International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Data Mining Techniques and its Applications : An Approach to Discover Knowledge in Data

Paramjeet Kaur
Assistant Professor of Computer Science Guru Nanak College Ferozepur

ABSTRACT

Data Mining (sometimes called data discovery) is the process which finds extraction of useful information and pattern from huge data, analyzing data from different perspective and summarizing it into useful information that can be used to increase revenue, cut costs, or both. Data mining is the process of extracting the useful data, patterns and trends from a large amount of data by using techniques like clustering, classification, association and regression. There are a wide variety of applications in real life. This paper discuss few of data mining techniques and applications involving some of the organizations which have adapted data mining technology to improve their business and found best results.

Keywords: Data Mining Techniques, Data Mining Applications.

I. INTRODUCTION

The development of Information Technology has generated large amount of data base in various areas. The research in database and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining".

Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining.

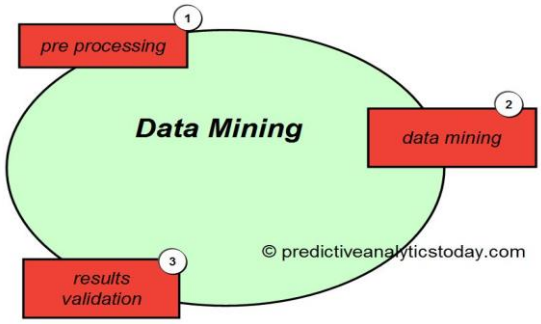

Figure 1.(a) Data Mining

What kind of data can be mined?

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different type of data. Data mining is being put into use and studied for databases, including relational database, object oriented database, data warehouses, transactional databases, unstructured and semi- structured repository such as World Wide Web, advanced database and textual databases, and even flat files.

DATA MINING PROCESS

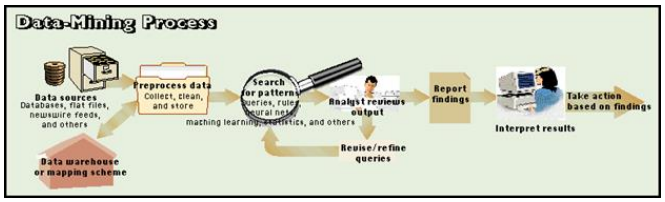

Figure 2.(b) Data Mining Process

Data mining is used to extract implicit and previously unknown information from data. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. So, many people use the term "knowledge discovery device" or KDD for data mining. Knowledge extraction or discovery is done in seven sequential steps used in data mining:

1) Data cleaning: we remove noise data and irrelevant data from collected raw data.

2) Data integration: At this step, we combine multiple data sources into single data store called target data.

3) Data Selection: Here, data relevant to analysis task are retrieved from data base as pre-processed data.

4) Data transformation: Here, data is consolidating into standard formats appropriate for mining by summarizing and aggregated operations.

5) Data Mining: At this step, various smart techniques and tools are applied in order to extract data pattern or rules.

6) Pattern evaluation: At this step, strictly identify tree patterns representing knowledge.

7) Knowledge representation: This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result. The goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge set of data and interpret useful knowledge and information.

DATA MINING 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 tuple. 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
- Neural Networks
- 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
- 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

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 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.

Types of neural networks

- Back Propagation

Data Mining Application

Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical and so on. Some of the main applications listed below:

1. Data Mining in Education Sector: We are applying data mining in education sector then new emerging field called “Education Data Mining”. Using these term enhances the performance of student, drop out

student, student behavior, which subject selected in the course. Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Use student's data to analyze their learning behavior to predict the results.

2. Data Mining in Banking and Finance: Data mining has been used extensively in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

3. Data Mining in Market Basket Analysis: These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping.

4. Data Mining in Earthquake Prediction: Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance).

5. Data Mining in Telecommunication: The telecommunications field implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highly competitive environment. Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks.

6. Data Mining in Agriculture: Data mining than emerging in agriculture field for crop yield analysis a with respect to four parameters namely year, rainfall, production and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network.

7. Data Mining in Cloud Computing: Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage. Cloud Computing uses the Internet services that rely on clouds of servers to handle tasks. The Data Mining techniques in Cloud Computing to perform efficient, reliable and secure services for their users.

II. CONCLUSION

Data mining is a "decision support" process in which we search for patterns of information in data. This paper provides a general idea of data mining, data mining techniques and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. Decision trees are a reliable and effective decision making technique which provide high classification accuracy with a simple representation of collected KDD. It helps experts to validate and classify the results and outcomes of tests and analyze various new symptoms of diseases based on data. Thus, data mining can help to play an important role in the field of medicine or health care and disease prediction.

III. REFERENCES

[1]. Lior Rokach and Oded Maimon,"Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and

Artificial Intelligence)", ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.
[2]. AnkitaAgarwal,"Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp57-61, April 2012.
[3]. Kalyani et al., International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X ,Volume 2, Issue 10, October 2012.
[4]. UmamaheswariK, SNiraimathi "A Study on Student Data Analysis Using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.
[5]. Yongjian Fu " data mining: task, techniques and application".
[6]. Han and Kamber, "Data Mining and Concepts".
[7]. www.wikipedia.com
[8]. Database Management System "Jaffrey Ullman."