Modern Data Mining: Tasks, Techniques, and Applications

J.Priya
Assistant Professor,
Department of Computer Science and Engineering,
Samskruti College of Engineering and Technology
Hyderabad

ABSTRACT

This paper deals with the study of Data Mining techniques, tasks and effective tools. Data Mining refers to the retrieving the knowledge from the large repository of data through data analysis. Data Mining is an important technique to provide a better solution. It is a powerful technology with higher potential use of the available data for competitive advantages. Data mining have been successful in many areas like, healthcare, Banking and finance, telecommunication, risk analysis management, fraud detection management etc. This paper describes major mining tasks, techniques, tools and some applications of data mining with real time examples.

Keywords:
Data Mining, KDD, Classification, Clustering, Association, Predictive, Descriptive

1. INTRODUCTION

Traditional techniques may be unsuitable due to enormity of data, high dimensionality of data, heterogeneous, distributed nature of data and much of the data is never analysed at all therefore data mining is needed(PN Tan et al 2005) . There are huge amount of data available in the Information Industry. By using this data we can't get any useful information until it is converted into useful information. So it is essential to analyze this huge amount of data and extract useful data from it. Data mining refers to extracting or "mining" knowledge from large amounts of data (J Han et al, second edition). Data mining is the part of the Knowledge Discovery process. Knowledge discovery in data bases frequently abbreviated as KDD. KDD and Data mining are often used interchangeably because data mining is a key part of the KDD process. KDD process consists of a sequence of the following steps: data selection, data pre-processing, data transformation, data mining interpretation and finding evaluation (J Han et al, second edition). The stages of KDD process is shown in Fig 1 (G. Piatetsky et al, 1996)

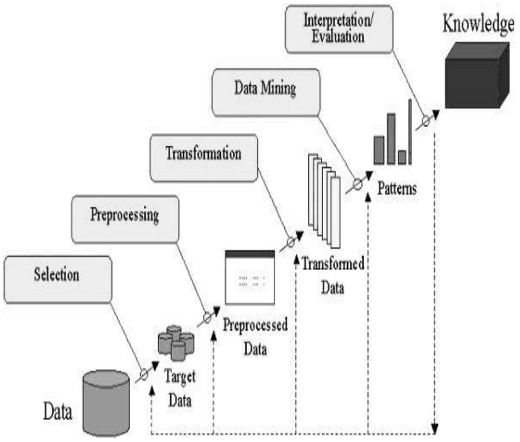

Fig 1. Stages of Knowledge Discovery Process

Data mining is a process used to convert raw data into useful information. By data mining in a large collection of data, the companies can able to understand about their customers and can develop more effective marketing strategies as well as increase sales and decrease costs. Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. The first step involved in the knowledge discovery process is Data selection, where it retrieves the information which is related to search from the database (J Han et al, second edition). Selection process that gathers the important data from which knowledge is to be extracted.The second step is Pre-processing step performs get rid of the noisy data, try to find the missing data or to develop a strategy for handling missing data, detect or remove outliers and resolve inconsistencies among the data. There are a number of different methods and

tools used for pre-processing, including: normalization sampling and feature extraction (Systematic Approach on Data Pre-Processing In Data Mining by S.S.Baskar, et al).The next step is Transformation, the generation of better data for the data mining is prepared and developed. Actually this step transforms the data into

Forms which is suitable for mining by performing Task like aggregation, smoothing, normalization, generalization, and discretization. (Sheenal Patel and Hardik 2016). The final step of the KDD process is interpretation and evaluation of the recover information with respect to the points defined in the first step. Interpretation of mined models to make the user to understand, such as visualization and summarization. Final is acting on the discovered knowledge: using the knowledge directly, organize the knowledge into another system for further process. Once all these processes have been completed, we can able to use this knowledge in many applications such as financial data Analysis, Retail Industry, Telecommunication industry, Fraudulent Detection, Research Analysis etc.

