A Survey on Data Mining Classification Algorithms

S.Umadevi [1], Dr.K.S.Jeen Marseline [2]
[1] Research Scholar, [2] Head, Department of IT ,
Sri Krishna Arts and Science College,Coimbatore,India
umacbe87@gmail.com,jeenmarselineks@skasc.ac.in

Abstract

Data mining refers to extracting or mining knowledge from large amount of data. It is also defined as finding hidden information from a database. It is a technique which is used primarily for discovering unknown patterns and that converts raw data into user understandable information. Nowadays it is being increasingly used in science and technology to extract the vast amount of data. Classification is the separating the given data according to their characteristics similar to one another. These are some of the classification methods Naïve Bayes Classifier, Decision tree, Neural Networks, and Support Vector Machine.

Keywords-Data mining, Classification, Decision tree, Naïve Bayes Classifier, Neural Networks (NNs), Support Vector Machine(SVM)

I. INTRODUCTION

Due to the advancement in technology there are large amount of unprocessed information. It is time consuming to view or extract the needed information. In such a situation we are in need to develop a strategy which is useful to obtain the necessary information. Since there are large amount of data decision making process is tedious. To overcome these pitfalls the concept of Data Mining is used. The techniques of data mining will help the users to acquire the essential information [1].

Data mining is the process of filtering relevant data according to one's business interests from the huge collection of data using different techniques and algorithms such as Association, Clustering and Classification [17]

The steps involved in knowledge extraction are as follows:

1. Data Cleaning: The information obtained may contain some errors which is preprocessed in this stage.
2. Data Integration: Data available in various forms that are to be integrated.
3. Data Selection: The data which is suitable for user application.
4. Data Reduction: Since they are large amount of data it occupies more space, so using this method we reduce the space but it achieves the same results.
5. Data Mining: A new methodology to extract the essential data.
6. Pattern Evaluation: It is the process in which a pattern is identified.
7. Knowledge Representation: This is the final stage in which the knowledge is represented using different visualization techniques.

The classification is one of the major tasks in data mining. The idea behind this is to classify the given data records into one of the many possible cases which are known already. Classification tasks can make use of any one strategy. If the data are classified without looking at the training data, this kind of classification is known as priori classification. But in converse if the data were classified with the help of training data this is known as posteriori classification.

II. CLASSIFICATION

Data mining techniques broadly classified into two categories. They are predictive and descriptive. Both of these methods are used to extract the hidden patterns from huge amount of data. Classification is the process of converting the data records into set of classes. It is divided into Supervised classification and unsupervised classification. In supervised classification, the data that are to be classified is previously known based on few assumptions. In Unsupervised classification, the set of cases were not predicted by the

users. By some assumption it is the job of the user to classify the given data and try to assign the name for those cases. This type of classification is known as clustering.

Classification involves predicting a certain outcome based on a given input. In order to predict the results, it needs to fetch the data already available. Based on this data the records are classified. The data sources can be categorized into training set and test set. The training set contains the data which are classified before and it used as a reference for classification purpose. With the help of the attributes the results are predicted. Next the test data is supplied to the algorithm. These data are checked against the attribute which are stored previously and based on these assumptions the data are classified. The algorithm analyses the data given and predicts the results.

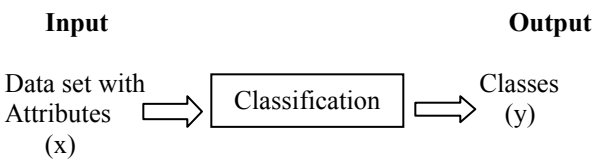

Figure: 1 Classification from x to y

Classification algorithm includes

- Decision Tree
- Naïve Bayes
- Neural Networks
- Support Vector Machine

III. DECISION TREE

A decision tree is widely used classification technique. The methodology used here is Divide and conquer. As there were huge amount of data, first we need to divide those data into sub data. The structure of the decision tree is organized in a manner that it contains the root the topmost node in the tree, Branches which are the internal nodes and leaf node is one which is not further classified. The internal nodes represent a question and the branch which connects the node denotes the solution and the leaf node tries to predict the solution. It is widely used in decision making process. Say for example to predict the patients who have swollen glands and diagnose the type of infection.

The users can be able to construct the decision tree with small amount of domain specific knowledge and therefore it is suitable for data mining process. The various decision tree algorithms are ID3, C4.5, C5, and CART. There are numerous methods for finding the feature that best divides the training data such as information gain (Hunt et al., 1966) and gini index (Breiman et al., 1984). They defined a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability[14]

During the construction of the decision tree there may be some of the uncovered errors. These errors would be rectified at the beginning itself, otherwise it leads to the wrong decision making process. Such errors can be corrected and evaluated using an approach called tree pruning. There are two methods of pruning were employed. One is pre pruning and the other is post pruning. In pre pruning method, pruning is done in the initial stage. It checks for the anomalies, if so it stops the construction of the tree at that stage itself. By halting the node becomes a leaf node. In post pruning method, the entire tree is built. Then from the root it starts pruning by removing the sub trees. An individual sub tree is taken from the original tree and if it finds any anomaly the corresponding branches are removed and replacing it with a leaf.

