International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 10 No. 1 Oct. 2014, pp. 1-5

http://www.ijisr.issr-journals.org/

A Survey on Data Mining Techniques

M. Suganthi

School of Computer Science and Engineering, Bharathidasan University, Trichy, India

ABSTRACT: This paper provides a survey of various data mining techniques. These techniques include Association rule, Fuzzy logic, Decision tree and neural network. The concept of data mining was summarized and its significance towards its methodologies was illustrated. This paper also conducts a formal review of the area of rule extraction from Association rule and Fuzzy Logic. This survey paper also conducts a formal review of the applications of data mining such as the education sector, marketing, fraud detection, manufacturing and telecommunication. This paper discusses the topic based on past survey paper and also studies the data mining techniques.

KEYWORDS: Data Mining, Neural Network, Decision Tree, Association Rule and Fuzzy Logic.

1 INTRODUCTION

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 an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. 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 [2]. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns. 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 functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in two categories-descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. A **descriptivemodel** presents, in concise form, the main characteristics of the data set. It is essentially a summary of the data points, making it possible to study important aspects of the data set. Typically, a descriptive model is found through undirected data mining; i.e. a bottom-up approach where the data "speaks for itself". Undirected data mining finds patterns in the data set but leaves the interpretation of the patterns to the data miner. The purpose of a **predictive model** is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable. If the target value is one of a predefined number of discrete (class) labels, the data mining task is called classification. If the target variable is a real number, the task is regression. The predictive model is thus created from given known values of variables, possibly including previous values of

Corresponding Author: M. Suganthi

the target variable. The training data consists of pairs of measurements, each consisting of an input vector x (i) with a corresponding target value y(i). The predictive model is an estimation of the function y=f(x; q) able to predict a value y, given an input vector of measured values x and a set of estimated parameters q for the model f. The process of finding the best q values is the core of the data mining technique [3].

At the core of the data mining process is the use of a data mining technique. Some data mining techniques directly obtain the information by performing a descriptive partitioning of the data. More often, however, data mining techniques utilize stored data in order to build predictive models. From a general perspective, there is strong agreement among both researchers and executives about the criteria that all data mining techniques must meet. Most importantly, the techniques should have high performance. This criterion is, for predictive modeling, understood to mean that the technique should produce models that will generalize well, i.e. models having high accuracy when performing predictions based on novel data.

Classification and prediction are two forms of data analysis that can be used to extract models describing the important data classes or to predict the future data trends. Such analysis can help to provide us with a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels, prediction model, and continuous valued function.

2 METHODOLOGIES OF DATA MINING

2.1 ARTIFICIAL NEURAL NETWORK

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 [4]. This powerful predictive modeling 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, nonlinearsystem 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 nonlinear characteristics of ANN provide it lots of flexibility to achieve input output map. Artificial Neural Networks, provide user the capabilities to select the network topology, performance parameter, learning rule and stopping criteria.