2. DATA MINING TASK

The data mining tasks can be categorized generally into two types based on what a certain task tries to achieve. Those two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data since predictive data mining tasks perform inference on the available information set to anticipate how a new data set will behave. Predictive data mining tasks is a model which is helpful in prognosticates unknown or future values of another data set from the available data. A medical practitioner trying to diagnose a disease depends on the medical test report of a patient can be considered as a predictive data mining task. Descriptive data mining tasks usually identifies data describing patterns and have new and vital information from the available data set as outcome. A retailer trying to identify a items that are purchased together can be considered as a descriptive data mining task. The classification of data mining task is displayed in the following figure 2 (J Han et al, second edition).

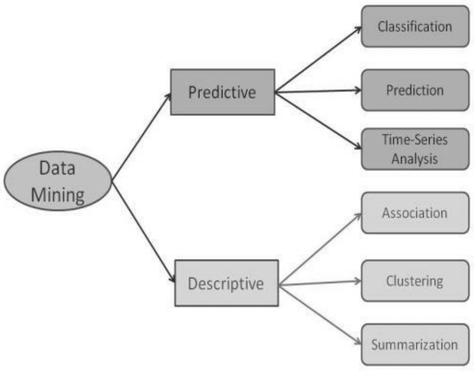

Fig 2. Data Mining Task Types

A. Predictive

Classification derives a model to identify the class of an object based on its properties. A collection of records will be available, each record have a set of attributes. One of the attributes will be a class attribute and assigning a class attribute to new set of records as accurately as possible is the goal of classification task. Classification can be used in direct marketing that is to reduce marketing costs by targeting a set of customers who always like to buy a new product. Using the data in hand, it is possible to know which customers purchased similar items and who did not bought in the past. Hence, {purchase, didn’t purchase} determination forms the class attribute in this case. Once the class attribute is assigned, by statistical analysis and lifestyle details of customers who purchased correspondent products can be collected and can be sent promotion mails to them directly. Prediction task guesses the possible values of missing or upcoming data. Prediction involves implementing a model depends on the available information and this model is used in anticipating future values of a new data set of interest (Divya Tomar, 2013). For example, a model can predict the income of an employee based on experience, qualification and other demographic factors like place of living, language etc. Also prediction analysis is used in various areas including medical marketing analysis, intrusion detection etc. Time series is a sequence of actions where the next event is determined by one or more of the earlier actions. Time series reflects the process being measured and there are confident components that affect the activities of a process. Time series analysis includes methods to examine time-series data in order to extract useful trends, patterns, statistics and rules. Stock market prediction is an important application of time- series analysis (Prakash Mahindrakar et al, 2013, vol-3)

B. Descriptive

Association determines the association or connection among a set of objects. Association identifies the associations among objects. Association analysis is used for product management, marketing, catalogue design, direct marketing etc. A retailer can recognize the products that generally customers purchase together or yet find the consumers who react to the promotion of same kind of goods. If a retailer finds that beer and nappy are bought collectively mostly, he can place nappies on sale to promote the sale of beer. Clustering is used to recognize data objects that are alike to one another. The similarity can be decided based on a number of factors like purchase actions, responsiveness to certain actions, physical locations and so on. For example, an insurance company can cluster its customers based on residence, age, income, education, profession etc.

This group information will be useful to know the customers better and hence provide better modified services (J Han et al, second edition).Summarization is the overview of data. A set of related data is summarized which result in a smaller set that provides collective information of the data. For example, the shopping done by a customer can be summarized into total spending, products, offers availed, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behaviour analysis. Data can be summarized in dissimilar abstraction levels and from different point of view. Different data mining tasks are the heart of data mining process. Different prediction and categorization data mining tasks actually extort the required information from the available data sets.

3. DATA MINING TECHNIQUES

Data mining incorporate approaches and techniques from various disciplines such as statistics, machine learning, artificial intelligence, data warehousing, database management, spatial data analysis, data visualization, probability graph theory etc. In short, data mining is a multi-disciplinary field.

A. *Statistical Approach*

Statistical approach includes a number of methods to analyze arithmetical data in bulky quantities. Different statistical tools used in data mining are cluster analysis, regression analysis, Bayesian network and correlation analysis. Statistical models are usually built from a training data set. Correlation analysis identifies the association 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 edges represent the relationship between the nodes whereas the nodes represent variables.

