ISSN: 2321-7782 (Online)
Volume 3, Issue 9, September 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Comparative Study of Popular Classification Techniques of Data Mining

Deepti Dalal
Worked as an Assistant Professor in Computer Science
KIIT, Gurgaon
Rohtak, India

Abstract: ***Data mining is a term related to the extraction of the unknown or hidden information which is previously unknown from the huge database. In data mining process various patterns are extracted and this is why it is also known as pattern discover. There are a number of data mining techniques including classification, clustering, decision trees, and association rules. In this paper a number of classification techniques are included which are used in diverse areas for data mining purpose and proved to be very helpful in the decision making process in their business. This paper is very useful to select a data mining technique for a specific application.***

Keywords: Neural networks, Multilayer perception, Naive bayes, Clustering, Predictive model, Neurons.

I. INTRODUCTION

Data is collected in the databases almost in every area and it is a difficult task to handle this huge data and extract useful information out of it. Manually extracting important and useful information in the form of knowledge required in the decision making process is not so easy. There is a need of automatic extraction of useful information from this huge data. Data mining is a very useful process and helps in the decision making process by finding the unknown patterns and future trends and there are a number of data mining techniques which can be used for the knowledge data discovery. It is one of the difficult tasks to choose one technique from a large number of available techniques which suits our requirements best because one technique may be suitable for the one process and the other technique for some other applications. Data mining process consist of a number of steps like storage of data in the data warehouse, management of the data in multidimensional database system, accessing the data, using appropriate application software analysis of the data to gather the useful information from large amount of data and finally presenting the extracted information in a useful way like graph or in the form of table. According to Jain and Srivastava, 2013 there are basically two types of tasks performed in the data mining, these tasks are classified as: Descriptive mining tasks and the predictive mining tasks. In descriptive mining task, the main characteristics of the data in the database is described and a bottom-up approach is used in this process. It is basically a process to concise the whole data into a meaningful information. Descriptive mining is considered to be an undirected approach in which the patterns are founded but the task of performing the predictions from the patterns is not done. The user does the predictions from the patterns. In predictive mining, predictions are made from the data to gather the unknown values and the future trends. The unknown values and the predictions are made from some specific variables and these variables are known as target variables. If the target variables used in the prediction process are from some discrete class then the data mining is of classification type and if the target variables are from real numbers then the data mining is regression. There are a number of data mining applications which are using the data mining techniques for their business like in the field of health care, agriculture, sports, education, web mining, detecting terror related activities, weather forecasting, fraud detection and many more.

There are a number of data mining techniques including classification, clustering, decision trees, and association rules. In this paper a number of classification techniques are included which are used in diverse areas for data mining purpose and

proved to be very helpful in the decision making process in their business. This paper is very useful to select a data mining technique for a specific application. The selection of a data mining techniques is based on the two main parameters; the data structure of the database and the purpose of data mining process.

II. CLASSIFICATION TECHNIQUES

Classification techniques of data mining used for the extraction of the knowledge and the patterns is based on the machine learning. In this technique the data from the databases is classified in predefined groups or set of classes that are used in the data mining process. We can make a set of class or a group according to some criteria decided in accordance to some rules like the grades of students in class are classified according to the marks obtained by the students in different subjects like O, A, B, C, D, F. The grouping of the data items is done by the software and the classification of the data in the database is done for further processing. For example, internet is used for shopping purpose and there are a number of online websites that are using data mining technique to assist their business process. Depending upon the choices made by the customer to buy a particular item the website suggest them a number of related products. The website suggests the customers a number of items depending upon their search history and the various other products that other customers buy along with products. This proved to be very helpful in decision making process for both the customers and the websites. A number of mathematical techniques are used in the classification techniques to assist the process like decision trees, neural networks, linear programming and the statistics.

In classification process outcome is predicted from a given input. For this purpose the algorithm process a training set which consists of an attribute set and the outcome is called prediction attribute. The algorithm progresses by finding the relationships between the input attribute set i.e. the training set. The input is then analysed to produce a prediction and how good an algorithm is depends upon the predictions made by the algorithm. Prediction rules are used for knowledge expression in the form of IF-THEN rules, IF part is known as antecedent which consists of a conjunction of conditions and the THEN part is known as consequent which gives the prediction attributes for an item depending upon the result whether it satisfies the antecedent or not. For example to find whether the student will pass the college examination or not can be find by a predicting rule: IF (Percentage_SE > 60) AND (Percentage_SSE > 60) THEN Response = PASS. There is a probability that the student will pass the examination if he/she secured more than 60 in the secondary and senior secondary examination. In general the prediction rule is not as simple as given in the above example, conjunction part is made up with a number of statements joined by OR and AND.

