ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

Data Mining: Task, Tools, Techniques and Applications

S.D.Gheware[1], A.S.Kejkar[2], S.M.Tondare[3]

Lecturer, Computer Technology, V.A.P.M.Almala , Latur, India[1]

Lecturer, Electronics and Telecommunication, V.A.P.M.Almala,Latur, India [2]

Assistant Professor, Electronics and Telecommunication, Sandipani Technical Campus F.E., Latur, India[3]

Abstract: This paper deals with detail study of Data Mining its techniques, tasks and related Tools. Data Mining refers to the mining or discovery of new information in terms of interesting patterns, the combination or rules from vast amount of data. It helps in classifying, segmenting data and in hypothesis formation. With such a vast amount of data, there is need for powerful technique for better interpretation of these data. Including commercial and open source, many program available to perform data mining. Data mining tools predict future trends and behaviours, allowing business to make proactive and present knowledge in the form which is easily understood to human.

Keywords: Data mining, KDD, Clustering, Tools, Regression.

I. INTRODUCTION

Traditional techniques may be unsuitable due to enormity of data, high dimensionality of data, heterogeneous, distributed nature of data [1] and much of the data is never analysed at all therefore data mining is needed. (Beerry and Linoff , 2000) stated that data mining is a process of analysis and exploration by means of automatic or semi automatic to discover the meaning patterns or rules. Data mining is the part of the Knowledge Discovery process [2]. Knowledge discovery in data bases frequently abbreviated as KDD. Data mining and KDD are often used interchangeably because data mining is the key part of the KDD process [3]. KDD process may consists several steps: like data selection, data cleaning, data transformation, pattern searching i.e. data mining, finding presentation, finding interpretation and finding evaluation. A typical KDD process is shown in Figure.1[4].

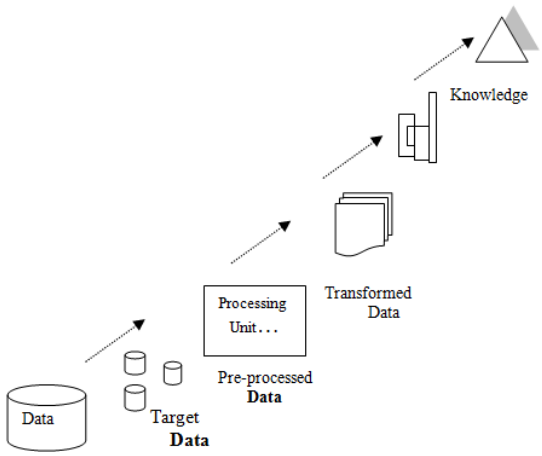

Figure 1: A typical Knowledge Discovery process

II. DATA MINING TASK

A. Summarization

Summarization is the generalization or abstraction of data. A set of relevant data is abstracted and summarized, resulting a smaller set which gives a general overview of data. For example, the long distance calls of customer can be summarized in to total minutes, total calls, total spending etc instead of detailed calls. Similarly the calls can be summarized in to local calls, STD calls, ISD calls etc.

B. Clustering

Clustering is identifying similar groups from unstructured data. Clustering is the task of grouping a set of objects in a such a way that object in same group are more similar to each other than to those in other groups. Once the clusters are decided, the objects are labelled their corresponding clusters, and common features of the objects in cluster are summarized to form a class description. For example, a bank may cluster its customer in to several groups based on the similarities of their income, age, sex, residence etc, and the command characteristics of the customers in a group can be used to describe that group of customers. This will the bank to understand its customers better and thus provide customized services.

C. Classification

Classification is learning rules that can be applied to new data and will typically include following steps: pre-processing of data, designing modelling, learning/feature selection and validation /evaluation. Classification predicts categorical continuous valued functions. For example, we can make classification model to categorize bank loan application as either safe or risky. Classification is the derivation of model which determines the class of an object based on its attributes. A set of object is given as training set in which every object is represented by vector of attributes along with its class. By analysing the relationship between attributes and class of the objects in the training set, classification model can be constructed. Such classification model can be used to classify future objects and develop a better understanding of the classes of the objects in the data base. For example, from the set

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

of loan borrowers (Name, Age, and Income) who serve as training set, a classification model can be built, which concludes bank loan application as either safe or risky. (If age = Youth then Loan decision = risky).

