Data Mining Application and Challenges

Shikha Dhoriyani[1]

[1] *B E Student, CE Department, Saffrony institute of technology, Gujarat, India*

ABSTRACT

*In this paper I have focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. As we are aware that many **MNC's** and large organizations are operated in different places of the different countries. Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions .The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today's business environment, efficiency or speed is not the only key for competitiveness. This type of huge amount of data's are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. This paper imparts more number of applications of the data mining and als**o** o focuses scope of the data mining which will helpful in the further research.*

Keyword: - *Data mining , application, Challenges*

1. INTRODUCTION

Data mining is a *process* that takes data as input and outputs knowledge. One of the earliest and most cited definitions of the data mining process, which highlights some of its distinctive characteristics, is provided by Fayyad, Piatetsky-Shapiro and Smyth, who define it as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Note that because the process must be non-trivial, simple computations and statistical measures are not considered data mining. Thus predicting which salesperson will make the most future sales by calculating who made the most sales in the previous year would *not* be considered data mining. The connection between "patterns in data" and "knowledge" will be discussed shortly. Although not stated explicitly in this definition, it is understood that the process must be at least partially automated, relying heavily on specialized computer algorithms (i.e., data mining algorithms) that search for patterns in the data.

It is important to point out that there is some ambiguity about the term "data mining", which is in large part purposeful. This term originally referred to the algorithmic step in the data mining process, which initially was known as the Knowledge Discovery in Databases (KDD) process. However, over time this distinction has been dropped and data mining, depending on the context, may refer to the entire process or just the algorithmic step. This entire process, as originally envisioned by Fayyad, Piatetsky-Shapiro and Smyth.

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection . a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves

effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

2. Application and Advancement in VR

a. Data Mining Applications in Healthcare

Data mining applications in health can have tremendous potential and usefulness [60]. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications

b. Future Directions of Health care system through Data Mining Tools

Data mining technique is used in MBA(Market Basket Analysis).When the customer want to buying some products then this technique helps us finding the associations between different items that the customer put in their shopping buckets. Here the discovery of such associations that promotes the business technique .In this way the retailers uses the data mining technique so that they can identify that which customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items.

c. The data mining is used an emerging trends in the education system in the whole world

In Indian culture most of the parents are uneducated .The main aim of in Indian government is the quality education not for quantity. But the day by day the education systems are changed and in the 21^{st} century a huge number of universalities are established by the order of UGC. As the numbers of universities are established side by side, each and every day a millennium of students are enrolls across the country. With huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, and increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes. In this current era we are using the KDD and the data mining tools for extracting the knowledge this knowledge can be used for improving the quality of education .The decisions tree classification is used in this type of applications.