a) Neural Networks

It is a biological system for detecting the unknown patterns and making the future predictions. Nodes of neural networks are analogous to neurons in the brain. Predicting unknown patterns and predicting future trends by using neural networks is known as artificial neural networks. The processing elements are interconnected with each other and are also known as nodes or neurons. In order to produce the output the nodes in the network work in parallel together. In this network structure there exist a fault tolerance mechanism and the output can be generated even after the failure of some nodes in the network. In neural networks there exists a hidden node between the input and the output nodes. Hidden node is a row of nodes which are invisible to the user and it has no predefined meaning. The prediction is done on the input nodes and the output is in the form of predictions done and is represented by the output node. There is a link between the various nodes that make connections between the nodes. An activation number is assigned to each node and weight is assigned to the link between the nodes in the network. IF-THEN rules are then applied iteratively on the input nodes in order to give the output node i.e. the result node which gives the useful information for future predictions. For clustering neural networks are used and clusters are created by forcing the system to compress the data by creating prototypes. The various application areas where neural network technique is used are: fraud detection in banking, detecting fraudulent use of the credit cards, automated driving. In neural networks a black-box approach is used for the prediction. Artificial neural network is non linear approach and is adaptive in nature. This technique is useful in when the data is big and poorly understandable. Complex relationships are formed between

the input and output to find the patterns and to make the future predictions. During learning phase artificial neural network changes its structure according to the flow of internal and external information in the system and that's why this is known as an adaptive system (Tewary, 2015).

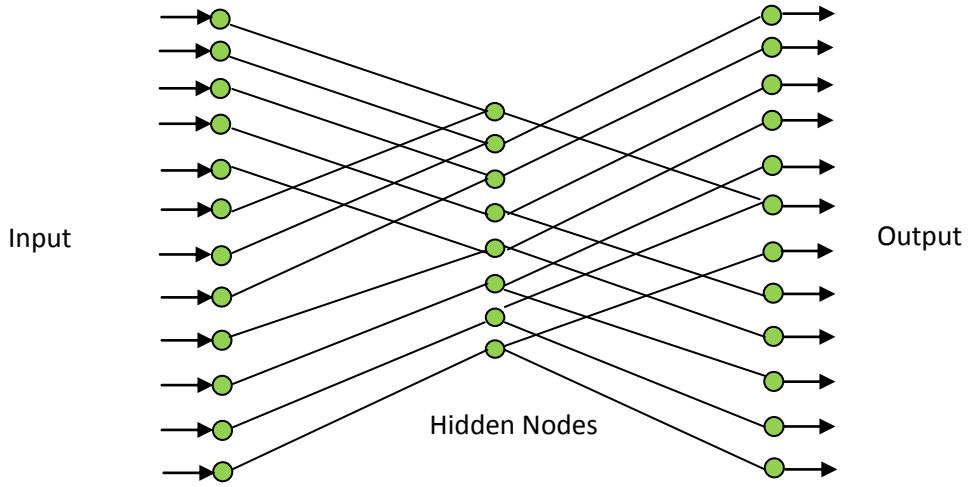

Fig. 1 Layout of a multilayered Artificial Neural Network

One widely used neural network known by Multilayer Perception (MLP) algorithm is given by Witten and Frank in 2000. Input layer in the network is formed by sensory elements set, output layer is of processing elements and the other interior hidden nodes consists of processing elements. This algorithm is most popular of its types and widely used for extracting patterns in various data mining applications. When the relationship between the input and the output attributes are not defined properly then this algorithm is suited best for the approximation of classification function.

b) Decision Trees

Decision tree is used for the prediction of trends and unknown patterns from the database. It is also known by a number of names like classification trees and the regression trees. A decision tree is a classification tree when the outcomes are predicted in terms of a class and if the outcome is predicted in terms of a real number then it is known as regression analysis. Decision tree is a tree like structure where there are parent nodes and the children nodes. The starting node in the tree is an attribute which is required for the prediction. The attributes are classified in terms of groups. The leaves in the decision tree represent the class labels and conjunction of these class labels are represented by the branches. This is used for supervised classification learning. A decision tree is a flow chart kind structure in which every internal node depicts a test on an attribute, where each branch represents an outcome of the test and every leaf node holds a class label. In decision tree each non terminal node represents a decision on the considered data item.

