MAT JOURNALS
Journal of Data Mining and Management
Volume 2 Issue 3

A Quick Review on Data Mining Techniques

Er. Ankit Sharma[1], Er. Surabhi Jain[2]
[1]Assistant Professor, CSE Department, Baba Farid College of Engg. And Tech., India
[2]Assistant Professor, Computer Department, Central University of Punjab, India

Abstract
In this paper, we studied about the concept of data mining and their techniques. Data mining is a process to extract the pattern in the data from the huge amount of data. For data mining there are some techniques those are apply on the huge amount of data and extract the useful information. In this paper we can study about the categories of these techniques of data mining also to improve their businesses and found excellent results.

Keywords**:** *Association rules, Classification, Clustering, Data Mining*

INTRODUCTION

IN the Data warehousing there are collection of many databases and huge data in every field. For use that data in further decision making the data mining on the basis of precious data. Data mining is a technique for extracting the useful information from the huge amount of data or data warehousing .data warehousing is a collection of data bases or data warehouse is like a store for storing the large amount of data. Data mining is used to extract the useful data or finding the patterns from the existing data for further used for making decisions to increase the business profits. Basically data mining is a process for discover the knowledge or check the pattern of the previous data. Once the patterns are found then these patterns are helps to make the decisions for the future strategies. It is also known as knowledge discovery through data (KDD).

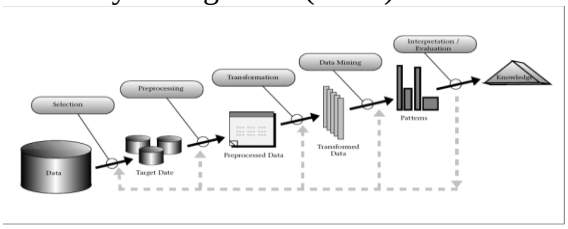

Fig. 1 *Knowledge Discovery through data (KDD) in Data mining*

For mining the data from the large amount of data we have following steps are:

- Exploration
- Pattern identification
- Deployment

Exploration: Exploration is a process for cleaned the data (means in well-mannered form or required data form), then transform the data into another form and after that determine the behavior of that data.

Pattern identification: After the exploration we can find the patterns of our previous data and choose the pattern which is make the best prediction.

Deployment: After pattern identification the patterns are deployed for best outcome.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Data Mining Techniques

For mining the data from huge amount of data we have some applications and

techniques. Data mining techniques like a Classification, Clustering, Regeration, Association rules etc. all are used for discover the knowledge or pattern from the precious data.

A. Classification

In classification we have two types of data first one is training data and second one is test data. In training data there are some predefined classes are there for classify which type of data is stored in which class. And test data is which have unknown class and we can apply a classifier algorithm on that data and extract the information about that test data is belongs to which class. For example: From a set of diagnosed patients ,who serve as the training set, a classification model can be built ,which concludes a patients disease from his/her diagnostic data. The classification model can be used to diagnose a new patient's disease based on the patient's diagnostic data such as age, sex, weight, temperature, blood pressure etc.

Types of classification algorithms :

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks

Decision trees: In data mining, decision trees are used for describe the data, but not make decisions. Through decision trees, resulting classification is used as an input for making the decisions.

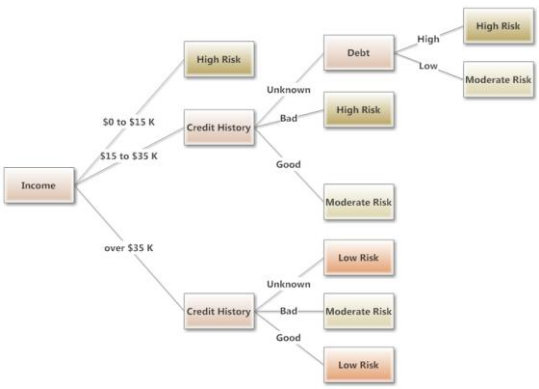

***Fig. 2** Classification by decision tree induction*

Bayesian classifications: Bayesian network can specify joint conditional probability distributions. There are two types of probabilities in baye's theorem –

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

Where X is data tuple and H is some hypothesis.

According to Baye's Theorem,

$P(H/X)= P(X/H)P(H) / P(X)$

Neural networks: Neural network is a set inputs and outputs units those are connected with each other and each connection has a weight. Neural networks are non-linear statistical data modeling tools. These are used to model complex relationships between inputs and outputs of the data or to find patterns in the existing data.

