International Journal on Emerging Technologies (Special Issue NCETST-2017) **8**(1): 373-376(2017)
(Published by Research Trend, Website: www.researchtrend.net*)*

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

DATA MINING: A Comparative Study on Data Mining techniques

Ms. Garima Joshi and Ms. Preeti Pandey
Assistant Professor, Amrapali Institute of Technology Haldwani, (Uttarakhand), India.

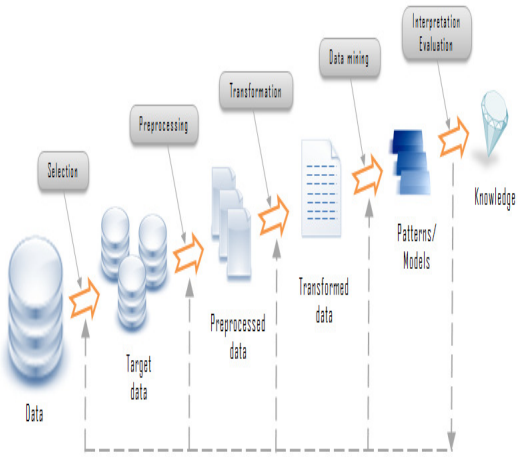
ABSTRACT: A Data Mining is a process of finding the patterns/ Knowledge from a given large set of data. In general, it can also be defined as, about how does one store, access, model, describe and understand very large amount of data sets. Various data mining techniques are presented which are used to extract the patterns out of the data sets, depending upon the conditions where they are been applied. This paper gives a brief study of different majorly used techniques in mining the data and introduces the best applicable environment for them. The paper provides comparative study on most common data mining methods.

I. INTRODUCTION

The field of data mining is finding its usage in almost every real world domain whether it is education, medical & health, business, science etc. The size of data in each field is increasing exponentially with time. Data mining is basically a method which provides us the best usage of this collected data over time series. Data mining has been used different methods with the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems. The objective of the data mining is to find out the knowledge from the large data sets and then convert them into human understandable patterns so that this knowledge can be used effectively in future. Data mining techniques which extract information from huge amount of data have been becoming popular in every real world domains, finding its major application in education and medical science. Suppose we have to analyze a new patient, depending upon his symptoms we can check in the hospital past records that if some patient will have that symptoms what treatment should be best to him so that he can recover quickly. Similar is the case of education depending on the past track/record we can come to know the best techniques which can help a student at its best in present or in future.
Section 2 describes literature review in brief. Section 3 explains the concept of data preprocessing. Section 4 contains introduction to data mining and types of data which can be minied. Section 5 summarizes the comparative analysis of different data mining techniques. Conclusion is shown in section 6 while references are mentioned in the last section.

II. LITERATURE REVIEW

Data mining is a method of solving a problem by analyzing a data which is already present in the database. This process generally known as knowledge discovery process. Knowledge is discovered in the form of patterns.



These patterns can be seen as a kind of summary of the input data and may be used in the future analysis. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. In the 1960s, statisticians had a bad practice of analyzing data without an a-priori hypothesis which is known as "Data Fishing" or "Data Dredging".

The term "Data Mining" appeared around 1990 in the database community which later on became more widespread in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably. We can say that data mining is a combination of statistics, Artificial Intelligence and Database research.

III. DATA PRE-PROCESSING

The actual objective of the mining process is the automatic or the semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as group of data records (Cluster analysis),finding unusual records(anomaly detection) , dependencies(association rule mining, sequential pattern mining) etc. The term data mining is a misnaming because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself.

Data Mining is a process which is totally dependent on data .Before using the data we should be aware about the following-

- The purpose of the data collection.
- How the data will be used
- Who will be able to mine the data and use the data
- The status of the security which is provided to the data
- How the collected data can be updated

Once we are done with the data collection part data preprocessing is done. Before data goes to the mining phase it is pre-processed. Real world data which is collected for the mining purpose that is not suitable, so it is preprocessed.

Data gathering methods are generally loosely resulting in general out of range values (income: -500 ; age: 150 etc),missing values(for some attributes values are not given generally in database theses values are replaced by NULL) etc. This type of collected data may lead to misconception, and hence affecting the results. Thus the quality and representation of data is very important before running an analysis.