A decision is made by following the Top-Down approach; starting from the root node reach to the terminal node by following the assertions down. Decision trees can also be defined as a special form of a rule set which is characterized by organisation in which rules are in hierarchical form. This is the case when the recursive data partitioning is used in the decision trees. It is a predictive model in which the input space is partitioned into cells and each cell belongs to a particular class. There are a number of applications in which for data mining decision tress is used. One such application of using data mining is given by Agarwal and *et al*., 2012. They used decision trees for educational data mining and the results obtained can be used for detecting the future placements of students, retention rate of the students, performance evaluation in the examination and to find the interest of the students for their career. All the results obtained are collected by analysing the past records of the students by proper mining technique.

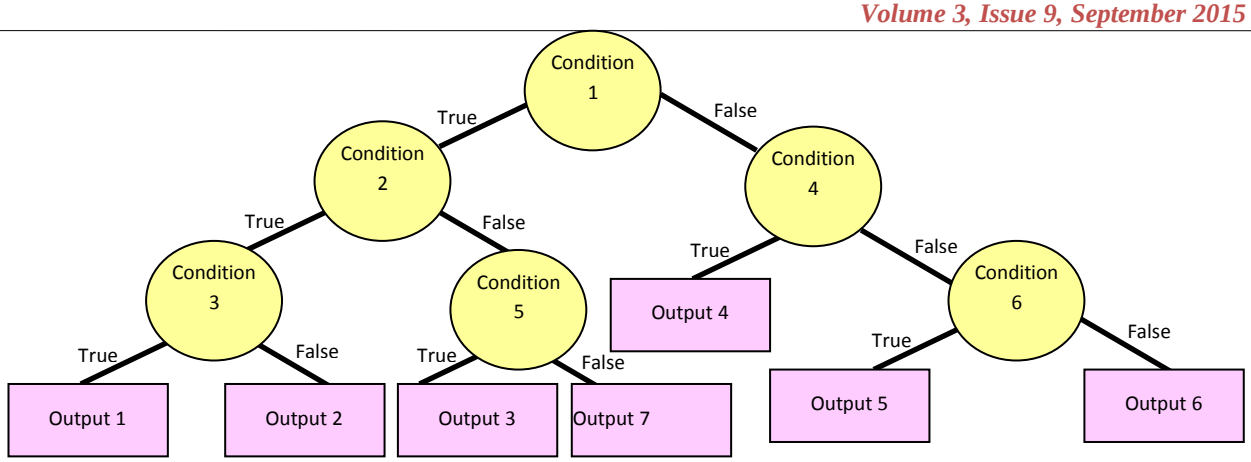

Fig. 2 Structure of a Decision Tree

c) Genetic Programming

Genetic programming is widely used classification technique to solve data mining problem because prediction rules used in this are represented naturally and give good results with global search problems. In genetic programming the prediction rules are represented naturally and give good results with the global search problems. It is a problem solving strategy and is best suited for the problems when there is little information is available and proved to be an optimal solution. The genetic algorithm adds the good feature of other solution in order to give the optimal solution of the problem in hand. In genetic algorithm a general approach is adopted for the problem solving and the algorithm is suited for all search space. This fitness value is used for the mating of the individuals and by combining two or more individuals, a new individual is created and the process of creating a new individual is known as crossover. How many times this process of mating is performed is decided by the user i.e. the number of generations. To avoid the infinite loop of mating of the individuals a parameter hit ratio is used. It is set to the minimum expected value and when a hit ratio greater than this minimum hit ratio is found the algorithm stop. Several solutions are generated for the same problem out of those the optimal one is chosen. Prediction rules are represented naturally in the search space there are a number of local peaks which creates a problem in the local search and perform badly. In genetic programming each and every candidate solution is represented an individual for which a unique fitness value is associated.

d) Genetic Programming

One another technique to deal with the problems relating to the data mining is Bayesian classifier. This technique of extracting patters is a statistical classifier and is unsupervised form of learning. From a long time a number of data mining techniques are used widely in approximately all areas. It is really a tough task to chose one technique from a number of available techniques; generally a hit and trial method is adopted to chose a particular technique to solve a problem. Bayesian classifier is also used for data mining from a long time; it is based on the probability theory. Bayesian classifier technique is unsupervised learning technique because the class variables in the data are not distinguished by the learner from the attribute variables. In this technique probability of class membership that a given tuple belongs to is predicted by statistical classifiers. A network is introduced to describe the probability distribution of the data over the training data set. Heuristic search is used to find the optimal and the best possible solution for a problem over the space of possible network. This technique is also called naive bayes algorithm because of the approach it adopts to solve the problem. The problem simplification is based on the two assumptions first it supposes that to affect the prediction process no hidden attributes are there in the system and secondly it assumes that there exists a conditionally independency for the prognosis attributes. According to this technique there exists an independency of the attributes; the Bayesian algorithm is proved to be very useful data mining technique to find the patterns that are previously unknown.