D. Regression

Regression is finding function with minimal error to model data. It is statistical methodology that is most often used for numeric prediction. Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so cautions advisable [5] for example, correlation does not imply causation.

E. Association

Association is looking for relationship between variables or objects. It aims to extract interesting association, correlations or casual structures among the objects i.e. the appearance of another set of objects in [3]. The association rules can be useful for marketing, commodity management, advertising etc. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness[6] and based on the concept of strong rules presented in [7], introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {Onions, potatoes} {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production, and bioinformatics.

III. DATA MINING TECHNIQUES

Data mining adopt its technique from many research areas, including statics machine learning, database systems, rough sets, visualization and neural networks.

A. *Statistical Approach*

Statistical models are built from a set of training data. Many statistical tools have been used for data mining including, Bayesian network, correlation analysis, regression analysis and cluster analysis. For example simple Bayesian network for traffic jam problem is given in figure 2. In the Bayesian network nodes represents states or variable while edges represents dependencies between nodes. From figure we can see that rush hour, bad weather or accident affect the traffic which in turn causes traffic jam.

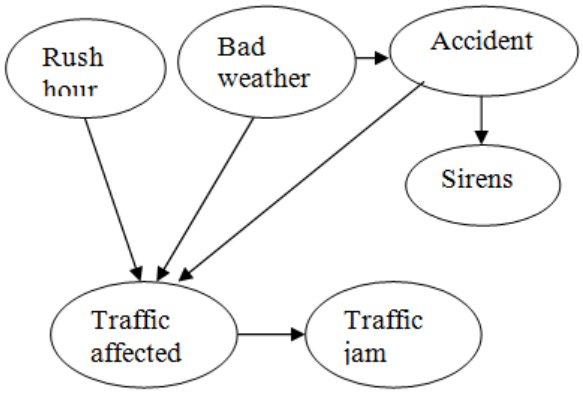

Figure 2. Example of an unacceptable low-resolution image

B. Machine Learning Approach

The most common machine learning methods used for data mining include conceptual learning, inductive concept learning and decision tree induction. By following the path from root to leaf node an objects class can be determine by decision tree. Decision trees are induced from the training set and decision trees give classification rules. A simple decision tree is given in Figure.3 [8], it determines the car's mileage from its size, transmission type and weight. The leaf nodes are in square boxes.

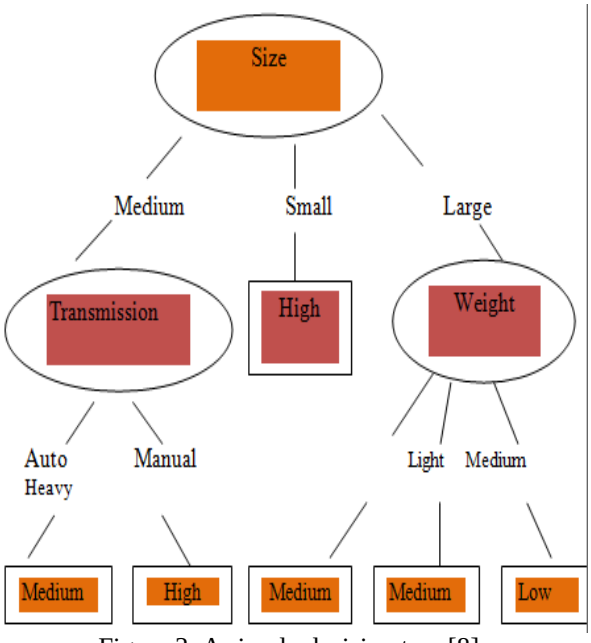

Figure 3. A simple decision tree [8]

From decision tree we can conclude, for example, large size; heavy weight car will have low mileage. Nodes represent three classes of mileage.

IV. DATA MINING TOOLS

Here are parts of the table with the active tools in [9] as License code: CO - commercial, OS - open source

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

TABLE I: Active tools used in Data mining

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.
Coheris SPAD	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
Revolution R Enterprise	Revolution Analytics	OS/CO	based on open source software R with many additional tools for big data (e.g. Hadoop support) and database coupling, some commercial parts also free for academic use
Salford Predictive Modeling Suite (SPM)	Salford Systems	CO	includes former separate tools CART, MARS, TreeNet, Random Forests
SAS Enterprise Miner	SAS Institute	CO	one of the world's leading tools, enterprise oriented

V. CHALLENGES

High dimensional spare data, uncertain data, incomplete data, scalability, complex and heterogeneous data, data quality, data ownership and distribution, privacy preservation, streaming data this are the challenges are faced by data mining.