D. Business & E-commerce Data. Back-office, front-office, and network applications produce large amounts of data about business processes. Using this data for effective decision making remains a fundamental challenge. b. Scientific, Engineering & Health Care Data. Scientific data and meta-data tend to be more complex in structure than business data. In addition, scientists and engineers are making increasing use of simulation and of systems with application domain knowledge. c. Web Data. The data on the web is growing not only in volume but also in complexity. Web data now includes not only text and image, but also streaming data and numerical data. In this section, we describe several such applications from each category. Business Transactions. Today, businesses are consolidating and more and more businesses have millions of customers and billions of their transactions. They need to understand risks (Is this transaction fraudulent? Will this customer pay their bills?) and opportunities (What is the expected profit of this customer? What product is this customer most likely to buy next?). Electronic Commerce. Not only does electronic commerce produce large data sets in which the analysis of marketing patterns and risk patterns is critical, but unlike some of the applications above, it is also important to do this in real or near-real time, in order to meet the demands of on-line transactions. Genomic Data. Genomic sequencing and mapping efforts have produced a number of databases which are accessible over the web. In addition, there are also a wide variety of other on-line databases, including those containing information about diseases, cellular function, and drugs. Finding relationships between these data sources, which are largely unexplored, is another fundamental data mining challenge. Recently, scalable techniques have been developed for comparing whole genomes. Sensor Data. Satellites, buoys, balloons, and a variety of other sensors produce voluminous amounts of data about the earth's atmosphere, oceans, and lands. A fundamental challenge is to understand the relationships, including causal relationships amongst this data. For example, do industrial pollutants affect global warming? There are also large terabyte to petabyte data sets being produced by sensors and instruments in other disciplines, such as astronomy, high energy physics, and nuclear physics. Simulation Data. Simulation is now accepted as a third mode of science, supplementing theory and experiment. Today, not only do experiments produce huge data sets, but so do simulations. Data mining, and more generally data intensive computing, is proving to be a critical link between theory, simulation, and experiment. Health Care Data. Health care has been the most rapidly growing segment of the nation's GDP for some time. Hospitals, health care organizations, insurance companies, and the federal government have large collections of data about patients, their health care problems, the clinical procedures used, their costs, and the outcomes. Understanding relationships in this data is critical for a wide variety of problems, ranging from determining what procedures and clinical protocols are most effective to how best to deliver health care to the most people in an era of diminishing resources. Multi-media Documents. Few people are satisfied with today's technology for retrieving documents on the web, yet the number of documents and the number of people accessing these documents is growing explosively. In addition, it is becoming easier and and easier to archive multi-media data, including audio, images, and video data, but harder and harder to extract meaningful information from the archives as the volume grows. The Data Web. Today the web is primarily oriented toward documents and their multi-media extensions. HTML has proved itself to be a simple, yet powerful language for supporting this. Tomorrow the potential exists for the web to prove equally important for working with data. The Extensible Markup Language (XML) is an emerging language for working with data in networked environments. As this infrastructure grows, data mining is expected to be a critical enabling technology for the emerging data web.

3. Challenges

Data Mining in a Network Setting

1. *Community and social networks*

Today's world is interconnected through many types of links. These links include Web pages, blogs, and emails. Many respondents consider community mining and the mining of social networks as important topics. Community structures are important properties of social networks. The identification problem in itself is a challenging one. First, it's critical to have the right characterization of the notion of "community" that is to be detected. Second, the entities/nodes involved are distributed in real-life applications, and hence distributed means of identification will be desired. Third, a snapshot-based dataset may not be able to capture the real picture; what is most important lies in the local relationships (e.g. the nature and frequency of local interactions) between the entities/nodes. Under these circumstances, our challenge is to understand (1) the network's static structures (e.g. topologies and clusters) and (2) dynamic behavior (such as growth factors, robustness, and functional efficiency). A similar challenge exists in bio-informatics, as we are currently moving our attention to the dynamic studies of regulatory networks. A questions related to this issue is what local algorithms/protocols are necessary in order to detect (or form) communities in a bottom-up fashion (as in the real world). A concrete question is as follows. Email exchanges within an organization

or in one's own mailbox over a long period of time can be mined to show how various networks of common practice or friendship start to emerge. How can we obtain and mine useful knowledge from them?

2. Mining in and for computer networks — high-speed mining of high-speed streams

Network mining problems pose a key challenge. Network links are increasing in speed, and service providers are now deploying 1 Gig Ethernet and 10 Gig Ethernet link speeds. To be able to detect anomalies (e.g. sudden traffic spikes due to a DoS (Denial of Service) attack or catastrophic event), service providers will need to be (several hundred GB) of data each day. One will need highly scalable solutions here. Good algorithms are, therefore, needed to detect whether DoS attacks do not exist. Also, once an attack has been detected, how does one discriminate between legitimate traffic and attack traffic so that it is possible to drop attack packets? We need techniques to
(1) detect DoS attacks,
(2) trace back to find out who the attackers are, and
(3) drop those packets that belong to attack traffic.

