RESEARCH ARTICLE OPEN ACCESS

A Survey of Data Mining: Concepts with Applications and its Future Scope

Dr. Zubair Khan[1], Ashish Kumar[2], Sunny Kumar[3]
M.Tech Research Scholar[2].
Department of Computer Science and Engineering[1]
Invertis University Bareilly, 243123
Uttar Pradesh - India

ABSTRACT

In this paper we have to focus on data mining concept and its tools and technology which help us for market perspective to take a proper decision and get a proper result. Data mining is a logical process that is used to analyze large amounts of information that can be in the form of document in order to find important data. The goal of data mining is to find patterns that were previously unknown. Once you have found out those patterns, you can use them to solve number of complex problems. Data mining [sometimes called data or knowledge discovery from data (KDD)] is the process of analyzing data from huge amount of data and summarizing it into useful information. Data mining is one of a number of analytical tools for analyzing data. It allows users to search and analyze data from many different source and transform into decision making data from which user can take decision. IT is the process of finding patterns among dozens of fields in large relational databases. Data mining is a powerful tool because it can provide relevant information. But it is not so easy to get relevant information that can help you to take proper decision. This is where data mining becomes a powerful tool that will help to extract useful information.

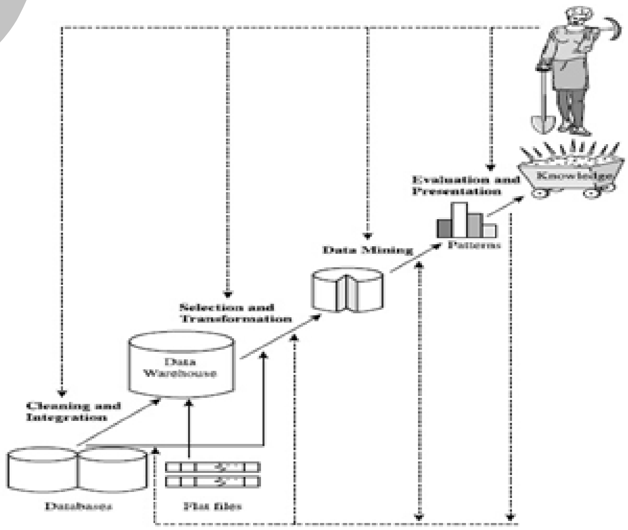
Keywords:- Data Mining, KDD, Data Mining Task, Data Preprocessing, Visualization of the data mining model, Data Mining: classification, methods and its application

I. INTRODUCTION

The primary aim of data mining is to extract the useful information for users from a large amount of data. As the data are available in the different formats such as graphically, audio, video or in the form of varies documents so that the proper action to be taken [1]. Not only to search or analyse these data but also take a good and proper decision for business perspective. When the user will required the data should be retrieved from the database and make the better decision. There is huge amount of complex data but we hardly able to transform them in to useful information and knowledge for managerial decision making for business. To generate useful information it requires massive collection of integrated data. It may be different formats like graphical, audio/video, text, numbers figures, and Hypertext formats. To take complete advantage of large data; the data retrieval from large database is simply not enough for proper decision, it requires a tool for automatic summarization of data, extraction from information stored, and the discovery of patterns in database. With the huge amount of data stored in files, databases, and other repositories system, it is important for that, to develop powerful tool for analysis and extracting the useful of such data or knowledge that could help in decision-making. The only answer to all above these is 'Data Mining'. It is the process of extraction hidden predictive information from large databases; it is a powerful tool and technology with great potential to help organizations focus on the most important information in large data sets. Data mining tools predict future growth and its behaviours.

II. STEPS OF DATA MINING

In general people feel helpless for analysed the large amount of data sets. Data mining can find the useful information that will help to users according to their needs for business perspective to take the proper decision. Data mining is the process of knowledge discovery [2]. KDD as a process is depicted in Figure 1 [8] and consists of an iterative sequence of the following steps:



A. Data Cleaning and Data Integration

Data cleaning is the process of to remove noise and inconsistent data. Fill in missing values, correct noisy data, remove or identify outliers, and resolve inconsistencies.

Data integration means integrate multiple database, files from different various sources. There is no need going to the next step unless complete this step.

B. Data Selection and Data Transformation

The next step is the data selection and then data transformation. Select relevant data from the data base to the analysis task are retrieved from the database.

