International Journal of Electrical Electronics & Computer Science Engineering
Special Issue - ICSCAAIT-2018 | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

Data Mining: Concepts, Techniques, Methods and Its Applications

Shivani Bansal[1], Vaishali[2], Meenu Gupta[3]
[1-2]PG Scholar, Lingaya's University, Faridabad, Haryana, India
[3]Assistant Professor, CSE Department, Lingaya's University, Faridabad, Haryana, India

Abstract: ***The paper discusses about Data mining and it concepts, It 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, mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. There are a wide variety of applications in real life. This paper talks about few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results. It also focuses on the classification and clustering techniques on the basis of algorithms.***

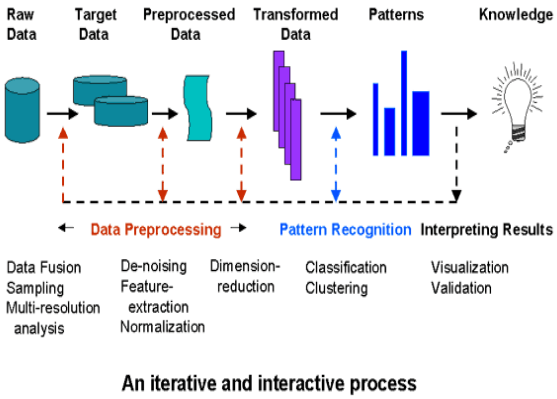
Keywords: Data mining, Algorithms, Clustering, Decision Tree, Data Mining Techniques and Its Applications.

I. INTRODUCTION

The development of Information Technology in these last few yew years has generated large amount of databases and huge data in various areas. Many researches in databases and information technology have given rise to an approach to store this data. This precious data is further manipulated for further decision making. Data mining is a**,** It is the process of extracting the useful data, patterns and trends from a large amount of data by using different techniques. It has other names too such as discovery process, knowledge mining from data, knowledge extraction or data pattern analysis.

The major steps involved in a data mining process are:

- Extract, transform and load data into a data warehouse.
- Store and manage data in multidimensional databases.
- Provide data access to business analysts using application software.
- Present analysed data in easily understandable forms, such as graphs.



An iterative and interactive process

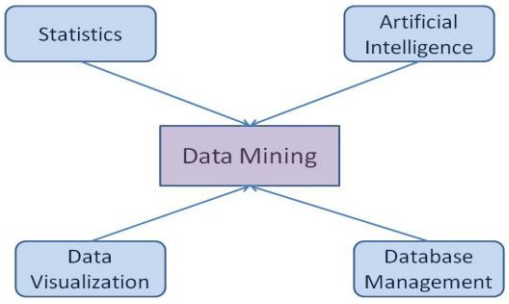
A. *Data Mining Process:*

The data mining process has five steps. In the first step the organizations collect data and load it into their data warehouses. The second step involves the storing and managing the data, either on in-house servers or the cloud. In the third step business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, in the fourth step the application software sorts the data which is based on the user's results, and finally, in the last step the end user presents the data in an understandable and easy-to-share format, such as a graph or table.

B. *Data Mining Software:*

Data mining programs analyse relationships and patterns in data based on what users' request. For example, data mining software can be used to create classes of information. For better understanding imagine a restaurant wants to use data mining to determine when they should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order. In other cases, data miners find clusters of information based on logical relationships, or they look at associations and sequential patterns to draw conclusions about trends in consumer behaviour.

II. TECHNIQUES USED IN DATA MINING



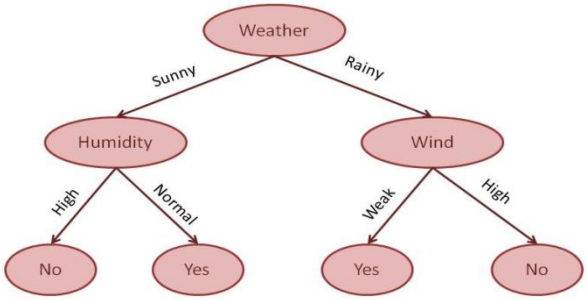
A. *Statistics:*

Statistics includes a number of methods to analyze numerical data in large quantities. Different statistical tools used in data mining are regression analysis, cluster analysis, correlation analysis and Bayesian network. Statistical models are usually built from a training data set. Correlation analysis identifies the correlation of

variables to each other. Bayesian network is a directed graph that represents casual relationship among data found out using the Bayesian probability theorem. Given below is a simple Bayesian network where the nodes represent variables whereas edges represent the relationship between the nodes.