Data pre-processing includes following steps:

- Data cleaning
- Missing values
- Noisy data
- Data Integration
- Use of Metadata
- Data transformation
- Normalization
- Smoothing
- Aggregation
- Generalization
- Data Reduction
- Data cubes aggregation
- Dimension Reduction
- Data compression
- Numerosity Reduction
- Discretization & Concept hierarchy generation

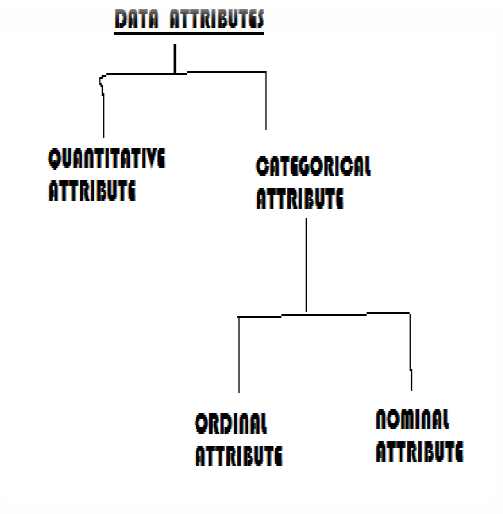
Above defined steps are implemented to make the data more accurate. The outcome of the data pre-processing is the final training data set in which the mining is to be applied. Hence it is the most important part of any mining process without preprocessing we will not be able to achieve our desired objective.

V. DATA AND DATA MINING

Data mining techniques are applied on the data. Hence data is represented in the data base in the form of table. A table consist of rows and columns where attributes plays very important role in storage of data inside a table. Data attributes are classified into two categories -- Quantitative data and Categorical data. Quantitative data is that data which takes number as their key values example age, income, marks, phone number etc. While Categorical data are further classified into two types Ordinal data and Nominal data.

Ordinal data attributes are those attributes which follow a natural order example qualification i.e, first a student will do 10^{th}, then 12^{th}, UG, PG and so on.

Nominal data attribute is an attribute that is having categories to hold, at a time it will able to hold only one value example Marital status i.e one is married or unmarried, gender –male or female etc.



Selection of data attribute while designing a database plays a vital role, in future accessing of data. Attributes generally describes the overall instances in the row of any datawarehouse.

Data mining techniques are the methods/algorithms which are applied on the database to extract out the knowledge, this process is known as knowledge discovery process in data mining.

Data mining techniques are broadly classified into two parts :

- Descriptive Data Mining
- Predictive Data Mining
- Prescriptive Data Mining

Descriptive Data mining analyzes the data on the basis of the past events and act accordingly for the future. Mining the historical data ,finding out the reasons for its success or failure and applying those analysis on the present data to come with the best and optimal results in the future.This technique is also known as post-mortam analysis. Area where it finds its applications are almost in every reporting like finance, marketing, sales & operations etc.This technique helps in finding out the relationship between data in such a way that it is used to classify customers or prospects into various different groups or categories. For example, descriptive mining technique analyzes past electricity usage data to help plan power needs and allow electric companies to set optimal prices accordingly in the future.

Descriptive Mining technique is further classified into following categories:

- Clustering
- Association Rule Mining
- Sequential Analysis

Predictive Data Mining technique converts the data into valuable and actionable information .It combines variety of techniques together like statistics, machine learning, game theory ,AI etc on the historical and present data and predict the future events. This mining technique is generally used to model out the relationship between different factors and then allow assessment of risk and potential associated with a particular set of condition , and hence allowing decision making for any situation.

Predictive analytics can be classified into two major categories:

- Classification
- Decision tree
- Rule Induction
- Neural Network
- Nearest-neighbor classification
- Regression

Prescriptive Data Mining is generally used to predict the future. It is more better and advance technique as compared to that of the above discussed techniques.

It is generally used to predict the outcomes by automatically synthesizing big data, mathematical sciences, machine learning etc and then suggest the different decision options to take advantage of those predictions and hence allowing the decision makers to know the implication of each decision option. We can say that prescriptive technique not only determines what will happen, when it will happen but also determines why it will happen. It provides a better decision support system where chances of failure gets minimized.

For example it can be used in determining benefits of healthcare, energy and utilities etc.

