Data Mining: Challenges for Future and Various Issues

Prof. Narendra J. Padole **Prof. Sameer Y. Thakur**

Abstract **—Generally mining of data is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data, however, it can also disclosure sensitive information about individuals compromising the individual's right to privacy.It is a process to extract the implicit information; knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. Therefore, privacy preserving data mining has becoming an increasingly important field of research. Nowadays, Data mining is emerging area to extract implicit and useful knowledge and also recognized as an important technology for businesses internationally and locally. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization". Undoubtedly, research in data mining will continue and even increase over coming decades. Hence this paper sketches vision of the future work to done in area of data mining. This paper elaborate various topics (starting from the classic definition of "data mining" and its basic terms)included various future challenges and issues in data mining which is important to do further more research in this emerging field.**

Keywords— **Data mining; Association rules; Clustering; Decision tree; Challenges; Limitations.**

I. INTRODUCTION

Data mining is discovering the methods and patterns in large databases to guide decisions about future activities [3]. It is expected that data mining tools to get the model with minimal input from the user to recognize. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. It often involves the analysis of data stored in a data warehouse.

In recent years, with the explosive development in Internet, data storage and data processing technologies; privacy preservation lias had been one of the greater concerns in data mining [14]. Generally three of the major data mining techniques are regression, classification and clustering. Data Mining also popularly known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Data mining and knowledge discovery in databases are frequently treated words used as synonyms in this paper. Knowledge discovery (a step of data mining process) in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. The following figure 1 shows data mining as a step in an iterative knowledge discovery process.

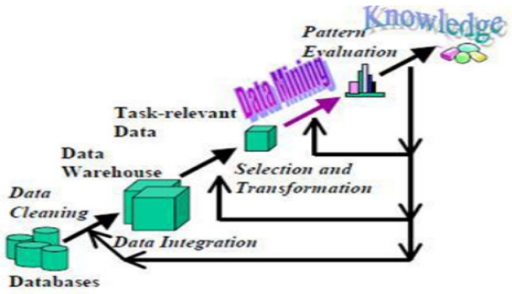

Fig.l: Data mining is the core of Knowledge Discovery Process

The KDD (Knowledge Discovery in Databases) process comprises of a few steps leading from raw data collections to some form of new knowledge.

a) Data cleaning : It also known as data cleansing, it is a phase in which irrelevant data and noise data are removed from the raw collection of data.

b) Data integration: in this multiple data sources, often heterogeneous may be combined in a common source.

c) Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

d) Data transformation: lt also known as data consolidation, in this process selected data is transformed into forms appropriate for the mining procedure.

e) Pattern evaluation: at this level, strictly interesting patterns representing knowledge are identified based on given measures.

f) Data Mining Processes: Data mining process consists of an iterative sequence of several steps/process:data preprocessing; data management; data mining tasks and algorithms, and post processing (Li, Li, Zhu, & Ogihara, 2002) [13].

g) Knowledge representation: It is final phase of KDD process in which discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

h) Now for a major problem in data mining, today's competition is one of the most important challenges facing by all organizations and industries in data mining issues.

i) So keeping above points in mind, this paper organized as follows: Section 2 describe related work. Section 3 defines about data mining goals. Challenges of data mining are explained in section 4. Limitations of data mining are defined in section 5. This paper discusses several future issues to concentrate on future problems in data mining in section 6, and finally conclude this paper in section

II. RELATED WORK

Data mining is the process of extracting and valuable interesting patterns from raw collection. However it can also disclosure sensitive information about individuals compromising the individual's right to privacy [1].

Generally it is a multidisciplinary field, drawing work from areas including database technology, pattern recognition, information retrieval, machine learning, statistics, neural networks, knowledge-based systems. High- performance computing, artificial intelligence, and data visualization [5].It involves the use of sophisticated data analysis tools to discover previously unknown valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods [3].

As explained today's privacy preserving in data mining has become an increasingly important field of research. Data mining can be performed on collected data represented in quantitative, textual, or multimedia forms etc. Now this section describes about various important issues and main contribution of this paper. This paper contribution can be summarized as follows:

a) Make an effort to mark process steps of knowledge discovery; goals of data mining and organize the current knowledge in future tasks with mentions important future issue.

b) Present the limitations and future work of data mining for future research directions.

c) Present the challenges of data mining to provide future research directions.