3. Distributed Data Mining and Mining Multi-Agent Data

The problem of distributed data mining is very important in network problems. In a distributed environment (such as a sensor or IP network), one has distributed probes placed at strategic locations within the network. The problem here is to be able to correlate the data seen at the various probes, and discover patterns in the global data seen at all the different probes. There could be different models of distributed data mining here, but one could involve a NOC that collects data from the distributed sites, and another in which all sites are treated equally. The goal here obviously would be to minimize the amount of data shipped between the various sites — essentially, to reduce the communication overhead. In distributed mining, one problem is how to mine across multiple heterogeneous data sources: multi-database and multi-relational mining. Another important new area is *adversary data mining*. In a growing number of domains — email spam, counter-terrorism, intrusion detection/computer security, click spam, search engine spam, surveillance, fraud detection, shopbots, file sharing, etc. — data mining systems face adversaries that deliberately manipulate the data to sabotage them (e.g. make them produce false negatives). We need to develop systems that explicitly take this into account, by combining data mining with game theory.

4. Data Mining for Biological and Environmental Problems

Many researchers that we surveyed believe that mining biological data continues to be an extremely important problem, both for data mining research and for biomedical sciences. An example of a research issue is how to apply data mining to HIV vaccine design. In molecular biology, many complex data mining tasks exist, which cannot be handled by standard data mining algorithms. These problems involve many different aspects, such as DNA, chemical properties, 3D structures, and functional properties. There is also a need to go beyond bio-data mining. Data mining researchers should consider ecological and environmental informatics. One of the biggest concerns today, which is going to require significant data mining efforts, is the question of how we can best understand and hence utilize our natural environment and resources — since the world today is highly "resource-driven"! Data mining will be able to make a high impact in the area of integrated data fusion and mining in ecological /environmental applications, especially when involving distributed/decentralized data sources, e.g. autonomous mobile sensor networks for monitoring climate and/or vegetation changes. For example, how can data mining technologies be used to study and find out contributing factors in the observed doubling of the number of hurricane occurrences over the past decades, as recently reported in *Science* magazine? Most of the data sources that we are dealing with today are fast evolving, e.g. those from stock markets or city traffic. There is much interesting knowledge yet to be discovered, as far as the dynamic change regularities and/or their cross-interactions are concerned. In this regard, one of the challenges today is how to deal with the problem of dynamic temporal behavioral pattern identification and prediction in: (1) very largescale systems (e.g. global climate changes and potential "bird flu" epidemics) and (2) human-centered systems (e.g. user-adapted human-computer interaction or P2P transactions). Related to these questions about important applications, there is a need to focus on "killer applications" of data mining. So far three important and challenging applications for data mining have emerged: bioinformatics, CRM/personalization and security applications. However, more explorations are needed to expand these applications and extend the list of applications.

5. Data Mining Process-Related Problems

Important topics exist in improving data-mining tools and processes through automation, as suggested by several researchers. Specific issues include how to automate the composition of data mining operations and building a methodology into data mining systems to help users avoid many data mining mistakes. If we automate the different data mining process operations, it would be possible to reduce human labor as much as possible. One important issue is how to automate data cleaning. We can build models and find patterns very fast today, but 90 percent of the cost is in pre-processing (data integration, data cleaning, etc.) Reducing this cost will have a much greater payoff

than further reducing the cost of model-building and pattern-finding. Another issue is how to perform systematic documentation of data cleaning. Another issue is how to combine visual interactive and automatic data mining techniques together. He observes that in many applications, data mining goals and tasks cannot be fully specified, especially in exploratory data analysis. Visualization helps to learn more about the data and define/refine the data mining tasks. There is also a need for the development of a theory behind interactive exploration of large/complex datasets. An important question to ask is: what are the compositional approaches for multi-step mining "queries"? What is the canonical set of data mining operators for the interactive exploration approach? For example, the data mining system *Clementine* has a nice user interface, but what is the theory behind its operations?