Milan Kumari.et al [8] used decision tree to predict cardiovascular disease and it classify the patients who have swollen glands and diagnose the patients whether they are infected from fever, cold or throat pain. The dataset used here is Cleveland cardiovascular disease dataset from UCI repository.

AI-Radaideh, et al [9] applied a decision tree method to predict the grade of the student. Even though various classification techniques were applied on the same, decision tree gives the better result. The data set is collected from Yarmouk University, Jordan who has taken C++ course in the year 2005.

Bangsuk Jantawan, Cheng-Fa Tsai[11] used one of the decision tree algorithm called J48. It is used to predict the status of the students after graduation. In this case some of them were classified as employed, some unemployed. The dataset is taken from Maejo University in Thailand. They obtained an accuracy of 98.31% by using J48 algorithm.

Yadav, Bharadwaj, and Pal [10] used decision tree to predict the students past performance. Some of them were excellent in their studies, some of them are average students and others fall in below average category. Those students are identified and teachers are requested to give special coaching to improve their performance

IV. NAÏVE BAYES CLASSIFIER

The base for the naïve bayes classifier is bayes theorem. A hypothesis is generated for the given set of classes. In Naïve bayes algorithm independence assumption is made. Based on the target value, the values of the attribute are chosen and it is independent to one another[7].

The approach used in Naïve Bayes classifier is very simple. With the help of small amount of training data it is possible to classify the given instances [16]. For example to predict the fruit as "apple", based on the color red, and its shape round it is classified as apple which shows it as an independent model. This method is also suitable for complex situations.

George Dimitoglou et al[3], applied naïve bayes classifier to predict the patients who are infected with lung cancer. This algorithm is applied over different kinds of training data. J48 algorithm is also used. The data source used here is from SEER. They achieve the accuracy rate of about 90% over the above mentioned algorithms.

Conditional probability serves as a base for the naïve bayes algorithm which is used to calculate present and past frequency occurrences.

$$P(A\backslash B)=P(B\backslash A)*P(A)/P(B)$$

Where

P(A) is the prior probability of A. It counts only the occurrences of A.

P(A\B) is the conditional probability of A, given B. It is also called as posterior probability which means A is derived from B

P(B\A) is the conditional probability of B, given A

P(B) is the prior probability of B.

V. NEURAL NETWORKS

Neural networks (NNs), more accurately called Artificial Neural Networks (ANNs). It is expressed in terms of biological neuron system. It consists of number of separate units. The individual units are communicated to each other by sending signals. It is similar to the brain composed of many processing components. It is organized as a directed graph which contains nodes and the edges connecting each node. The edges are the interconnections between each node.

Consider a firing rate of each and every neuron. As we said before, the neurons are interconnected. It receives 'm' inputs from 'n' nodes. Each edge connecting the node contains weight. The sum of the weights is calculated. The threshold value is assigned to each neuron. If the weighted sum is greater than the threshold value it produces the output 1 otherwise 0.

S.Gopika [13] used SVM and ANN for diagnosing the renal disease. Comparatively ANN achieves the better result in prediction. UCI repository is used as the dataset

Dr. S. Vijayarani et al [5] focused on predicting the kidney disease with the help of SVM and ANN. The efficiency of these two algorithms was compared based on accuracy and execution time. Finally the results proved that ANN is better than SVM.

The topologies of Artificial Neural Network are FeedForward and Feedback. In FeedForward approach, the data flows only in one direction so it does not receive any acknowledgement from the receiver side. Feedbacks cannot be sent in case of indication of the errors. In case of recognizing images or identifying the fingerprint patterns this method is appropriate. The inputs and outputs are not changed. In Feedback approach, it is possible to send the feedbacks. It is little more efficient because of passing the indications then and there.

A classic application [4] for NN is image recognition.

- Quality assurance, this strategy is applied to test whether the material is of good standard.
- Medical diagnostics, with the help of the pictures for identifying diagnosis.
- Detective tools, by analyzing the fingerprints to a database.

VI. SUPPORT VECTOR MACHINE

Support Vector machine is a supervised machine learning technique which is extensively used in pattern recognition, classification or regression challenges. It is particularly used in noisy and complex domains. The parameters are identified by solving a quadratic equation which involves equality and inquality constraints. The data item is plotted in n-dimensional space with the value of each feature being the value of a particular coordinate. The hyper plane has to created so as to differentiate the classes.