III. DATA MINING GOALS

In general, Data mining is used for a variety of purposes in both the private and public sectors. Industries such as insurance, banking, medicine, and retailing commonly use data mining to increase sales, enhance research, and reduce costs. Hence the goals of data mining instead applications also as:

a) **Data Processing:** Depending on the goals and requirements of the KDD process, analysts may select, filter, aggregate, sample, clean and/or transform data [10], Automating some of the most typical data processing tasks and integrating them seamlessly into the overall process may eliminate or at least greatly reduce the need for programming specialized routines and for data export/import, thus improving the analyst's productivity.

b) **Association rule learning:** It searches for relationships between variables [14] for e.g. a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

c) **Prediction:** Given a data item and a predictive model, predict the value for a specific attribute of the data item. For example, given a predictive model of credit card transactions, predict the likelihood that a specific transaction is fraudulent. Prediction may also be used to validate a discovered hypothesis.

d) **Regression:** For a given set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item and the automatic production of a model that can predict these attribute values for new records for e.g. given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.

e) **Classification:** Given a set of predefined categorical classes, determine "which of these classes a specific data item belongs" for e.g. given classes of patients that corresponds to medical treatment responses; identify the form of treatment to which a new patient is most likely to respond.

f) **Link Analysis /Associations:** in this, items; identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. These relations may be associations between attributes within the same data item (out of the shoppers who bought milk, 64% also purchased bread) or associations between different data items (Every time a certain stock drops 5%, a certain other stock raises 13% between 2 and 6 weeks later). The investigation of relationships between items over a period of time is also often referred to as 'sequential pattern analysis.

g) **Clustering:** in this for a given a set of data items, partition this data items into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar for e.g. given a data set of customers; identify subgroups of customers that have a similar buying behavior.

h) **Exploratory Data Analysis (EDA):** It is the interactive exploration of a data set without heavy dependence on preconceived assumptions and models, thus attempting to identify interesting patterns. Graphic representations of the data are used very often to exploit the power of the eye and human intuition. While there are dozens of software packets available that were developed exclusively to support data exploration, it might also be desirable to

integrate these approaches into an overall KDD environment [1, 10].

i) **Model Visualization:** Visualization plays an important role in making the discovered knowledge understandable and interpretable by humans. Besides, the human eye-brain system itself still remains the best pattern-recognition device known. Visualization techniques may range from simple scatter plots and histogram plots over parallel coordinates to 3D movies.

j) **Summarization:** in short, it provides a compact description for a subset of data.

k) **Dependency modeling:** It describes significant dependencies among variables.

IV. CHALLENGES IN DATA MINING

This section discusses various challenges in area of data mining to processing and extracting valuable data from collected data in real world that are discussed by Hristidis et al.(2010), as preprocessing, post processing; data mining tasks and algorithms.. I nil also According to article in the June issue of Data Analytics magazine [3], there are three most one challenge: Dirty Data ;Explaining Data Mining to Others; Difficulty Accessing Data. Hence each one challenges can be discussed now as:

As advances in storage media and network architectures have made it possible and affordable to collect and store huge amounts of data in various media types (text, image, video, graphics etc.) formats, and structures from multiple information sources. These data typically include a significant amount of missing values and noises and may have multi-level confidences, multi-level completeness and may be inconsistent. In order to improve the efficiency and accuracy of knowledge discovery and data mining process, ensuring data quality is a big challenge.

a) **Data Cleaning and preprocessing:** It is an important step to ensure the data quality and to improve the efficiency and ease of the mining process. Because Real-world data tend to be incomplete, inconsistent, noisy, high dimensional and multi-sensory etc means which are not suitable for mining directly. Actually Data preprocessing includes data cleaning to remove noisy data and outliers, data reduction to reduce the dimensionality and complexity of the data, data integration to integrate data from multiple information sources, and data transformation to convert the data into suitable forms for mining etc.