6. Security, Privacy, and Data Integrity

Several researchers considered privacy protection in data mining as an important topic. That is, how to ensure the users' privacy while their data are being mined. Related to this topic is data mining for protection of security and privacy. One respondent states that if we do not solve the privacy issue, data mining will become a derogatory term to the general public. Some respondents consider the problem of knowledge integrity assessment to be important. We quote their observations: "Data mining algorithms are frequently applied to data that have been intentionally modified from their original version, in order to misinform the recipients of the data or to counter privacy and security threats. Such modifications can distort, to an unknown extent, the knowledge contained in the original data. As a result, one of the challenges facing researchers is the development of measures not only to evaluate the knowledge integrity of a collection of data, but also of measures to evaluate the knowledge integrity of individual patterns. Additionally, the problem of knowledge integrity assessment presents several challenges." Related to the knowledge integrity assessment issue, the two most significant challenges are: (1) develop efficient algorithms for comparing the knowledge contents of the two (before and after) versions of the data, and (2) develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtainable by broad classes of data mining algorithms. The first challenge requires the development of efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data. The second challenge is to develop algorithms to measure the impact that the modification of data values has on a discovered pattern's statistical significance, although it might be infeasible to develop a global measure for all data mining algorithms.

7. Dealing with Non-Static, Unbalanced and Cost-Sensitive Data

An important issue is that the learned models should incorporate time because data is not static and is constantly changing in many domains. Historical actions in sampling and model building are not optimal, but they are not chosen randomly either. This gives the following challenging phenomenon for the data collection process. Suppose that we use the data collected in 2000 to learn a model. We then apply this model to select inside the 2001 population. Subsequently, we use the data about the individuals selected in 2001 to learn a new model, and then apply this model in 2002. If this process continues, then each time a new model is learned, its training set has been created using a different selection bias. Thus, a challenging problem is how to correct the bias as much as possible. Another related issue is how to deal with unbalanced and cost-sensitive data, a major challenge in research. Charles Elkan made the observation in an invited talk at *ICML 2003 Workshop on Learning from Imbalanced Data Sets*. First, in previous studies, it has been observed that UCI datasets are small and not highly unbalanced. In a typical real-world dataset, there are at least 10^5 examples and $10^{2.5}$ features, without single well-defined target class. Interesting cases have a frequency of less than 0.01. There is much information on costs and benefits, but no overall model of profit and loss. There are different cost matrices for different examples. However, most cost matrix entries are unknown. An example of this dataset is the direct marketing DMEF data library. Furthermore, the costs of different outcomes are dependent on the examples; for example, the false negative cost of direct marketing is directly proportional to the amount of a potential donation. Traditional methods for obtaining these costs relied on sampling methods. However, sampling methods can easily give biased results.

4. CONCLUSIONS

In this paper we briefly reviewed the various data mining applications. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining. Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns

and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory.

Since its conception in the late 1980s, data mining has achieved tremendous success. Many new problems have emerged and have been solved by data mining researchers. However, there is still a lack of timely exchange of important topics in the community as a whole. This article summarizes a survey that we have conducted to rank 10 most important problems in data mining research. These problems are sampled from a small, albeit important, segment of the community. The list should obviously be a function of time for this dynamic field. Finally, we summarize the 10 problems below:

- Developing a unifying theory of data mining
- Scaling up for high dimensional data and high speed data streams
- Mining sequence data and time series data
- Mining complex knowledge from complex data
- Data mining in a network setting
- Distributed data mining and mining multi-agent data
- Data mining for biological and environmental problems
- Data Mining process-related problems
- Security, privacy and data integrity
- Dealing with non-static, unbalanced and cost-sensitive data

5. REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & 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] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[6] Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", *Pearson Education, New* Agrawal, R., Imielinki, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM-SIGMOD International Conference* on *Management of Data*, 207-216, Washington, DC.

[7] Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 1994 International Conference on Very Large Databases,* 487-499, Santiago, Chile.