ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

A Review on Data Mining Techniques and its Recent Development

R.Subhulakshmi[1], P.Sangeetha[2], V.Rohini[3], P.Manjula[4]

#V.Sathyabama MSc., M.Phil.,
Assiatant Professor
Dept of Information Technology
manonmaniam sundaranar university
G.Venkataswamy Naidu College (SFC),Kovilpatti,Tamilnadu
#P.Manjula
M.Sc, Information Technology, Scholar
manonmaniam sundaranar university
G.Venkataswamy Naidu College (SFC),Kovilpatti,Tamilnadu.
[1]sathyamit16@gmail.com
[4] pmanjula1995@gmail.com

**P.Sangeetha, V.Rohini*
M.Sc, Information Technology, Scholars
manonmaniam sundaranar university
G.Venkataswamy Naidu College (SFC),Kovilpatti,Tamilnadu
[2]sangithasara@gmail.com
[3] rohiniraman3095@gmail.com

Abstract- **Data and Information/knowledge has an important role in every aspect of human activities. In this paper discussed the data mining techniques, tools and advantages.**

Keywords- Data mining, Techniques.

I.INTRODUCTION

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 techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time.

Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: [1]

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data layer: as mentioned above, data layer can be a database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in the form of reports or another kind of visualization.

Data mining application layer is used to retrieve data from the database. Some transformation routine can be performed here to transform data into the desired format. Then data is processed using various data mining algorithms.

Front-end layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer. [2]

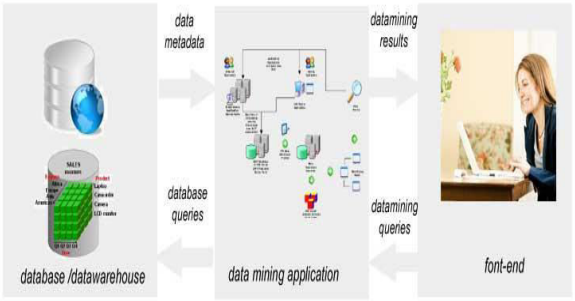

Fig1.1Architecture of Data Mining

II. KNOWLEDGE DISCOVERY IN DATABASES PROCESS

KDD processes are depicted in the following diagram. It is important to note that KDD is not accomplished without human interaction. The selection of a data set and subset requires an understanding of the domain from which the data is to be extracted. For example, a database may contain customer address that would not be pertinent to discovering patterns in the selection of food items at a grocery store. Deleting non-related data elements from the dataset reduces the search space during the data mining phase of KDD. If the dataset can be analyzed using a sampling of the data, the sample size and composition are determined during this stage.

Databases are notoriously "noisy" or contain inaccurate or missing data. During the reprocessing stage the data is cleaned. This involves the removal of "outliers" if appropriate; deciding strategies for handling missing data fields; accounting for time sequence information, and applicable normalization of data.

In the transformation phase attempts to limit or reduce the number of data elements that are evaluated while maintaining the validity of the data. During this stage data is organized, converted from one type to another (i.e. changing nominal to numeric) and new or "derived" attributes are defined.

At this point the data is subjected to one or several data mining methods such as classification, regression, or clustering. The data mining component of KDD often involves repeated iterative application of particular data mining methods. "For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulation manager might need to first use clustering to segment the subscriber database, then apply rule induction to automatically create a classification for each desired cluster.Various data mining methods will be discussed in more detail in following sections.

Interpretation and documentation of the results from the previous steps. Actions at this stage could consist of returning to a previous step in the KDD process to further refine the acquired knowledge, or translating the knowledge into a form understandable to the user. A commonly used interpretive technique is visualization of the extracted patterns. The results should be critically reviewed and conflicts with previously believed or extracted knowledge resolved.

Understanding and committing to all phases of the data mining process is crucial to its success. [3]

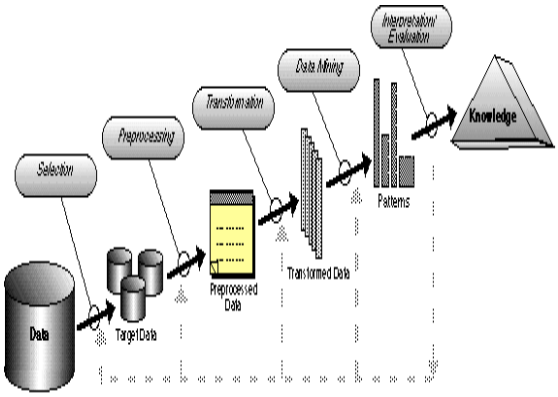

Fig 1.2 Steps for KDD Process

III . ADVANTAGES OF DATA MINING

Data mining is an important part of knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Data mining is applied effectively not only in the business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc. Data mining has a lot of advantages when using in a specific industry. Besides those advantages, data mining also has its own disadvantages e.g., privacy, security and misuse of information. We will examine those advantages and disadvantages of data mining in different industries in a greater detail.

- *Marketing / Retail*

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through

market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

- *Finance / Banking*

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

- *Manufacturing*

Data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of the golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

- *Governments*

Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

IV. Disadvantages of Data Mining

- *Privacy Issues*

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

- *Security Issues*

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

- *Misuse /inaccurate information*

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

In addition, data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence. [5]

V. Challenges in Data Mining

- 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 Costsensitive Data.[4]

VI. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

- *Classification*

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data

classification process involves learning and classification.

In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

✓ *Types of classification models:*

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

➢ *Clustering*

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

✓ *Types of clustering methods*

- Partitioning Methods
- Hierarchical Agglomerative(divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

➢ *Predication*

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

✓ *Types of regression methods*

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

➢ *Association rule*

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

✓ *Types of association rule*

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

➢ *Neural networks*

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real

world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. [6]

✓ *Types of neural networks*

- Back Propagation

VI.REAL TIME EXAMPLES

➢ *Education*

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

➢ *Research Analysis*

History shows that we have witnessed revolutionary changes in research. Data mining is

helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

➢ *Customer Segmentation*

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

➢ *Criminal Investigation*

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

➢ *Bio Informatics*

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction

➢ *Corporate Surveillance*

Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

➢ Market Basket Analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

➢ *Future Healthcare*

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse. [7]

VII.DATA MINING TOOLS

- *weka*

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modeling, which currently is not included.

- *KNIME*

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis.

Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

- *NLTK*

NLTK provides a pool of language processing tools including data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks. All you need to do is install NLTK, pull a package for your favorite task and you are ready to go. Because it's written in Python, you can build applications on top if it, customizing it for small tasks. [8]

VIII.CONCLUSION

In this paper overview of Data mining is discussed. The analysis of the various Data Mining techniques and application. Finally, this review will motivate a new direction for future research.

IX. REFERENCE

[1] http://www.thearling.com/text/dmwhite/dmwhite.htm

[2] http://www.zentut.com/data-mining/data-mining-architecture/.

[3] https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm

[4] http://www.ijera.com/papers/Vol4_issue5/Version%203/G045033841.pdf

[5] http://www.zentut.com/data mining/advantages-and-disadvantages-of-data-mining

[6] http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf

[7] http://bigdata-madesimple.com/14-useful-applications-of-data-mining/

[8] http://www.dereak.com/newsletter/2nd%dataminingtool.htm