STUDY OF DATA MINING METHODS AND ITS APPLICATIONS

S.A.R. NIHA

M.Tech (CSE) Computer Science Department, MuffakhamJah College of Engineering and Technology, Hyderabad, India.

---***---

Abstract - *Data mining is a process which finds useful patterns from large amount of data. The paper discusses classification of data mining techniques and the applications in real world. It shows how data mining can be used to find solutions of business problems and improving them.* It is also called as knowledge discovery process

Key Words: Data mining methods; Data mining algorithms*;*

INTRODUCTION:

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

Data Mining is as process developed to explore large amounts of data to discover useful patterns . The goal of data mining is extracting knowledge from large database, which can be used in major decision making business applications.

Below are the steps involved in the knowledge discovery process

Data Selection- In the first step ,data is cleaned by removing the inconsistent and relevant data.

Data Preprocessing - Relevant and useful data is selected from multiple sources, combined together and prepared for mining

Data Transformation – In this step, data is transformed into appropriate forms for mining by performing various operations such as aggregation etc.

Data Mining – In this step, methods are applied to extract data patterns.

Pattern Evaluation – In this step, data patterns are evaluated and knowledge is represented.

Data mining is the exploration and analysis of large quantities of data in order to discover , valid and potentially useful knowledge hidden in the database.

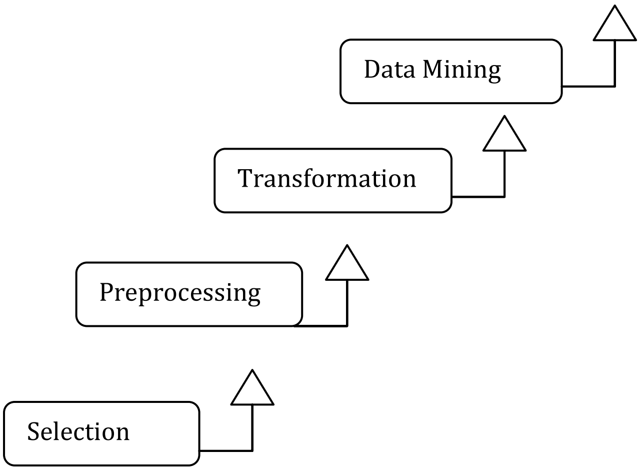

Figure 1: Knowledge Discovery Process

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

2. DATA MINING TECHNIQUES:

There are several major data mining techniques developed. The data mining tasks can be classified into two types . The two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave. There are a number of data mining tasks such as classification, prediction, association etc, which can be grouped either as predictive data mining tasks or descriptive data mining task

Predictive data mining is used to predict unknown or future values of another data set . A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task. Descriptive data mining tasks usually finds data describing patterns and comes up with new, significant information from the available data set.

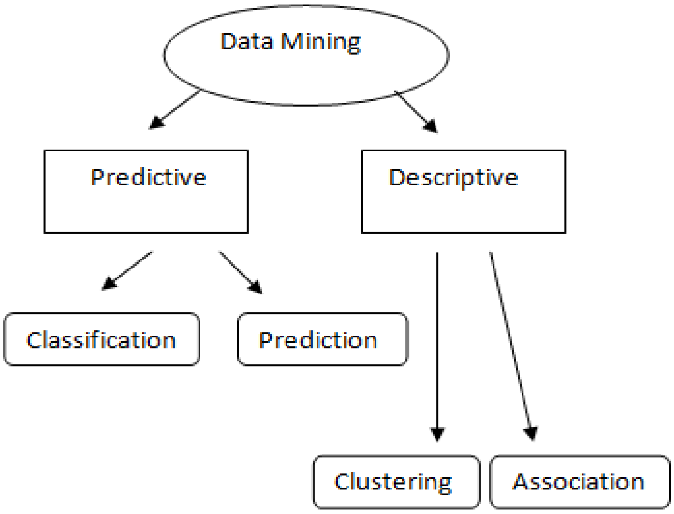

Figure 2: Classification of Data Mining Techniques

For example a doctor using medical test results tries to diagnose a disease can be considered as a predictive data mining task. A retailer identifying products that are purchased together based on customer purchasing behavior can be considered as a descriptive data mining task.

We will briefly discuss those data mining techniques in the following sections.

2.1. Classification:

Classification is the data mining technique in which classifies the dataset into different classes. The classes are called the classifiers. For example, books can be classified based on the genre. The classified based on the similar attributes. Examples include science fiction, satire, drama, action , adventure ,romance, mystery, horror, self help. In medical field a classifier can be used to analyze health conditions of the patient. It is used to predict the risk factor involved and how critical the patient is. It can be also used in banking sectors, where you can classify the loan applications for loan approval, fraud detection etc

The various types of classification techniques are:

- Classification by decision tree induction
- Classification Based on Associations
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)

2.2. Clustering:

Clustering is grouping of similar objects to form a cluster. A cluster is a collection of similar entities which are distinguishable from other entities which belong to other clusters. can be said as identification of similar classes of objects. By using clustering we can identify overall distribution pattern and correlations among data entities. Classification approach can also be used for distinguishing groups For example, segregation of customers based on purchasing patterns. They are different types of clustering methods, which includes the following:

- Partitioning methods
- Hierarchical clustering
- Fuzzy clustering
- Density-based clustering
- Model-based clustering Grid-based methods

2.3. Prediction:

It is used to predict unavailable data values. Regression Analysis is usually used for prediction. Prediction can also be used for identification of trends based on data. Regression analysis is used to show the relationship between independent variables and dependent variables. In data mining independent variables are known attributes and the attributes we predict are called response variables are what we want to predict. Sometimes, Unfortunately, many realtime problems cannot be predicted. Hence, here combination of data mining methods can be used, for example both regression and classification can be used.