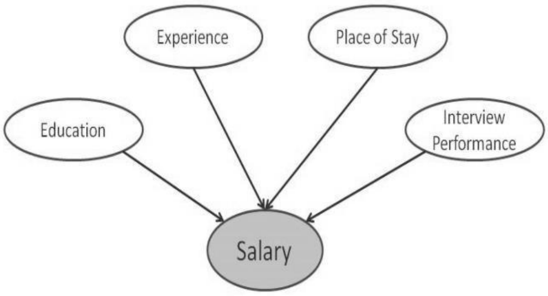

Fig 3. Simple Example of Bayesian network

B. *Machine Learning Approach*

Machine learning is the collection of methods, principles and algorithms that permits learning and guess on the basis of past data. Machine learning is used to build new models and to search for a best model. Machine learning methods usually apply heuristics while searching for the model. Data mining uses a number of machine learning methods including conceptual clustering, inductive concept learning and decision tree induction (J.R. Quinlan, Programs for Machine Learning. San Mateo, 1993). A decision tree is a classification tree that selects the class of an object by following the path from the root to a leaf node. A simple decision tree that is used for weather forecasting is displayed in the following figure.

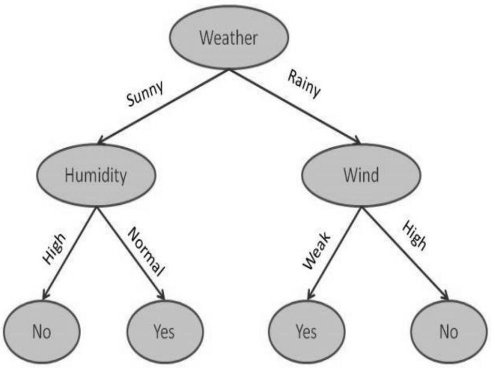

Fig 4. Simple Example of Decision Tree

C. *Neural Networks*

A neural network or artificial neural network is a mathematical function performed on a set of connected nodes called neurons. A neuron is a processing element that computes some requirement of its inputs and the inputs can even be the outputs of other nurodes or neurons. A neural network can be trained to find the association between input attributes and output attribute by adjusting the associations and the parameters of the nodes. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed (Gajendra Sharma, S.K. Kataria & Sons).

D. *Data Visualization*

The information extracted from bulky volumes of data should be presented well to the end user and data visualization techniques make this possible. Visual interpretation of complex relationship in Multidimensional data (Introduction to Data Mining and its Applications by S. Sumathi et al 2006).

Data is changed into different visual objects such as lines, dots, 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.

4. DATA MINING TOOLS

Here are parts of the table with the active data mining tools in (Nen-Fu Huang et al,2005) as License code:
CO - commercial, OS - open source

Tool	Company	License	Remarks
11 Ants	11Ants Analytics	CO	family of data mining tools with a focus on business applications
ADAPA	Zementis Inc.	CO	develops the ADAPA decision engine which is a framework to deploy, integrate, and execute predictive models in PMML, add-ins for Excel, IBM cloud solution (Software as a Service - SaaS)
Coheris SPAD Data Mining	Coheris	CO	company provides also solutions for text mining, former company SPAD
D2K - Data to Knowledge	U. of Illinois	CO/OS	additional tools for EA and text mining, tool I2K for images under development, free academic version, see Alcala09, no developments since 2004
Data Applied	Data Applied	CO	web service for Data Analysis, SAAS
DataDetective	Sentient	CO	with tools for fuzzy matching, applications on CRM, crime analysis, fraud detection
GhostMiner	FQS Poland / Fujitsu	CO	multi model support
IBM SPSS Modeler	IBM	CO	former Clementine, now in cooperation with IBM, Predictive Analytics Software (PASW), SPSS is an IBM company since 2009
InfiniteInsight	KXEN	CO	Providing predictive software tools to application providers and system integrators
JMP	SAS Institute	CO	free trial, additional special tools for genomics
KnowledgeStudio	ANGOSS Software	CO	PMML support and code generation
Model Builder	FICO	CO	company's former name Fair Isaac Corporation
Oracle Data Mining (ODM)	Oracle	CO	provides GUI, PL/SQL-interface, and Java-interface to Attribute Importance, Bayes Classification, Association Rules, Clustering, SVM
Partek Discovery Suite	Partek Incorporated	CO	additional special solutions for genomics, free demos
PolyAnalyst	Megaputer	CO	from Goebel99, support for text mining
Predixion Enterprise Insight	Predixion Software	CO	data mining suite with a focus to standard worksflows, big data support, cloud options, OEM options possible