Dr. S. Vijayarani, Mr S. Dhayanand [13] used SVM for detecting the liver disease. Naïve Bayes algorithm is also employed for the same purpose. The efficiency of these two algorithms are compared and shows that SVM produces better prediction results. But naïve bayes completes its execution with minimum running time.

For a linear dataset, hyperplane *f(x)* is defined which passes through the classes and hence separating it [2].

VII. PERFORMANCE METRICS

For each and every algorithm applied so far the performance has to be evaluated. Only based on these evaluation criteria we will able to find out which algorithm suits for certain kinds of application. The tool used here is confusion matrix and receiver operating curve [6].

The confusion matrix shows the number of accurate and inaccurate predictions made by the model. After predicting, the training and test data are to be compared. The attributes of the confusion matrix are true and false.

True positive Rate (TP Rate) is the fraction of positive cases predicted as positive.
False positive Rate (FP Rate) is the fraction of negative cases predicted as positive.
True negative Rate (TN Rate) is the fraction of negative cases was correctly classified as negative.
False negative Rate (FN Rate) is the fraction of positive cases that were incorrectly classified as negative.

ROC is another criterion for evaluating the performance of the algorithms. The graph is plotted against true positive and false positive rates. False positive rate denotes the x-axis and true positive rates for the y-axis. The point(x,y) on the graph denotes any one of the following below:

- The point (0,1) in the graph shows that the instances are classified exactly as positive and negative cases;
- The point (1,1) denotes all cases to be positive;
- The point (1,0) indicates a classifier that classifies all instances incorrectly.

Accuracy

Accuracy [4] is defined as number of correctly classified instances divided by the total number of instances present in the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

TP Rate

It is the ability to find the high true-positive rate. The true-positive rate is also called as sensitivity/recall.

$$TPR = \frac{TP}{TP+FN}$$

Precision

Precision is defined as instances classified as positive divided by number of entire modules classified fault-prone.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F-Measure

It is the combination of both precision and recall which is used to compute the score.

$$\text{F-Measure} = \frac{2*recall*precision}{recall+Precision}$$

VIII. CONCLUSION

This paper gives the basic idea about classification techniques. Decision tree, naïve bayes, Neural Networks and support vector machine were discussed. These algorithm performances can be

evaluated using the criteria Sensitivity, Accuracy, Error Rate, precision, recall and f-measure. Decision tree algorithm is used when the user doesn't know the in depth knowledge of the domain. SVM is used in the context of minimum execution time. Naïve bayes is used if the application follows independent feature model. The goal of Classification algorithms is to produce precise and accurate results.

REFERENCES

[1] Aarti Sharma Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, "Application of Data Mining: A Survey Paper", International Journal of Computer Science and Information Technologies, Vol. 5(2), 2014, 2023-2025 ISSN:0975-9646.

[2] S.Neelamegam,Dr.E.Dharmaraj, "Classification Algorithm in Datamining:An Overview", International Journal of P2P Network trends & technology, Vol.4,issue 8-Sep 2013.

[3] George Dimitoglou, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability".

[4] Fiona Nielsen,"Neural Networks-Algorithms and Applications".

[5] Dr.S.Vijayarani,Mr.S.Dhayanand,"Datamining Classification Algorithm for Kidney Disease Prediction", International Journal on Cybernetics & Informatics, Vol 4,Nov 4,Aug 2015.

[6] Cristina Opera,"Performance Evaluation of the Datamining Classification Methods"

[7] Jaiwen Han,Micheline Kamber," Data Mining Concepts and Techniques"

[8] Milan Kumari, Sunila Godara," Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. 2, Iss ue 2, June 2011

[9] AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[10] S.K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," International Journal of Innovative Technology and Creative Engineering, vol. 1(12), pp. 13–19, 2011

[11] Bangsuk Jantawan, Cheng-Fa Tsai," The Application of Data Mining to Build Classification Model for Predicting Graduate Employment,International Journal of Computer Science and Information Security,Vol. 11, No. 10, October 2013

[12] Dr. S. Vijayarani, Mr.S.Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, ISSN: 2278 – 7798, April 2015.

[13] S.Gopika,Dr.M.Vanitha,Survey on Prediction of Kidney Disease by using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 1, January 2017

[14] Nikita Jain, Vishal Srivastava, " data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology

[15] Anunciacao Orlando, Gomes C. Bruno,Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining Approach for detection of high risk breast cancer groups", Advances in soft computing, Vol. 74, pp. 43-51,2010

[16] R. Savundhariyalachmi, N. Pandimeena, P. Ramya," Study of Classification algorithm in Data mining", International Journal of Science and Research

[17] Samiddha Mukherjee, Ravi Shaw, Nilanjan Haldar, Satyasaran Changdar, " A survey of Data Mining Applications and Techniques", International journal of Computer Science and information Technologies, Vol.6(5), 2015