b) **Post processing:** It is the process to refine and evaluate the knowledge derived from mining procedure (Brulia and Famili, 2000). As example Post processing include the simplification of the extracted knowledge, evaluating the extracted knowledge and visualizing it, or documenting it. Generally Challenges in post-processing such as(a) how to evaluate the discovered patterns (b) how to present the mining results to the domain experts in a way that is easy to understand and interpret (c) how to convert the discovered patterns into knowledge etc.

c) **Tasks and Algorithms:** Data mining tasks and algorithms are essential steps of knowledge discovery. Data mining typically involve a wide range of tasks and algorithms such as pattern mining for discovering interesting associations and correlations; clustering and trend analysis to understand the nontrivial changes and trends and classification to prevent future reoccurrences of undesirable phenomena. These different data mining tasks may use the same database in different ways.

Generally Challenges in tasks and algorithms (a) how to achieve the efficient data mining of different kinds of knowledge using different data mining algorithms(b) many datasets needed for emergency evacuation planning are of geospatial type. It is just not a complicated task to define a local neighborhood for mining geospatial patterns that includes space, time and semantic information, but also challenging to incorporate domain specific information (e.g., semantic ontology) into the mining process without compromising the underlying performance[15,16], Hence Data mining across multiple information sources is a critical and challenging task because of the vast difference in data type, dimension and quality.

Also, the data may contain inherent uncertainty and impreciseness due to the random nature of data generation and collection process. Although there lias been much research work in data uncertainty management and in querying data with uncertainty, there is only limited research work in mining uncertain data. Unpredictable events often indicate suspicious situations. However, these events are extremely difficult to detect because they don't occur often or they occur at a time/location where they are not expected. For domain specific applications, utilizing the domain knowledge to guide data mining process or improve data mining performance is a challenging issue.

d) **Dirty Data:** Here no surprise that dirty data tops the list, because it lias been at the top of the list for the past several years [2] in area of data mining as a challenging issue. Many data miners provided input as "How they have tried to overcome the problem and how to provide a clear theme emerges those involving business users". Data miners use descriptive statistics and visualization to assist business users in understanding their data and identifying problem areas etc. Helping users understand their data "hands on" and helps everyone to gain a shared understanding about the quality of the data. This can help manage expectations about providing potential results of a data modeling exercise and also create action plans to improve quality of data.

e) **Explaining Data Mining to Others:** Some data miners expressed frustration with executives who don't support solutions because they do not have the knowledge about background to understand about data mining, but at the same time refuse to sit through more than a brief presentation on the topic. Data miners recommended finding support one level down from the executive i.e. identifying someone who is willing to invest time to understand the solutions and willing to champion solutions with the senior executive. Other data miners went even lower in the organization, and convinced key users to identify a problem and work interactively with the data miner on the solution. This allows the business users to see the power and capability of data analytics first-hand and to be able to get answers to questions "on the fly."

f) **Difficulty Accessing Data:** To accessing data that is so typical and have a major challenge for data mining also, for e.g it is scattered throughout an organization, more commonly accessing data because it does not exist. Data miners generally agreed that difficulty accessing data is due to the lack of a plan or strategy for data i.e. "how it can be obtained", "what data is needed", "how quality can be assured or improved"

and "how it can be maintained" etc. Again data miners suggest working directly with business users to match business problems with data requirements, and to use this as way to begin developing a broader plan for data collection and data accessibility.

g) **Other Challenges in Data Mining**: Despite above challenges, to do research in field of data mining, there are some other challenges also discuss as:

1) The patterns described by the data mining algorithms are still too abstract for being understood. However, a pattern that is misinterpreted is of great danger for e.g. many data mining algorithms do not distinguish between co-occurrence and causality [9, 16], Consider an application that aims at finding the reason for a certain type of disease. There is a great difference between finding the origin of the disease and finding just an additional symptom. Therefore, a very old challenge will remain very important for the data mining community: developing systems which derive understandable patterns and making already derived patterns understandable [9],

2) When working with future patterns, it increased the number of valid patterns, and finding a large data set of complex objects. Therefore, the number of potentially valid patterns will be too large to be handled by a human user, without a system organizing the results. Thus, future systems must provide a platform for pattern exploration where users can browse for knowledge they might consider as interesting.