After selection data must transformed or consolidated into forms appropriate for mining by performing aggregation or summary operations for instance.

C. Data Mining

It is essential process where intelligent methods are applied in order to extract patterns.

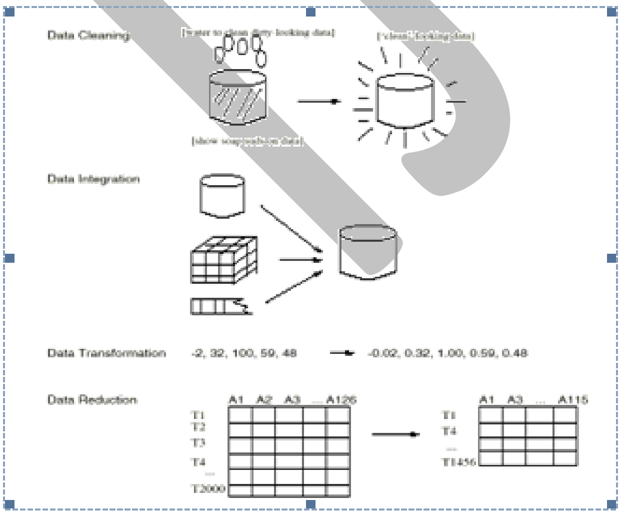
D. Pattern Evaluation and Knowledge Presentation

Evaluating to identify the correct interesting patterns representing knowledge based on some interestingness measures according to the user requirements.

III. DATA PREPROCESSING

In above step A B are different form of data pre-processing, where the data or information are ready or prepared for mining. The data mining task may interact with the user. The truly interesting patterns are presented to the end user and may be stored in the knowledge base. Figure 3 shows the processing of data [9].

Forms Of Data Pre Processing



IV. DATA MINING TASK

The following are the task of Data Mining:-

A. Classification:

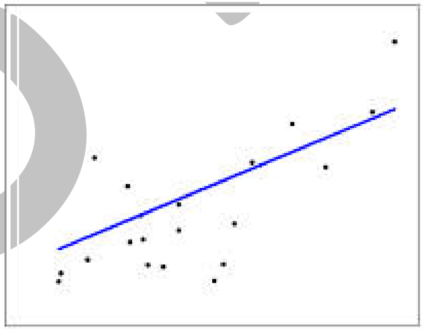
It predicts categorical class labels (nominal or discrete). Learning a function that maps an item into one of a set of predefined classes. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

B. Regression:

It is a statistical process for estimating or predicting the relationships among items or variables. It includes many techniques for analyzing and modeling of several variables, when focuses on the relationship between a dependent variable and one more independent variables.

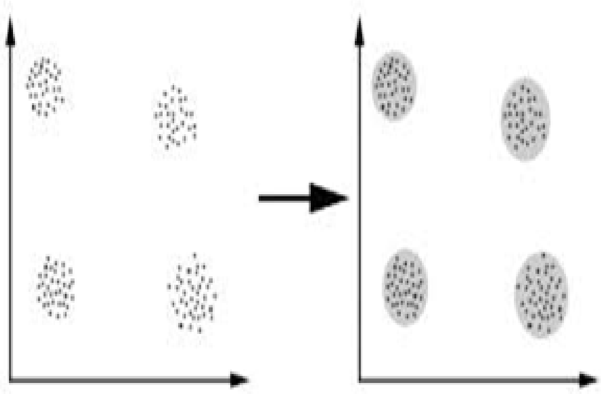
Linear Regression

- w0 + w1 x + w2 y >= 0
- It computes wi from data item to minimize squared error to 'fit' the data
- Not flexible enough



C. Clustering:

The process of identify or grouping a set of physical objects or items into classes of similar objects or we can say that identify a set of groups of similar items.



D. Dependencies and associations:

Identify the significant dependencies between data attributes.

Find those attributes in which dependency are occurred and associates them to each other.

E. Summarization:

Find a summarized or compact description of the dataset or a subset of the dataset.

V. VISUALIZING DATA MINING MODEL

The main objective of visualization of data is the overall idea about the data mining model .In data mining most of the times we are extract the data from the data repositories which are in the hidden form. It is very difficult task for an end user. So this visualization of the data mining model helps us to provide topmost levels of understanding and trust. Because the user does not know what the data mining process has discovered [1].

The data mining models are categorized in two types:

- Predictive and Descriptive.