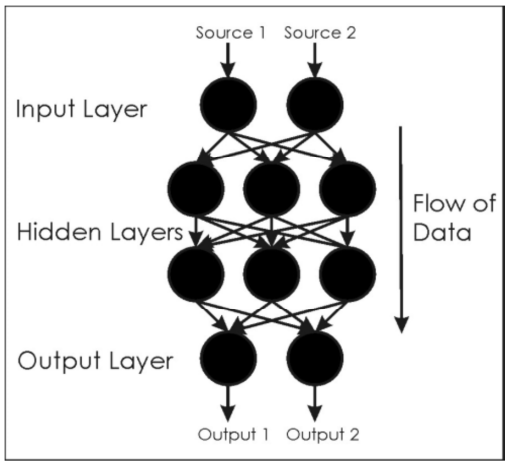

Fig.1. Neural Network with hidden layers

2.2 DECISION 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 [5].Decision tree is represented in figure 2.

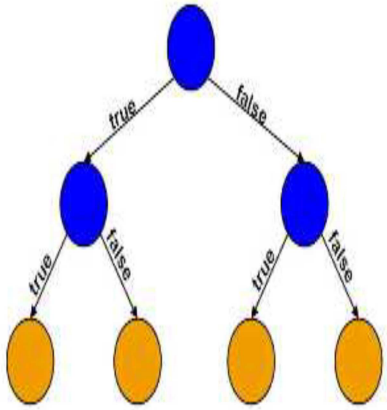

Fig.2. Decision tree

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 [6]. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus marketing managers make use of segmentation of customers, products and sales region for predictive study. These predictive segments derived from the decision tree also come with a description of the characteristics that define the predictive segment. Because of their tree structure and skill to easily generate rules the method is a favored technique for building understandable models.

2.3 ASSOCIATION RULE

Association rule discovery from large databases is one of the tedious tasks in data mining. Association rule mining has a wide range of applicability such as market basket analysis, suspicious e-mail detection, library management and many areas. The conventional algorithm of association rules discovery proceeds in two steps. All frequent item sets are found in the first step. The frequent item set is the item set that is included in at least minimum support transactions. The association rules with the confidence at least minimum confident are generated in the second step. Apriori algorithm uses transaction data set and uses a user interested support and confidence value and produce the association rule set. These association rule sets are discrete and continue therefore weak rule set are required to prune [7].

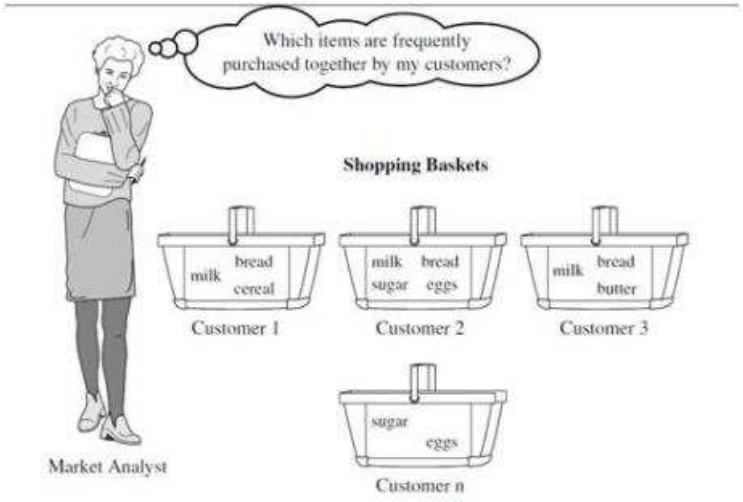

Fig.3. Market Based Analysis

A broadly-used example of association rule mining is market basket analysis. In Market Basket Databases consist of a large no. of records and in each record all items bought by a customer on a single purchase transaction are listed. Managers would be paying attention to know that which groups of items are constantly purchased together. This data is used by them to adjust store layouts (placing items optimally with respect to each other), to cross-sell, to promotions, to catalog design and to identify customer segments based on buying patterns [5]. For example, suppose a shop database has 200,000 point-of-sale transactions, out of which 4,0000 include both items A and B and 1600 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 1600 transactions (alternatively 0.8% = 1600/200,000) and a confidence of 4% (=1600/4,0000). The probability of a randomly selected transaction from the database will contain all items in the antecedent and the consequent is known as support, whereas the conditional probability of a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent is known as confidence. Now a day's products are coming with bar codes. A large amount of sales data is produced by the software supporting these barcode based purchasing/ordering system which is typically captured in "baskets". Commercial organizations are interested in discovering "association rules" that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This "market basket analysis" result can be used to suggest combinations of products for special promotions or sales[8]. Market Based Analysis shown in figure 3.

2.4 FUZZY LOGIC

Fuzzy set theory is an extension of conventional set theory that deals with the concept of partial truth. Fuzzy logic aims to model the vagueness and ambiguity in complex systems. In many image processing applications, expert knowledge must be used for applications such as object recognition and scene analysis. Fuzzy set theory and fuzzy logic provide powerful tools to represent and process human knowledge in the form of fuzzy IF-THEN rules[9].