VI. DATA MINING TECHNIQUES

a) Classification: It is a supervised mining technique which is used to determine the categorical target. This method is mainly used in the bank systems to identify whether the loan applicant as high medium or low credit risk. Classifications are generally discrete in nature and do not necessarily support any order. There are different methods used for implementing a classification technique in data mining:

1. DECISION TREE
2. BAYESIAN CLASSIFICATION
3. NEURAL NETWORK

b). Clustering: It is an unsupervised mining technique, mainly used to discover groupings in the data. It finds out the same types of data elements together on basis of some common properties hold by them and form one group. This group is known as a cluster. The elements which lie inside a cluster are known as inliers and those which lie outside a cluster are known as outliers. The clusters are designed in such a way keeping in mind that the inter cluster similarities are very low and the intra cluster similarities are high. This process of forming a cluster is sometimes refer as detection.

c) Association: It is an unsupervised mining technique which is genrally used to determines set of rules. These rules are known as association rules. The main application area of this technique is sales transaction. This analysis is termed as market basket analysis. In this we study about the items purchased by a customer during his visit. A customer who purchased item A and B is 80% likely to purchase item C. On the basis of these analysis rules are made and implemented in the system.

d) Prediction: It a supervised data mining technique which is generally used to predict the result from the given set of historical data. It directly use the data to

determine the class value of a new instance. This technique finds its major application in weather forcasting. It works on the principle of predictive model.

It is made up of predictors which are generally variable factors that are likely to influence the future behavior or result. For example in any sales system on the basis of a users age, gender and purchase history we can predict the future profit.

e) Time series analysis: This analysis works on time series data. It is used to extract the statistical property hidden inside a data element. Time series data is data which is very much influenced by the time period. We can also say that the data which changes with the time is also known as time series data. This analysis is done to predict the future values depending upon the observations done on the previously observed values.

f) Sequential Patterns : It is method of finding relevant patterns between the data. It generally works on the discrete values. It is a type of structured data mining. This method find its best application in building efficient and effective databases ,indexes for sequential information recovering from missing sequence etc.

VII. CONCLUSION

This paper gives us a basic understanding of what data mining is, areas where it is applied and purpose of applying the mining technique. Introduces with the different types of techniques which are generally used .Depending upon the data environment i.e., supervised or unsupervised a particular technique is applied.

REFERENCES

[1]. Bharati M. Ramageri, "Data Mining Techniques and Application", *Indian Journal of Computer Science and Engineering*, Vol. **1** No. 4; pp. 301-305.

[2]. Velmurugan T. et al., "Performance Evaluation of K-Means & fuzzy C-means Clustering Algorithm for Statistical Distribution of Input Data Points", European Journal of Scientific Research, Vol. 46, 2010.

[3]. Kavitha P., T. Sasipraba, "Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining", International Journal on Computer Science & Engineering (IJCSE),

[4]. Hamidah Jantan, "Classification and Prediction of Academic Talent using Data Mining Technique", 14 *International Conference on Knowledge based and Intelligent Information and Engineering Information* pages 491-500.

[5]. Tai Chang Hsia, "Course Planning of Extension Education to meet Market Demand by using Data Mining Techniques", *Expert System with Applications: An International Journal*, Vol. **34**, Issue 1, Jan 2008.

[6]. Jiawei Han and Michelire Kamber, "Data Mining Concept and Technique", Published by Morgan Kaufman, 2006.

[7]. Monika Goyal and Rajan Vohra, "Application of Data Mining in Higher Education", *International Journal of Computer Science (IJCSI)* Issues, Vol. **9**, Issue 2, No.1, March 2012; pp-113-120.

[8]. Arun K Pujari, "Data mining Technique", Published by Universities Press (I) Pvt. Ltd, Hyderabad, India.

[9]. Gajendra Sharma, "Data mining and Data Warehousing and OLAP", Published by S.K. Kataria & Sons, New Delhi, India.

[10]. Arvind Sharma *et al*., "Data Mining Techniques and Their Implementation in Blood Bank Sector-A Review", *International Journal of Engineering Research and Application (IJERA),* Vol. **2**, Issue-4, July-August 2012; pp.1303-1309.

[11]. R.K. Somani, "Data Mining & Warehousing", College Book Centre, Chaura Rasta, Jaipur, India.

[12]. Venkatadri.M, "A Review on Data Mining from Past to Future", *International Journal of Computer Applications (IJCA)*, Vol. **15**, No.7, Feb 2011.