h) Generally existed current algorithms mostly focus on a limited set of standard patterns. However, deriving these patterns often does not yield a direct and complete solution to many problems where data mining could be very useful. Furthermore, with an increasing complexity of the analyzed data, it is likely that the derived patterns will increase in complexity as well.

i) Thus, a future trend in data mining will be to find richer patterns.

j) Scalability

k) Data Quality

l) Data Ownership and Distribution

m) Dimensionality

n) Privacy Preservation

o) Streaming Data

p) Complex and Heterogeneous Data

Hence this section describes existed various challenges in area of data mining to do concentrate more research in as future work. Now next section discusses various limitations issues in data mining.

V. LIMITATIONS IN DATA MINING

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. Data mining products can be very powerful tools; they are not self sufficient applications. To be successful, data mining requires analytical specialists and skilled technical that can structure the analysis and interpret the output that is created.

Consequently, the limitations of data mining are:

a) Primarily data or personnel related, rather than technology-related. Although data mining can help reveal patterns and relationships, it does not tell the

user the value or significance of these patterns. These types of determinations must be made by the user [15, 16], Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances for e.g. to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

b) While it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship for e.g. an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use.

However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variables such as occupation (the need to make trips on hobby (taking short However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variables such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a advantage of last minute discounts to visit new destinations).

Hence this section describes several limitations in data mining. Now next section describes several future issues to do further research in this area.

VI. FUTURE WORK

Today's competition is one of the most important challenges facing by all organizations and industries in data mining issues. That is hard to find in a particular organization or industry which has no rival to him. As explained Data mining is the extraction of interesting patterns or extracts new knowledge from existed knowledge data set for further research. So the following points will work for futureworks are as:

To address these issues, following problem should be widely studied:

Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.

In distributed privacy preserving data mining areas, efficiency is an essential issue. We should try to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost [14],

Side-effects are unavoidable in data sanitization process. How to reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring the side-effects resulted from data processing.

The data to be studied is usually drawn from several sources. For this reason another important trend in data mining will be the growing importance of data preprocessing and integration, ensuring that the "patterns in data" found are "valid" on the complete set of data objects and not just on a particular subset.

How to deploy privacy-preserving techniques into practical applications is also required to be further studied.

a) Developing a unifying theory of data mining
b) Data mining for biological and enviromenental problems

c) Data Mining process-related problems
d) Mining complex knowledge from complex data
e) Data mining in a network setting
f) Security, privacy and data integrity
g) Dealing with non-static, unbalanced and cost-sensitive data
h) Scaling up for high dimensional data and high speed data streams
i) Mining sequence data and time series data
j) Distributed data mining and mining multi-agent data

An ultimate trend that data mining faces is increased usability to detect "understandable patterns", and to make data mining methods more user-friendly. If future data mining methods have to handle all this complex input and intelligent preprocessing, it is very likely that the user will have to adjust more and more switches and knobs before getting any result. Hence, achieving user-friendliness with transparent or even reduced parameterization is a major goal. Usability is also enhanced by finding new types of patterns that are easy to interpret, even if the input data is very complex.

a) We started with the type of "patterns in data" which knowledge discovery is examining. While original data mining concentrated on pectoral data, future data will predominantly be stored in much more complex data types and data mining will have to cope with this increasing volume of structured data. Another aspect of "patterns in data" in the future is the increasing importance of studying their evolution over time. Considering time, allows observing the dynamics of patterns as well as the behavior and the interactions of data objects.

Although no human being can foretell the future, so there are plenty of interesting new challenges ahead of present research, and quite a few of them cannot be foreseen at the current point of time.

b) To discover any hidden patterns or valuable information from vast data, the data mining methods and algorithms have to be efficient and scalable. We also need to consider localized facts applying data mining in emergency evacuation planning as often some situations are very unique to a specific region

c) Real time information (such as hourly weather forecast and real time traffic data) also needs to be considered, i) Data needed to build a proper and effective emergency evacuation plan come from multiple sources including census data, disaster management data, traffic and transportation data, map data, and so on[13]. Also in recent years the volume of all data has grown tremendously both in size (i.e., number of instances) and dimensionality (i.e., number of items). It is a critical and challenging task to data mining across multiple data sources because of the huge difference in data type, dimensionality and quality (Grossman &Mazzucco, 2002; Wu, Oviatt, & Cohen, 1999). Data come from multiple sources are also often very inconsistent. So for proper data mining in emergency evacuation planning, there must be an effective data integration technique in place. Data cleaning and preprocessing is also important, discussed in Section 4 to make the data consistent.