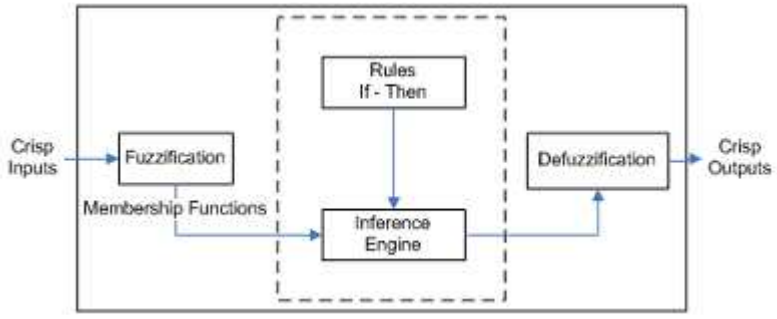

Fig 4 Fuzzy Logic- IF-Then

Over the past few decades, fuzzy logic has been used in a wide range of problem domains. The areas of applications are very wide: process control, management and decision making, operations research, economies and pattern recognition and classification. In the lack of precise mathematical model which will describe behavior of the system, Fuzzy Logic is a good "weapon" to solve the problem: it allows using logic if-then rules to describe the system's behavior [10].

Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form. IF condition THEN conclusion. Example is R1, R1: IF age=youth AND student = yes THEN buys computer=yes. The "IF" part of a rule is known as the rule antecedent or precondition. The "THEN"part is the rule consequent. In this rule antecedent, the condition consists of one or more attribute tests (such as age=youth, and student=yes) that are logically AND. The rule's consequent contains a class prediction (we predicting whether a customer will buy a computer). R1 can also be written as R1: (age=youth)^(student=yes)=>(buy computer=yes). If the condition in a rule antecedent holds true for a given tuple, we says that the rule antecedent is satisfied and that the rule covers the tuple.

A rule R can be assessed by its coverage and accuracy. Given a tuple, X, from a class labeled data set, D, let n_{covers} be the number of tuple covered by R; $n_{correct}$ be the number of tuple correctly classified by R; and |D| be the number of tuple in D. we can define the coverage and accuracy of R as

$$\text{Coverage}(R) = n_{covers/} |D|$$
$$\text{Accuracy}(R) = n_{correct}/ n_{covers}$$

That is, a rule's coverage is the percentage of tuples that are covered by the rule. For a rule's accuracy, we look at the tuple that it cover and see what percentage of them the rule can correctly classify. The Fuzzy logic IF-THEN process shown in figure 4.

3 CONCLUSION

If the conception of computer algorithms being based on the evolutionary of the organism is surprising, the extensiveness with which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new and important area of research. In this survey paper shows that association mining and fuzzy logic gives a accurate solution then other two techniques.

REFERENCES

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

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

[3] Zhao, Kaidi and Liu, Bing, Tirpark, M Thomas. 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.

[4] 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.

[5] 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.

[6] M. Venkatadri 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.

[7] K. Saravana kumar, R. Manicka chezia, "A Survey on Association rule mining using Apriori Algorithm", International Journal of Computer Application(IJCA), vol-45 No:5.May 2005.

[8] Nikita jain, Vishal Srivastava, "Data Mining Techniques: A Survey paper", IJRET ISSN:2319-1163, vol-02 Issue:11, Nov 2013.

[9] S. Sanjeev Sannakki, S. Vijay Rajpurohit, S. Arunkumar, "A Survey on Application of Fuzzy logic in Agriculture",Journal of Computer Applications (JCA) ISSN: 0974-1925, Volume IV, Issue 1, 2011.

[10] Jorge Ropero, Carlos León, Alejandro Carrasco, Ariel Gómez, Octavio Rivera," Fuzzy Logic Applications for Knowledge Discovery: a Survey", International Journal of Advancements in Computing Technology Volume 3, Number 6, July 2011.