The predictive model makes prediction about unknown or missing data values by using the known values. Ex. Classification, Regression, Prediction, Time series analysis etc.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Association rule, Summarization, Sequence discovery etc.

Many of the data mining applications are aimed to predict the future state of the data.

- Prediction is the process of analyzing the current and past states of the variables or attribute and prediction of its future state.
- Classification is a data mining technique of mapping the target data to the classes or predefined groups, this is a supervise learning because the classes are predefined before the examination of the target data.
- The regression technique involves the learning of function that map data item or data set to real valued prediction variable.
- In the time series analysis technique the value of an attribute or item is examined as it varies over time.
- The term clustering means analyzes the set of different data objects without consulting a known class levels. It is referred to as unsupervised learning or segmentation. It is the segmentation or partitioning of the data set in to similar type of groups or clusters. The clusters are defined by grouping of similar type of objects into one cluster. The term segmentation is a process of partitioning of database into disjoint grouping of similar tuples.
- Summarization is the technique of representing summarize or accurate information from the data.

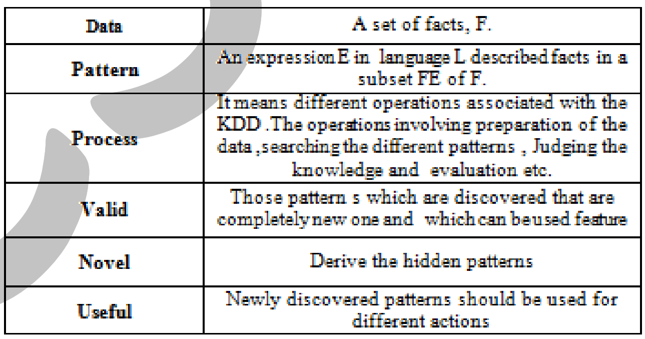
The association rule finds the association between the different attributes.

- Association rule mining is a process in two-step: Finding all frequent item sets, Generate strong association rules from the frequent attributes or item sets.
- Sequence discovery is a process of finding the Sequence patterns in data set. This sequence can be used to understand the trend.

1) A new way to defined the KDD process

We have found the broader meaning of the followings Patterns, data, Process, Valid, Novel, and Useful Understandable of KDD. The Knowledge discovery in data repository or databases is the non-trivial process of identifying valid, useful, novel, and ultimately understandable patterns in data.

TABLE I. TO DESCRIBE THE NEW FORM THE WORD



Data	A set of facts, F.
Pattern	An expression E in language L described facts in a subset FE of F.
Process	It means different operations associated with the KDD .The operations involving preparation of the data ,searching the different patterns , Judging the knowledge and evaluation etc.
Valid	Those pattern s which are discovered that are completely new one and which can be used feature
Novel	Derive the hidden patterns
Useful	Newly discovered patterns should be used for different actions

VI. DATA MINING METHODS

Following are the popular data mining methods:

a. Decision Trees and Rules
b. Nonlinear Regression and Classification Methods
c. Example-based Methods
d. Probabilistic Graphical Dependency Models
e. Relational Learning Models

These are some famous data mining methods are broadly classified as: On-Line Analytical Processing ,(OLAP), Classification, Association Rule Mining, Clustering, Temporal Data Mining, Time Series Analysis, Web Mining, Spatial Mining, etc. These types of methods use different types of algorithms and data. The data source coming from data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest

Neighbour, Neural Network based (ANN), Genetic Algorithms based, Ruled based, and Support Vector Machine etc.

Generally the data mining algorithms are dependent of the two factors these are:

(i) Which type of item or data sets are using
(ii) What type of requirements of the end user?

Basing upon these above two factors the data mining algorithms are used. A knowledge discovery (KDD) process involves pre-processing of data, choosing a data-mining algorithm, and pre- processing the mining results.