Types of regression methods are as follows

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

2.4. Association rule Mining:

Association rule mining is finding correlation or to find frequent item sets among data sets. Such information is helpful to make certain decisions for marketing activities, such as catalogue design, marketing and analysis of customer shopping. Let us consider an example of transactional dataset, if a customer buys bread, he is more likely to buy milk or eggs. This information is used for designing marketing strategies Association Rule algorithms need to be able to generate rules with confidence values less than one. Types of association rules are as below

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

2.5. Time- Series Analysis

Time Series a method in which there is a sequence of events, where the next event is determined by the preceding events.

Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.

3. DATA MINING APPLICATIONS:

In this world of digitalization and computerization, there are huge amounts of data we come across everyday. Data mining is used are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns that would be difficult to find.

3.1 Medicine and Healthcare:

Data mining is used in many fields. We see the applications of data mining in various domains such as health care, hospitals, business, banks, education, fraud detection, corporate firms etc. There are various data and analytics best practices in data mining holds great potential in various domains. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

3.2 Biological Data Analysis:

In recent times, we have seen growth in the field of biology. such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis

- Analysis of heterogeneous and homogeneous information
- Analysis of comparative and similar nucleotide sequences.
- Study of structural patterns of genetic and protein pathways.
- Association and path analysis.
- Genes finding and analysis.
- classification of plants and animals given their features
- Study of life sciences such as medicine etc
- Exploring cellular information

3.4 Education:

In the field of education, data mining can be used in knowledge discovery. We can use the techniques for students' learning. In predicting the students performance. You can develop methods that discover knowledge from data originating from educational Environments. The goals are identified as predicting students' future learning behavior, studying the effects of educational support, and increasing scientific knowledge about learning. Data mining can be used by educational institutions to take accurate decisions and also to predict the results of the student. With the prediction of the performance of the students, the teachers can focus on concentration of teaching. We can use the learning pattern to develop techniques to teach them.

3.5 Sports:

Now a days there is lot of importance given to sports. Tremendous amount of data and statistics about sports and players are collected. From cricket, hockey scores, basketball, golf, tennis, swimming, chess etc, all the data are stored. This information can be used for reporting and to improve performance by the trainers.

3.6 Customer Behavior Analysis:

Analysis of the behavior of a customer is done so as to acquire new customers, that Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a relationship with a customer a business need to collect data and come up with certain strategies and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. them to learn more about their customers and make smart marketing decisions.

3.7 Fraud Detection:

With huge amounts of information, there also comes a risk of fraud. Billions of cash have been lost to the action of frauds. Hence data mining plays a role in fraud detection. Traditional methods of fraud detection are time consuming and complex. A fraud detection system should protect information of all the users. The information can be classified into fraudulent or non-fraudulent and the fraudulent information can be detected

3.8 Financial Data Analysis:

The data in banking and financial industry is very prone to leakage. Hence generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Prediction of eligible candidates for loan sanction analysis and data mining.
- Loan payment and credit card prediction and customer credit policy analysis.
- Classification of customers for targeted marketing.
- Detection of financial crimes.
- maintaining a profitable customers.

3.9 Intrusion Detection:

Any action which effects the integrity and confidentiality of an entity is an intrusion. The system which uses defensive methods to avoid an intrusion is called Intrusion Detection System (IDS). Data mining can help in aiding intrusion detection , to identify , hidden data to identify abnormal activity of interest for the user effectively. The plays a vital role in detecting anomalies and attacks in the network

3.10 Customer Segmentation:

Data mining helps in market research to segment customers effectively. Data mining helps to differentiate customers into various distinct segments according to the purchasing behavior of the customers. This information can be used to retain the customers for a business by offering them with special offers and enhance satisfaction.

3.11 Research:

Advancement in technology has brought evolution in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can extract find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known .Research is used to understand various issues A research study is a scientific way to understand various issues, to improve or develop new methods of health care. Research studies are important because they produce to knowledge which would be helpful to mankind

4. CONCLUSION:

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in identifying the patterns and to predict the trends in any domain. Data mining has wide applications in almost every industry where huge amounts of data is found

5. References:

1. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

2. Sunil Kumar Khatri, Intrusion Detection using Data Mining, IEEE 2015

3. Patricia B. Cerrito, University of Louisville, Louisville, KY (2008), The Difference Between Predictive Modeling and Regression published by www.mwsug.org

4. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

5. Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.

6. Customer Successes in your industry from http://www.spss.com/success/?source=homepage&hpzone=nav_bar.

7. https://www.allbusiness.com/Technology /computer-software-data-management/ 633425-1.html, last retrieved on 15th Aug 2010. 6. http://www.kdnuggets.com/.

BIOGRAPHIES

S. A .R. Niha holds a Master's degree in Computer Science Engineering and Technology from Osmania University, India. Her Research interests include Data Mining and its applications.