Using these algorithms we can classify the data which data is belongs to which class.

B. Clusters

Clustering is a process of portioning a set of data or objects into subclasses called clusters. Clusters are helps to users for understanding the structure of data set .Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped one class and dissimilar objects are grouped in another class. Clustering is a identification or verification of classes, also called a groups .Clusters are identify the set of objects whose classes are unknown.

Following the types of clusters are:

- Partitioning Methods
- Density based methods
- Grid-based methods
- Model-based methods

Portioning method: portioning method is constructing a portion of documents into a set of clusters. Partitioning methods are relocate data by moving from one cluster to another, starting from an initial partitioning. Such methods are required the number of clusters will be pre-set by

the user.

Density based method
Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution (Benfield and Raftery, 1993). The overall distribution of the data is assumed to be a mixture of several distributions.
Grid Based Method: This method divides the data into the form of rows and columns know as grid. This makes the clustering very easy.
Model based clustering: This method follows a well-defined formula for portioning.

C. Association Rules

Association discovers the probabilities of the co-occurring items are expressed as association rules Association rules are often used to analyze sales transaction. An example of an association rules would be "if customer buys a Burger, he is 80% likely to also purchase coke". This application is association modeling is called Market Basket analysis. It is valuable for direct marketing, sales promotion. Association and correlation is usually to find frequent item set findings among large data sets. Association Rule algorithms need to be able to generate rules with confidence values less than one.

Types of association rule:
- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Multilevel association rules: In multiple level association there are items are described in hierarchy way. When items are expressed in the lower level of hierarchy then the support of the item is going lower also. The transaction of data base can be encoded based on dimensions and levels. Multiple level association rule mining can work with two types of support- Uniform and Reduced.

1. Uniform Support: The minimum support threshold gives the outcome and is used at each step. The minimum support threshold has to be appropriate.
2. Reduced Support: In this approach reduced minimum support is used at lower levels

Multidimensional association rules: Items in the rule refer to two or more dimensions or predicates. e.g., "buys", "time_of_transaction", "customer category". In the following example: nationality, age, income. There are two categories of multi-dimensional association rules:

Inter-dimension association rules: in this there are no predicates are repeated and combination or attributes are interrelated. For example:.

$$age(X,"19\text{-}25") \wedge occupation(X,"student") \Rightarrow buys(X,"coke")$$

Hybrid-dimension association rules: in this category predicates are repeated and attributes or objects are not related with each other. For example:

$$age(X,"19\text{-}25") \wedge buys(X, "popcorn") \Rightarrow buys(X, "coke")$$

Quantitative association rules: These rules are used for mining the large database of sales transactions. In the Database all the transactions consists the identifier of the customers, the items bought. In the transaction and the quantity associated with each purchased item.

D. Regression

This technique is utilized to predict number values like age, height, etc. Regression technique can be analysis the model the relationship between one or more independent variables and dependent variables.

Types of regression methods:

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

Out of these few are:

Linear regression: Linear regression is the most basic and commonly used predictive analysis. Regression determines the relationships between two items.

Multivariate Linear Regression: For predict the value of one or more responses from a set of predictors use the regression analysis. This is also helps to estimate the linear association between the predictors and Reponses. Predictors can be continuous or categorical or a mixture of both.

CONCLUSION

Now a days, in business filed there are huge amount of data To use that data in an efficient manner for making the decisions to increase the sales of the products based on the finding pattern from the large amount of data .Data mining techniques are such as classification, clustering, association rules and reggeration those all are used for help to find the patterns and discover the knowledge form the large data amount or data warehouse. Data mining techniques are further divide in some categories those are categorized on the basis of which type of data we discover from the data warehouse.

REFERENCES

1. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
2. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.
3. Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP 0800.pdf.
4. Customer Successes in your industry from http://www.spss.com/success/?source=homepage&hpzone=nav_bar.
5. https://www.allbusiness.com/Technology /computer-software-data-management/ 633425 1.html, last retrieved on 15th Aug 2010.
6. http://www.kdnuggets.com/.
7. R. Agarwal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. In Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining), 2000.
8. H. Bhavsar, A. Ganatra,"Variations of Support Vector Machine Classification: A survey", International Journal of
9. Advanced Computer Research, Volume 2, Number 4, Issue 6 (2012) 230–236.
10. [10] Ms. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu "Survey on common Data Mining Classification Techniques", International August 2012, ISSN: 2319 – 1058.