VII. DATA MINING APPLICATIONS

We have focused on the applications of data mining and its techniques are analyzed respectively Order[3]. They are categorized in the following way:

- Health Care
- Business
- Web Education
- Biological Data Analysis
- Intrusion Detection

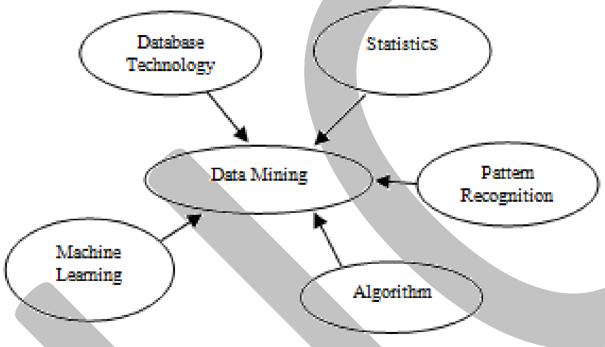


Fig .1 Data Mining: confluence in many discipline

A. Data Mining Application in Health care

In the medical and health care areas Data Mining application can have tremendous potential and usefulness, due to regulations and availability of computers, a huge amount of data is becoming available. Such a large amount of data cannot be analyzed or processed by humans in a short period of time to make diagnosis, prognosis and to make treatment schedules .This problem is overcome with the help of data mining so that applications of data mining in this field results timely and accurate decisions for diagnosis the treatment.

B. Data Mining In Business Analysis

In business, bank as well as financial institutions offer a wide range of services so the useful information collected is said to be complete, reliable , accurate and high quality where data mining need to provide security to help in fraud detection.

C. Web Education

Data mining methods are used in the Web Education field which is used to improve courses in institutions. The relationships are discovered among the usage data selected during students' sessions. This knowledge is very useful for the faculty or teacher and the author of the course, who could decide what modifications will be the most suitable or appropriate to improve the effectiveness of the course.

D. Biological Data Analysis

In the biomedical field analysis ranging from the development of pharmaceuticals and in cancer therapies to the identification and analysis of human genome by discovering the large scale of sequencing patterns and genetic functions so the data mining applications helps to do DNA analysis for the discovery of genetic causes for many diseases with the help of large databases.

Intrusion Detection

Data Mining is one of the most popular techniques for detecting intrusion. It can be classified on the basis of their strategy of detection. Data mining technologies applied to intrusion detection to invent a new pattern from the massive network data as well as to reduce the stains of the manual complications of the intrusion [1].It is helpful in detecting new vulnerabilities and intrusions, it discover previous unknown patterns of attacker behaviours and also provide decision support system for intrusion management

VIII. SCOPE OF DATA MININGS

Data mining extract the useful information and provide accurate data for decision making. Data mining help in predict the future trends and its behavior for the business purpose. It extract valuable business information in a large database for example, finding linked products in gigabytes or terabytes of store scanner data .Given databases of sufficient size and good quality, data mining technology can generate new decision making business opportunities by providing these capabilities.

Data Mining automates the process of finding predictive information from large data bases. It uses the current or past promotional mailings data to identified the most likely to maximize the return on investment on future mailings. On the other hand Data Mining technologies detect the fraud detection and identifying segments of population likely to respond of similar events or task.

a) Artificial neural networks:

It is Non-linear predictive models that learn through training set and resemble biological neural networks in structure.

b) Decision trees :

Tree-shaped type structures that represent sets of decisions. Each node is judgment and separately represent for the decision. These decisions generate rules for the classification of a dataset.

c) Genetic algorithms:

Genetic algorithm is optimization techniques that use process such as genetic combination, natural selection and mutation in a design based on the concepts of evolution.

d) K-Nearest neighbor method:

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. It is technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.

IX. CONCLUSIONS

In this paper we briefly reviewed the various data mining concepts, its techniques and applications. Data Mining is not a new term, but in the recent years its growth day by day touches great horizons. It has spread in almost all areas nowadays. It is clear that Data Mining tools helps in extracting useful or meaningful knowledgeable attributes or information from the unimaginable massive data. This review would be helpful for the researchers to focus on the various issues of data mining. In future, we will review the popular classification algorithms and significance of their evolutionary computing approach in designing of efficient classification algorithms for data mining.

REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra “The Survey of Data Mining Applications And Feature Scope” International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[2] Zhang Xiaodan Shenyang University of Chemical Technology “Plain Discussion of Data Mining Technology Research” 2011 IEEE

[3] V.K.Deepa “Rapid Development of Applications in Data Mining” Proceedings of 2013 International Conference on Green High Performance Computing March 14-15, 2013, India

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

[5] Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2,John Wiley & Sons, Inc., 2005.

[6] https://www.google.co.in/search?q=kdd+process

[7] https://www.google.co.in/search?q=forms+of+data+preprocessing.

.