TABLE I: Active tools used in Data mining

5. DATA MINING APPLICATIONS

Data mining techniques have been applied successfully in many areas from business to science and sports. Data mining has been used in database marketing, retail data analysis, stock selection, credit approval, etc. Data mining techniques have been used in astronomy, molecular biology, medicine, geology, and many more fields. It has also been used in health care management, taxfraud detection, money laundering monitoring, and even sports. The following table shows some basic applications of data mining and most popular data mining functionalities (Feng et al, 2015).

Application	Classification	Clustering	Association analysis	Time series analysis
e-commerce		✓	✓	
Industry	✓	✓	✓	
Health care		✓	✓	
City governance	✓	✓	✓	✓

Table 2: The data mining application and most popular data mining functionalities.

6. CONCLUDING REMARKS

In this paper presented the detail information about data mining tasks, tools, techniques and applications for the efficient mining of data, and easily understand the real time examples in data mining for the better understanding to develop the future applications in the modern trends of computer engineering fields, the concept of the data mining is explained in detailed. This paper provided the comparative analysis of different data mining techniques. In this, easily observed data mining tool which comes under commercial and open source, this provides some common applications of data mining with data mining techniques which have been used. Here discussed the summary of data mining techniques, tools, tasks and applications; it is used to develop the future Engineers and researchers in the modern world.

REFERNCE

1) S.S. Baskar, et al, (2013) An international journal of advanced computer technology, Systematic Approach on Data Pre-Processing in Data Mining, Volume-2, no.9, PP.-335-339,

2) Jiawei Han and Micheline Kamber(2005) “Data Mining: Concepts and Techniques”, Second Edition by , MK Publisher, pp-5-50

3) Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare”, International Journal of Bio-Sci-ence and Bio-Technology Vol.5, No.5 (2013), pp. 241-266.

4) G.Piatetsky-shapiro, U.Fayyed and P.Smith. (1996)“Data mining to Knowledge discovery: An overview. Advances in knowledge Discovery and Data Mining”, MIT Press, pp. 1-35

5) Gajendra Sharma, “Data mining and Data Warehousing and OLAP”, Published by S.K. Kataria & Sons, New Delhi, India.

6) Hindawi Publishing Corporation International Journal of Distributed Sensor networksVolume 2015, Article ID 431047, 14 pages-Data Mining for the Internet of Things: Literature Review and Challenges by Feng et al

7) International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016, Sheenal Patel and Hardik, PP 53-60

8) Introduction to Data Mining and its Applications by S. Sumathi, S.N. Sivanandam,PP-27-30

9) J.R. Quinlan, Programs for Machine Learning. San Mateo, USA: organ Kaufmann, 1993.

10) Nen-Fu Huang, Chia-Nan Ka, Hsien-Wei Hun, Gin-Yuan Jai and Chia-Lin Lin,” Apply Data Mining to Defense-in-DepthNetworkSecuritySystem”. Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA’05), 2005.

11) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar “*Introduction to Data Mining”,* Addison-Wesley (2005). PP. 4-6

12) Prakash Mahindrakar & Dr. M. Hanumanthappa, “Data Mining In Healthcare: A Survey Of Techniques And Algorithms With Its Limitations And Challenges”, Int. Journal of Engineering Research and Applications:, Volume: 03 no.06 pp.937-941, , Nov-Dec 2013.