International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

A SURVEY ON USE OF DATA MINING METHODS TECHNIQUES AND APPLICATION

K.Prema[1],A.Kumar Kombaiya[2]

[1]Research scholar, Chikkanna Government Arts College,(India)

[2]Assistant Professor, Chikkanna Government Arts College,(India)

ABSTRACT

Data mining is helpful in acquiring knowledge from large domains of databases, data warehouses and data marts. The process of data mining to find useful patterns from large volume of data. Techniques used in data mining helps in classifying, segmenting data with open source and in hypothesis formation. There is need for powerful technique for better interpretation with large volume of data with commercial and open source, to perform data mining. Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. This is an important and evolving research area and used by the biologists to statisticians and computer scientists as well.. Knowledge /information are conveying the message through direct or indirect. This paper provides a survey of various data mining techniques and applications areas include association, correlation, clustering and neural network. This paper discusses the topic based on past research paper and also studies the data mining techniques.

Keywords: Data mining, knowledge discovery database, Data mining Techniques, clustering, Data mining applications.

I.INTRODUCTION

Data mining is a logical process that is used to search through large amount of data in order to find useful data. Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as ‘Knowledge mining from data’ or “Knowledge mining”. Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining..Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns .

II.METHODOLOGIES OF DATA MINING

2.1 Neural Network

Neural network is a set of connected input/output units and each connection has a weight present with it.During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications . This powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in fig.1. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non linear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the non linear characteristics of ANN provide it lots of flexibility to achieve input output map.

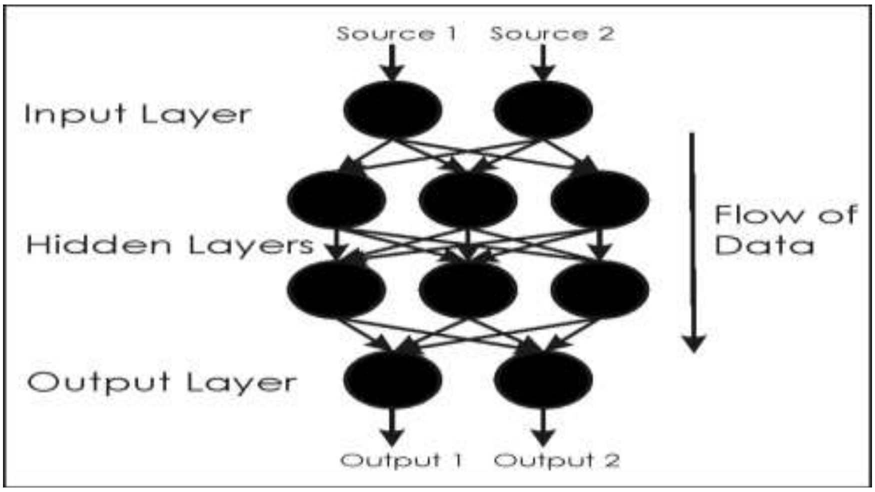

Fig: 1 Neural Network with hidden layers

2.2Decision Trees

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distribution. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached .Decision tree is represented in figure 2.

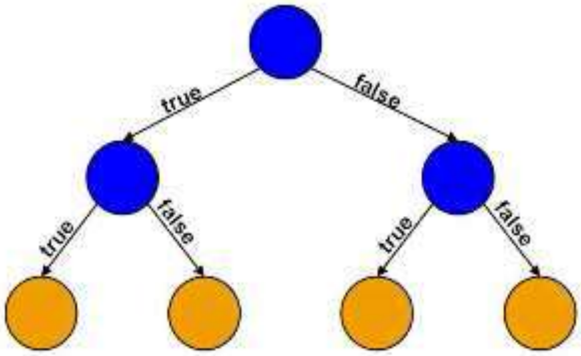

Fig 2 Decision tree

Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. The author defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability.

2.3 Genetic Algorithm

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions (schemata theory), just like nature does by combining the DNA of living beings. Genetic Algorithm is basically used as a problem solving strategy in order to provide with a optimal solution. They are the best way to solve the problem for which little is known. They will work well in any search space because they form a very general algorithm. The only thing to be known is what the particular situation is where the solution performs very well, and a genetic algorithm will generate a high quality solution. Genetic algorithms (GAs) are based on a biological applications; it depends on theory of evolution.