d) As defined. Data mining is a very powerful tool and very limited research lias been done on using advanced and effective data mining techniques specifically in emergency evacuation planning [16], There is high potentiality conducting further research using advanced data miming in emergency evacuation planning. For e.g. from past research, GIS

seemed to be an effective tool for emergency evacuation planning. So applying effective data mining methods and algorithms using various other tools such as GIS, ArcMap, etc., and data from multiple heterogeneous data sources (like census, map, disaster, traffic, etc.), collaborating among emergency management professionals and data mining specialists, can lead to develop a proper and effective emergency evacuation plan to avoid casualties and financial crisis during any natural disasters or terrorist attacks [13].

Hence to conclude, future data mining should generate a large variety of well understandable patterns. Due to variations in the parameterizations, the number of possibly meaningful and useful patterns will dramatically increase and thus, an important aspect is managing and visualizing these patterns. Hence this section discusses several future issues on data mining to do research. Now finally, next section concludes this paper in short.

VII. CONCLUSION

Data mining seeks to extract hidden knowledge from large amount of data. Data mining is the process of extracting and valuable interesting patterns from raw collection of data. Data mining can be used to uncover patterns in the data but it is often carried out only on the samples of data. This mining process will be ineffective if the samples are not a good representation of the larger body of the data. And beside this, today's competition is one of the most important challenges facing by all organizations and industries in data mining issues. That is hard to find in a particular organization or industry which lias no rival to liimTliis paper describes various tasks; goals and limitations of data mining. Additionally this paper also discussed about the various valuable problems; future challenges and issues in field of data mining which is important to do further more effective research in this emerging field.

REFERENCES

[1] Clifton. Christopher "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[2] Ian H. Witten; Eibe Frank; Mark A. Hall. "Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.)". Elsevier. 30 January 2011.

[3] Kantardzic, Mehmed "Data Mining: Concepts. Models. Methods, and Algorithms". John Wiley & Sons. 2003.

Gimnemann, S.; Kremer. H.; Seidl. T. "An extension of the PMML standard to subspace clustering models". Proceedings of the 2011 workshop on Predictive markup language modeling - PMML '11.

[5]Ralf Mikut; Markus Reischl. "Data Mining Tools". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Retrieved October 21. 2011.

[6] Domenico Talia: Paolo Tranfio"How distributed data mining tasks can thrive as knowledge services". Communications of the ACM . 2010

[7] Karl Rexer. Heather Allen. & Paul Gearan "Data Miner Survey - Overcoming Data Mining's Top Challenges". 2011.

[8] Hans-Peter Kriegel . Karsten M. Borgwardt et.al "Future trends in data mining", sprinzer. 2007.

[9] QIANG YANGXINDONG WU. "10 Challenging Problems in Data Mining Research".

[10] Michael Goebel.Le Gruenwald "A Survey of Data Mining and Knowledge Discovery Software Tools". ACM. 1999.

[11] DianshengGuo. Jeremy Mennis. "Spatial data mining and geographic knowledge discovery—An introduction". 2009 Elsevier

[12] Pavel Berkhin Accrue Software. Inc."Survey of Clustering Data Mining Techniques".

[13] MuhammedMiah, "Survey of Data Mining Methods in Emergency Evacuation Planning", Conference for Information Systems Applied Research 2011. CONISAR Proceedings Wilmington North Carolina. USA.

[14] Pingshui WANG, "Survey on Privacy Preserving Data Mining", International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010.

[15] Jeffrey W. Seifert, "Data Mining: An Overview" CRS Report for Congress, 2004.

[16] Venkatadari M., Dr.Lokanataha C. Reddy, "A Review on Data Mining from Past to Future", International Journal of Computer Applications, pp.19-22. vol. 15. No. 7. Feb 2011.