III. Data Mining Techniques In Application Areas

There are a number of data mining applications which are using the data mining techniques for their business like in the field of health care, agriculture, sports, education, web mining, detecting terror related activities, weather forecasting, fraud detection and many more.

According to the studies of Gupta *et al.*, 2011 done on medical data by using a number of classification techniques, it is found that the accuracy rate of a technique is different for different type of disease. In their study a number of data mining classification techniques are used. A similar study of Abirami *et al.*, 2013 to find the best data mining classification technique on medical dataset for heart disease find the best technique. It is found that classification of dataset is best in the case where the results show minimal error rate and maximal accuracy rate for a particular technique.

Educational data mining is an area where the students' performance is analyzed and the results are helpful for both the teachers and the students. The results obtained after applying the mining techniques can be used to improve the retaining students for higher education, improving teaching learning process, reducing student drop out and their failing rate and many more. Osmanbegović, 2012 studied various data mining classification techniques including decision tree, neural networks, Naïve Bayes classifier for predicting the students performance. According their study Naïve Bayes classifier proved to be a better among all other.

To predict the employees' performance Qasem *et al.*, 2012 studied various classification techniques for finding the factors that contribute to the performance of the employee. It is found that there are a number of factors that affects the performance of an employee to a great extent. According to their study title of the job is considered to be the most effective and several other factors like complexity of job and the responsibilities hold are related to this factor. The various factors are found to be interrelated to each other like performance of an employee and motivation is directly affected by the high responsibilities. There are several other factors like educational degree and grade have less affect the performance of the employee. Several other factors that are considered in this study are age, sex, marital status, gender, experience and kids which can affect the performance of an employee. These factors are very important to consider candidates for the selection and prove to be helpful to avoid hiring of poorly performed employees according to the job profile.

Gibert *et al.*, 2010 findings give a way to find the good and most useful data mining technique from the available. It helps the users to choose a technique which is best suited for their application.

IV. Conclusion

There are a number of classification data mining techniques available each has its own algorithm to find the patterns which are previously hidden and for future prediction. In this paper four classification techniques are used these are artificial neural networks, Genetic algorithms, decision tree and the Bayesian network. There is a lot of research work done to find which algorithm is best suited for a particular application. It is found that depending on the nature and size of the attributes different data mining classification techniques perform different tasks. To select the technique, trial and error methods is best to use for a small dataset. Depending upon the results obtained from different techniques that technique is considered to be the best for which the error rate is minimal and accuracy rate is maximal. Then a set of rules is generated for that particular dataset.

References

1. Agarwal S., Pandey G. N., Tiwari M. D., "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, 2012.
2. Tewary Gaurav, "EFFECTIVE DATA MINING FOR PROPER MINING CLASSIFICATION USING NEURAL NETWORKS", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, 2015.
3. Osmanbegović E., Suljić M., "DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE", Economic Review – Journal of Economics and Business, Vol. X, Issue 1, 2012.
4. Gupta S., Kumar D., Sharma A., "PERFORMANCE ANALYSIS OF VARIOUS DATA MINING CLASSIFICATION TECHNIQUES ON HEALTHCARE DATA", International Journal of Computer Science & Information Technology (IJCSIT) Vol. 3, No 4, 2011.

5. Abirami N., Kamalakannan T., Muthukumaravel A., "A Study on Analysis of Various Datamining Classification Techniques on Healthcare Data", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 7, pp. 604-607,2013.
6. Qasem A. Radaideh A. Nagi A., "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.
7. Gibert K., Sànchez-Marrè M., Codina V., "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation", International Environmental Modelling and Software Society (iEMSs) 2010.
8. Witten I. H., Frank E., "Data mining: Practical machine learning tools and techniques with Java implementations". Morgan Kaufmann, San Francisco, CA. USA, 2000.
9. Jain N., Srivastava V., "DATA MINING TECHNIQUES: A SURVEY PAPER". IJRET: International Journal of Research in Engineering and Technology, Vol. 02 Issue 11, 2013.

