International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 01, January 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

THE DECISION TREE LEARNING ALGORITHM USING NBTREE FOR CLASSIFICATION OF DATA

Subhendu Kumar Pani[1], Dr.Maya Nayak[2]

[1]Associate Professor, [2]Professor, Dept. of CSE, OEC, BBSR

ABSTRACT

Data mining is the technique of maintaining a large amount of data stored in the database. Classification is a supervised learning approach, which maps a data item into predefined classes. Number of researchers used many data mining techniques for prediction purpose and many of them were in practice. Decision tree is a method of data mining which classify the data and produces important results. These results are used in the study and future prediction. The prime objective of this research work is to present an NBTree decision tree algorithm that classifies the data more efficiently and effectively than existing other decision trees. These results are used in analysis and future prediction. We apply existing decision tree classifiers NBTree on a IRIS data.

Keywords: Classification, NB Tree, Data Cleaning, Data Mining, Decision Tree

I. INTRODUCTION

Classification is a tree based structure which is a concept of data mining (machine learning) technique. It used to predict data instances through attributes [1]. Classification is a method where one can classify future data into known classes. In general this approach uses a training data set to build a model and test data set to validate it. Popular classification techniques include decision trees, Naïve Bayes, Logistic regression, etc. The accuracy of the supervised classification will be much better than unsupervised classification, but depends on prior knowledge. J48 tree algorithm basically uses the divide-and-conquer algorithm by splitting a root tree into a subset of two partitions of child nodes [10]. Random Forest is a machine learning classifier that works over much iteration of the same technique but with a different approach [2,6]. Reduced Error Pruning performed as well as most of the other pruning methods in terms of accuracy and better than most in terms of tree size [3]. It is very difficult to select any prediction techniques in practical situation, because prediction depends on many factors like nature of problem, nature of data set, uncertain availability of data. Machine learning algorithms are most significant classifiers to solve a variety of problems in software development and mainly in software fault prediction. Prediction of faulty and non-faulty modules has been done by so many researchers and organizations.

II. CLASSIFICATION LEARNING ALGORITHMS

Classification techniques can be compared on the basis of predictive accuracy, speed, robustness, scalability and interpretability criteria [4]. In data mining classification tree is a supervised learning algorithm. So one can prepare popular classifiers: J48, Random Forest, Reduce Error Pruning, and Logistic Model Tree. For

comparison purpose, authors have also prepared the fault-prone filtering techniques. A classification model is able to identify the fault-prone (fp) module correctly. The algorithm C5.0 is superior to C4.5. J48 is the enhanced version of C4.5 but the working of algorithms are very similar [7,8]. The goal of decision tree is to predict to response on a categorical dependent variable to measure a more predictor**.** The iris dataset uses a 4 attributes and 150 instances.

2.1 Decision Trees

A decision tree is a flow-chart-like tree structure. The internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution [4][9]. The top most nodes in a tree shown by oval is a root node. Further internal nodes are represented by rectangles, and leaf nodes are denoted by circles. A decision tree consists of nodes [2]. Each node represents some information. Decision tree learning is started from root node and discrete values are produced at each node by testing the values of attribute. These discrete values acts as target function. Then by using target function, value of attribute for next node is evaluated. This process is repeated for each new node. The learned tree is represented by if-then rules. Decision tree algorithms [3] such as ID3, C4.5, J48 [4] NBTree can be applied on large amount of data and valuable predictions can be produced. These predictions evaluate future behavior of problem. Decision tree are preferred because they can evaluate information more accurately than other methods. In this research work following decision tree algorithms are used: (1) ID3: ID3 means Iterative Dichotomiser 3. It is a decision tree algorithm which is developed by Ross Quinlan. The steps of ID3 algorithm are as following: (a) ID3 is a greedy algorithm in which the tree created from top to bottom. (b) At each node, the appropriate attribute is selected which best classifies the data. Data is in the form of training examples. (c) The above process is repeated until the complete tree is generated or until all the attributes used. (2) J48: J48 is the open source Java implementation [5] of C4.5 decision tree algorithm in Weka data mining tool. Following are the steps of J48 algorithm: (a) this algorithm uses basic algorithm which create trees by using recursive top down divide and conquer approach. (b) First of all, the training examples are at the root node. (c) Test attribute is selected based on some measures such as information gain, entropy etc. (d) Examples are divided repeatedly by using test attribute. (e) The process continued until no sample leaf is leaf. (3) NBTree: NBTree (Naive Bayesian tree) consists of [6] naïve Bayesian classification and decision tree learning. An NBTree classification sorts the example to a leaf and then assigns a class label by applying a naïve bayes on that leaf. The steps of NBTree algorithm are: (a) At each leaf node of a tree, a naive bayes is applied. (b) By using naive bayes for each leaf node, the instances are classified. (c) As the tree grows, for each leaf a naive bayes is constructed. (d) This process repeated until no example is left.

III. PERFORMANCE MEASURES FOR CLASSIFICATION

One can use following performance measures for the classification and prediction of fault prone module according to his/her own need.

Confusion Matrix: The confusion matrix is used to measure the performance of two class problem for the given data set. The right diagonal elements TP (true positive) and TN (true negative) correctly classify Instances as well as FP (false positive) and FN (false negative) incorrectly classify Instances. Confusion Matrix Correctly Classify Instance TP+TN Incorrectly Classify Instances.