2.4 Rule Extraction

The taxonomy of Rule extraction contains three main criteria for evaluation of algorithms: the scope of dependency on the black box and the format of the extract description. The first dimension concerns with the scope of use of an algorithm either regression or dimension. The second dimension focuses on the extraction algorithm on the underlying black-box with independent algorithms. The third criterion focuses on the obtained rules that might be worthwhile algorithms. Besides this taxonomy the evaluation criteria appears in almost all of these surveys rule; Scalability of the algorithm &consistency.Generally a rule consists of two values. A left and a right hand consequent. An antecedent can have one or multiple conditions which must be true in order for the consequent to be true for a given accuracy whereas a consequent is just a single condition. In a database antecedent, consequent, accuracy, and coverage are all targeted. Sometimes "interestingness" is also targeted used for ranking. The situation occurs when rules have high coverage and accuracy but deviate from standards. It is also essential to note that even though patterns are produced from rule induction system, they all not necessarily mean that a left hand side ("if "part) should cause the right hand side ("then") part to happen. Once rules are created and interestingness is checked they can be business where each rule performs a prediction keeping a consequent as the target and the accuracy of the rule as the accuracy of the prediction which gives an opportunity for the overall system to improve and perform well. For data mining domain, the lack of explanation facilities seems to be a serious drawback as it produce opaque model, along with that accuracy is also required. Experience from the field of expert systems has shown that an explanation capability is a vital function provided by symbolic AI systems. In particular, the ability to generate even limited explanations is absolutely crucial for user acceptance of such systems. Since the purpose of most data mining systems is to support decision making, the need for explanation facilities in these systems is apparent.

III.DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases.

3.1. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a
classifier.

3.2. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

3.3. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

3.4. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

3.5. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it.During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

IV.DATA MINING APPLICATIONS

Data mining is a relatively new technology that has not fully matured. Despite this, there 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 and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found using data mining technology.

4.1. FBTO Dutch Insurance Company

Challenges

1 To reduce direct mail costs.

2 Increase efficiency of marketing campaigns.

3 Increase cross-selling to existing customers, using inbound channels such as the company's sell center and the internet a one year test of the solution's effectiveness.

Results

1. Provided the marketing team with the ability to predict the effectiveness of its campaigns.
2. Increased the efficiency of marketing campaign creation, optimization, and execution.
3. Decreased mailing costs by 35 percent.
4. Increased conversion rates by 40 percent.

4.2. ECtel Ltd., Israel

Challenges

1. Fraudulent activity in telecommunication services.

Results

1. Significantly reduced telecommunications fraud for more than 150 telecommunication companies worldwide.

2 Saved money by enabling real-time fraud detection.

4.3. Provident Financial's Home credit Division, United Kingdom

Challenges

1. No system to detect and prevent fraud.

Results

1 Reduced frequency and magnitude of agent and customer fraud.

2.Saved money through early fraud detection.

3. Saved investigator's time and increased prosecution rate.

4.4. Standard Life Mutual Financial Services Companies

Challenges

1. Identify the key attributes of clients attracted to their mortgage offer.
2. Cross sell Standard Life Bank products to the clients of other Standard Life companies.
3. Develop a remortgage model which could be deployed on the group Web site to examine the profitability of the mortgage business being accepted by Standard Life Bank.

Results

1. Built a propensity model for the Standard Life Bank mortgage offer identifying key customer types that can be applied across the whole group prospect pool.
2. Discovered the key drivers for purchasing a remortgage product.
3. Achieved, with the model, a nine times greater response than that achieved by the control group.

V.CONCLUSIONS

At present data mining is a new and important area of research and ANN itself is a very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies. Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

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.] 4.Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, 2007.

[5.] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: MonitoringMeasurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[6.] Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao,"A Visual Data Mining Framework for Convenient Identification of Useful Knowledge", ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1,pp.- 530- 537,Dec 2005.

[7.] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems,vol.- 8,no.-6, pp.-378-389,1995.

[8.] Lior Rokach and Oded Maimon,"Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)", ISBN:981-2771-719, World Scientific Publishing Company, 2008.

[9.] 9.Venkatadri.M and Lokanatha C. Reddy ,"A comparative study on decision tree classification algorithm in data mining" , International Journal Of Computer Applications In Engineering ,Technology And Sciences (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.

[10.] 10. AnkitaAgarwal,"Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[11.] 11. Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, "The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation", Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications,pp- 593-604,sept 2005.

[12.] 12. Fu Xiuju and Lipo Wang "Rule Extraction from an RBF Classifier Based on Class-Dependent Features ",ISNN'05 Proceedings of the Second international conference on Advances in Neural Networks ,vol.-1,pp.-682-687,2005.

[13.] 13. H. Johan, B. Bart and V. Jan, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models".In Open Access publication from Katholieke Universiteit Leuven, pp.1-56, 2006

[14.] 14 Tan, Steinbach, and Kumar, "Introduction to Data Mining".2004

[15.] 15 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.

[16.] 16 Er. Rimmy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012