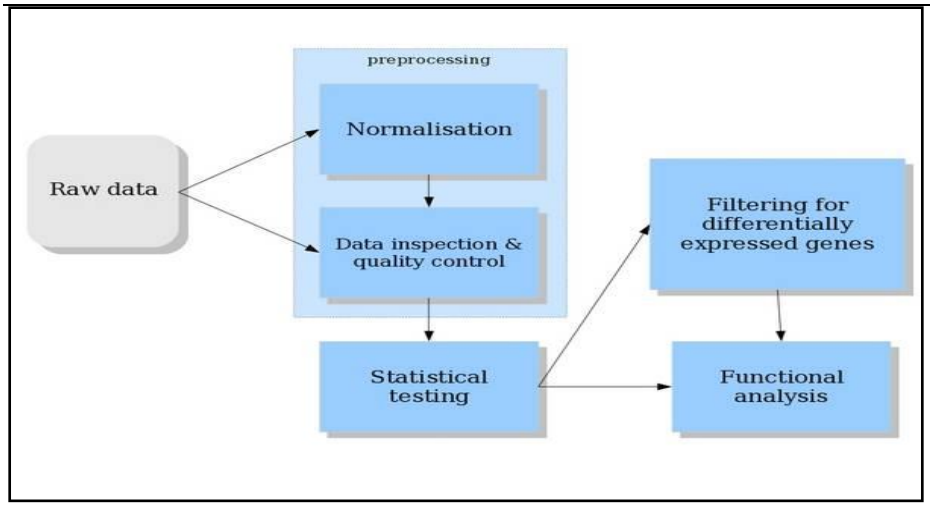

Fig.5 Data Mining in Bioinformatics

Biologists, like scientists are trained primarily to collect new information. Until, biology lacked the tools to analyze massive repositories of information such as the human genome database. Fortunately, the discipline of computer science has been developing methods and approaches which is well suited to help biologists to manage and analyze the incredible amounts of data that promise to profoundly improve the human condition. Data mining is one such technology.

IV. CONCLUSION

Data Mining is not a new term and application, but in the recent years its growth touches great horizons. It has spread its wings in almost all areas nowadays. This paper describes its applications in few areas namely Intrusion Detection, Diabetes Prediction, Business process development and Biological data analysis. It is clear that Data mining tools helps in extracting meaningful knowledgeable attributes from the unimaginable massive data.

REFERENCES

[1] Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.
[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2,JohnWiley & Sons, Inc, 2005
[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education,New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.
[4] Kapil Wankhade, Sadia Patka, Ravindra Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques", In Proceedings of 2013 International Conference on Communication Systems andNetwork Technologies , IEEE,2013, pp.626-629.
[5] Kapil Wankhade, Sadia Patka, Ravindra Thool, "An Efficient Approach for Intrusion Detection UsingData Mining Methods", In Proceedings of 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE,2013, pp.1615-1618.
[6] Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Enginebased on Internet, Emerging Technologies and Factory Automation," Proceedings ETFA '03, IEEEConference, 16-19 Sept. 2003.
[7] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery inDatabases," AI Magazine, American Association for Artificial Intelligence, 1996.
[8] Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process",Working Paper of the Center for Digital Economy Research, New York University and also presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.
[9] Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent:An Application to Spatial Data, volume 5, ISSN 1307-6884," Proceedings of World Academy of Science,Engineering and Technology, April 2005.