IV. EXPERIMENTAL WORK AND DISCUSSIONS

NBTree with Add classification Filter in weka. The supervised weka filter Add Classification is applied to the data set for preprocessing. Then the NBTre Weka classifier is applied on the IRIS dataset with 150 instances and 4 attributes of the format (arff) Attribute relationship file format which supports weka tool. The result obtained is given below. This implementation uses 10-fold cross-validation test mode. The time taken to build the model is 0.23 seconds. The accuracy obtained by NBTree with Add classification Filter in weka is 94.66 % and the other measures were given in

Table-1. NBTree visualize is depicted in Figure1.

Time taken to build the model	0.23 seconds
Correctly Classified Instances	142(94.6667%)
Incorrectly Classified Instances	8(5.3333%)
Kappa statistic	0.92
Mean absolute error	0.0497
Root mean squared error	0.1709
Relative absolute error	36.2534
Root relative squared error	36.2534
Total Number of Instances	154

Table-1: NB tree with add classification filter in WEKA.

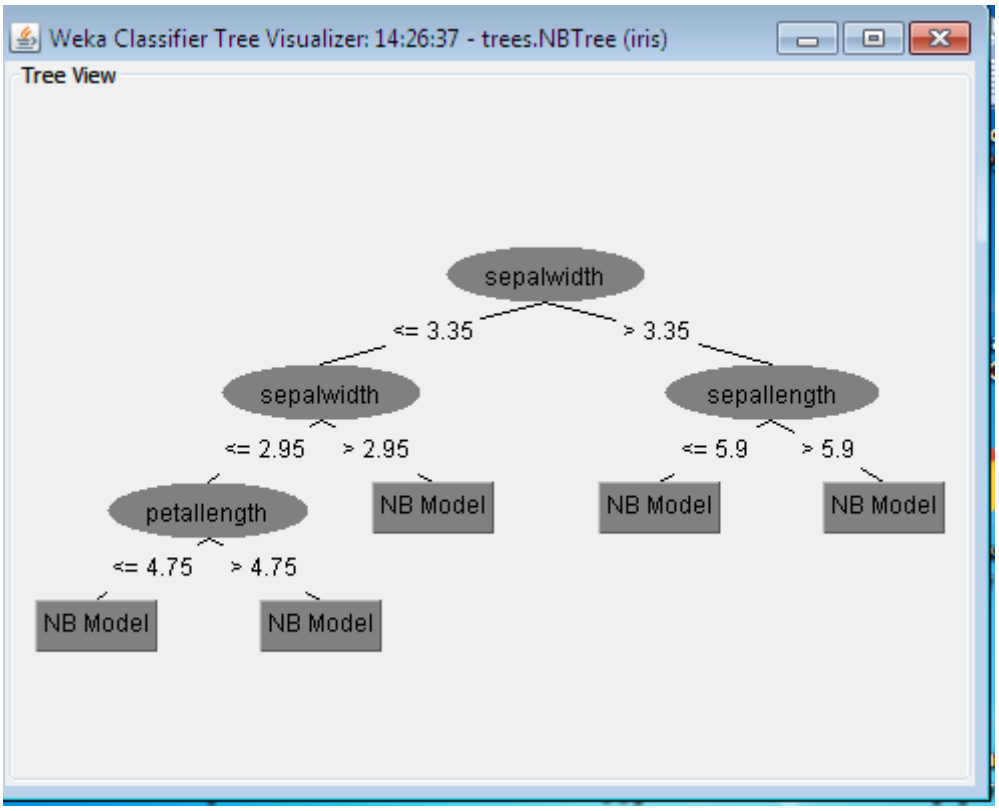

Figure1: NBTree visualize

V. CONCLUSIONS

In this paper, we have applied supervised weka filters Add classification for preprocessing the data on the IRIS dataset. The preprocessed data is given as input to the NBTree classifier and the result obtained is analyzed for measuring performance. Hybrid data preprocessing techniques can be applied in future to improve the accuracies still in prediction systems for further research.

REFERENCES

[1] Jiawei han and micheline kamber. Data mining concepts and techniques, second edition,285-291

[2] J. R. Quinlan, 'Introduction of decision tree', Journal of Machine learning.

[3] Mrs. Swati .V. Kulkarni, ''Mining knowledge using Decision Tree Algorithm'', International Journal of Scientific & Engineering Research, Volume 2, Issue 5.

[4] Youvrajsinh Chauhan, Jignesh Vania, "J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient", International Journal of Computer Trends and Technology (IJCTT), volume 5 number, 2013.

[5] Bangsuk Jantawan and 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.

[6] Gehrke, J., Ramakrishnan, R., Ganti, V. (1998). RainForest - a Framework for Fast Decision Tree Construction of Large Datasets.Proceedings of the 24th VLDB conference, New York, USA. pp.416- 427 [7] Hunt, E.B., Marin. and Stone,P.J. (1966). Experiments in induction, Academic Press, New York.

[8] Khoshgoftaar, T.M and Allen, E.B. (1999). Logistic regression modeling of software quality. International Journal of Reliability, Quality and Safety Engineering, vol. 6(4, pp. 303-317.

[9] Kufrin, R. (1997). Decision trees on parallel processors. In J. Geller, H. Kitano, and C. B. Suttner, editors, parallel Processing for Artificial Intelligence 3 Elsevier Science.

[10] Tzung-I tang,Gang Zheng, Yalou huang,Guangfu Shu,Pengtao wang. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS vol.4,no.1,pp-102-108,june 2005