A. High dimensional spare data:

High dimensional spare data significantly deteriorate the reliability of the models derived from the data [10] common approaches are to employ dimension reduction or future selection [11] to reduce data dimension or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

B Uncertain Data

Uncertain data are special type of data reality where each data field is subjected to some random/error distribution. For example, each recording location of GPS system is represented by mean value and variances to indicated expected errors; for uncertain data major challenge is that each data item is represented as sample distributions but not as single value, so most of the existing data mining algorithms cannot be directly applied. Error aware data mining utilizes the mean and variance values with respect to each single data item to build Naive Bayes model for classification. Similar approaches have also been applied for decision tree or database queries.

C Incomplete Data

Incomplete data [10] refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values. For example, dropping some sensor node readings to save power for transmission. While most modern data mining algorithms have in-built solutions to handle missing values, data imputation is an established research field that seeks to impute missing values to produce improved models.

VI. DATA MINING APPLICATIONS

Various fields uses data mining technologies because of fast access of data and valuable information from vast amount of data. Data mining technologies have been applied successfully in many areas like marketing, telecommunication, fraud detection, and finance, medical and so on. Some of the application is listed below.

A. Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases: Design and construction of data warehouses for multidimensional data analysis and data mining. Loan payment prediction and customer credit policy analysis. Classification and clustering of customers for targeted marketing. Detection of money laundering and other financial crimes.

B. Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web. The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

C. Telecommunication Industry

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, e-mail, web data transmission etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services as

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

D. Biological Data Analysis

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.

E. Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modelling, chemical engineering, fluid dynamics etc. Following are the applications of data mining in field of Scientific Applications.

VII. CONCLUSION

In this paper, we have discussed detail study of data mining with various studies like tasks, tools and techniques. The implementation of data mining techniques will allow users to retrieve meaningful information from virtually integrated data. These techniques provide variety of applications for industries like retail, telecommunication, Bio-medical etc. These tools predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human.

REFERNCES

[1] Tan, Steinbach, and Kumar, "Introduction to Data Mining".2004.
[2] D.W. Cheung, S.D.Lee and B.Kao, "A general incremental technique for maintaining discovered association rule". Proc. In fifth international conference on data base system for advanced applications, Australia, 1997.
[3] Y.Fu , Data Minig : Tasks, Techniques and Applications.
[4] G.Piatetsky-shapiro, U.Fayyed and P.Smith. From data mining to Knowledge discovery: An overview. Advances in knowledge Discovery and Data Mining, pages 1-35, MIT Press, 1996.
[5] R.Kaur, S.Kaur, A.Kaur, R.Kaur, A.Kaur, "An Overview of Database management System, Data warehousing and Data Mining". IJARCCE, Vol.2, issue.7, July 2013.
[6] K. Maheshwar and D.Singh, "A Review of Data Mining based instruction detection techniques". International Journal of Application or Innovation in Engineering & Management.Vol.2, Issue.2, Feb.2013.
[7] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process". International Journal of Soft Computing and Engineering,Vol.2,Issue.3, July.2012.
[8] M.S.Chen,J.Han,and P.S.Yu. Data Mining: An Overiew from a database Prespective. IEEE Transactions on Knowledge and Data Engineeirng, Vol.8, pp.866-883,1996.
[9] Nen-Fu Huang, Chia-Nan Ka, Hsien-Wei Hun, Gin-Yuan Jai and Chia-Lin Lin," Apply Data Mining to Defense-in-Depth Network Security System". Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05), 2005.
[10] X.Wu, X.Zhu, Gong-Qing.Wu and W.Ding , " Data mining with big data". IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, Jan. 2014.
[11] H.Wang, G.Nie and K.Fu , "Distributed data mining based on semantic web and grid". IEEE International Conference on Computational Intelligence and Natural Computing, 2009.

ACKNOWLEDGMENT

The authors would like to thanks the Department of Computer Technology, Electronics and Telecommunication, V.A.P.M. Almala and S.T.C.F.E. Latur, India for their guidance and cooperation.