B. Machine Learning:

Machine learning is the collection of methods, principles and algorithms that enables learning and prediction on the basis of past data. Machine learning is used to build new models and to search for a best model matching the test data. Machine learning method uses heuristics while searching for the model. Data mining uses a lot of machine learning methods which includes inductive concept learning, conceptual clustering and decision tree induction. A decision tree is a classification tree. It decides the class of an object by following the path from the root to a leaf node. Given below is a simple decision tree that is used for weather forecasting.



C. Database Oriented Techniques:

Advancements in database and data warehouse implementation helps data mining in a number of ways. Database oriented techniques are used mainly to develop characteristics of the available data. Iterative database scanning for frequent item sets, attribute focusing, and attribute oriented induction are some of the database oriented techniques widely used in data mining. The iterative database scanning searches for frequent item sets in a database. Attribute oriented induction generalizes low level data into high level concepts using conceptual hierarchies.

D. Neural Networks:

A neural network is a set of connected nodes called neurons. A neuron is device that computes the requirement of its inputs and that inputs can even be the outputs of other neurons. A neural network can be trained to find the relationship between input attributes and output attribute by adjusting the connections and the parameters of the nodes.

E. Data Visualization:

The information extracted from large volumes of data should be presented well to the end user and data visualization techniques make this possible. Data is transformed into different visual objects such as dots, lines, shapes etc and displayed in a two or three dimensional space. Data visualization is an effective way to identify trends, patterns, correlations and outliers from large amounts of data.

III. Benefits Of Data Mining

- Automated prediction of trends and behaviors.
- It can be implemented on new systems as well as existing platforms.
- It can analyze huge database in minutes.
- Automated discovery of hidden patterns.
- There are a lot of models available to understand complex data easily.
- It is of high speed which makes it easy for the users to analyze huge amount of data in less time.
- It yields improved predictions

IV. Applications Of Data Mining

Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

A. Market Analysis and Management:

Listed below are the various fields of market where data mining is used –

1) Customer Profiling: Data mining helps determine what kind of people buy what kind of products.

2) Identifying Customer Requirements: Data mining helps in identifying the best products for different customers. It uses prediction technique to find out the factors that may attract new customers.

3) Cross Market Analysis: Data mining performs Association/correlations between product sales.

4) Target Marketing: Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.

5) Determining Customer Purchasing Pattern: Data mining helps in determining customer purchasing pattern.

6) Providing Summary Information: Data mining provides us various multidimensional summary reports.

B. Corporate Analysis and Risk Management:

Data mining is used in the following fields:-

Finance Planning and Asset Evaluation: Cash flow analysis and prediction, contingent claim analysis to evaluate assets.

Resource Planning: It focuses on summarizing and comparing the resources and spending.

Competition: It involves monitoring competitors and market directions.

C. Fraud Detection:

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

V. MAJOR ISSUES OF DATA MINING

1) Mining Different Kinds of Knowledge in Databases: The need of different users is not the same. It is necessary for data mining to cover broad range of knowledge discovery task.

2) Interactive Mining of Knowledge at Multiple Levels of Abstraction: The process should be interactive so that it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

3) Incorporation pf Background Knowledge: To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.

4) Data Mining Query Languages and Ad Hoc Data Mining: It allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

5) Presentation and Visualization of Data Mining Results: Once the patterns are discovered it needs to be expressed in high level languages, visual representations. This representation should be easily understandable by the users.

6) Handling Noisy or Incomplete Data: The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

7) Pattern Evaluation: It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

8) Efficiency and Scalability of Data Mining Algorithms: In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

9) Parallel, Distributed, and Incremental Mining Algorithms: The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed parallel. Then the results from the partitions are merged. The incremental algorithms, updates databases without having mine the data again from scratch.

VI. CONCLUSION & FUTURE WORK

This paper presents a detailed description of data mining techniques and algorithms. Therefore, Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. The various algorithms used for the mining of data are specified in detail. The future scope provides enhancement and efficiency of data in the system. They could lead to better, faster and qualitative exaction of data with better tools and techniques.

VII. REFERENCES

[1] Jiawei Han and Michelin Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd edition.

[2] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[3] VikramPudi,PRadha Krishna “Data Mining”,Oxford University Press, First Edition,2009

[4] Margaret H. Dunham, “Data Mining Introductory and Advanced Topics”,Dorling Kindersley Pvt.Ltd.India,Sixth Edition,2013.

[5] Phyu, Thair Nu. "Survey of classification techniques in data mining."International MultiConference of Engineers and Computer Scientists, 2009.

[6] Buddhinath, Gaya, and Damien Derry."A simple enhancement to One Rule Classification." Department of Computer Science & Software Engineering. University of Melbourne, Australia (2006).

[7] Ali, Shawkat, and Kate A. Smith."On learning algorithm selection for classification." Applied Soft Computing, 2006.

[8] Grabmeier. J, Rudolph. A, “Technique of Clustering Algorithms in Data Mining”, Data Mining and Knowledge Discovery,2002.