International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 5 Issue: 1
ISSN: 2454-4248
77 – 80

Forest Tree Algorithm- An Efficient Approach of Data Mining Over Decision Tree

Anuradha
Computer Science & Engineering
Golden College of Engineering and Technology
Gurdaspur, India

Suraj Pal
Computer Science & Engineering
Golden College of Engineering and Technology
Gurdaspur, India

Mohit Angurala
Asst. Prof. GNDU Regional Campus,
Gurdaspur,
Punjab, India

***Abstract*—** Mining of Data (DM) is a way to display different models, summaries and values derived from a given data collected. The DM itself works in the process of searching for analytical information on the large number of available databases. An example of a predictive enigma is targeted marketing. There are many factors that affect data mining performance in large data sets. In this article we will use the forest tree technique to improve performance in search for data and implementation, surely overcome the previous work performance that includes the approach of the existing tree decision tree algorithm.

***Keywords*—** *DM, ITS, AI, DB*

*****

I. INTRODUCTION

Remote resources such as computers, databases, files, etc., along with people like analysts, professionals, end users are often involved in the overall data analysis process. This analysis is omnipresent and is very important for an application that deals with finance, process control, defense, and many other domains. The ability to analyze large amounts of data is the demand for these applications. Technique Decimal tree are CART, ID3 and C4.5 minerals that are scalable and fast and are for ubiquitous device flow monitoring data like computers, handhelds, etc.

Data Mining Functions and Methods - There are some data mining systems to provide a single data mining function as a classification, while others provide multifunction based mining, such as concept description, OLAP discovery, mining associations, linkage analysis, Statistical analysis, classification, prediction, clustering, outlier analysis, search by analogy, etc.

Database data mining pairing or data storage systems - data mining systems to be paired with a database or data warehouse system. The paired components are integrated into a uniform data processing environment. Here are the types of links listed below -

- Without joint
- paired coupling
- semi-rigid
- Coupling
- Narrow coupling

Scalability - There are two scalability issues in data mining:

Row scalability: a system mining is considered scalable when the number or lines are magnified 10 times. It does not take more than 10 times to run a query.

(size): A system mining is considered scalable in columns if the run time of the mining query increases linearly with the number of columns.

II. LITERATURE REVIEW

Gyozo Eesha Goel et al. 2017 Ensemble is a data mining technique composed of number of individual classifiers to classify the data to generate new instances of data. Random Forest is the most popular ensemble technique of classification because of the presence of excellent features such as Variable importance measure, Out-of-bag error, Proximities etc. In this paper, the developments and improvements of Random Forest in the last 15 years are presented. This paper deals with the approach proposed by Brieman since 2001. This paper also presents the description of usage of Random Forest in various fields like Medicine, Agriculture, Astronomy, etc.

Prajwala T R et al. 2015 This paper discusses two classification algorithms namely decision trees and Random forest.. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Random forest includes construction of decision trees of the given training data and matching the test data with these. Rattle an open source R-GUI is used for analysis of weather data for prediction of rainfall using 256 data samples. Based on results obtained a comparative analysis is done.

Zakariah et al. 2014 This paper discusses many applications which use Random Forest to classify the dataset like Network

 ISSN: 2454-4248
77 – 80

intrusion detection, Email spam detection, gene classification, Credit card fraud detection, and Text classification. In this paper each application is briefly introduced and then the dataset used for implementation is discussed and finally the real implementation of Random Forest algorithm with steps wise procedure and also the results are discussed. Actual Random Forest Algorithm and its features are also discussed to highlight the main features of Random Forest Algorithm more clearly.

Barrett Lowe et al. 2015 Random Forest grows many decision trees for classification. To classify a new object, the input vector is run through each decision tree in the forest. Each tree gives a classification. The forest chooses the classification having the most votes. Random Forest provides a robust algorithm for classifying large datasets. The potential of Random Forest is not been explored in analyzing multispectral satellite images. To evaluate the performance of Random Forest, we classified multispectral images using various classifiers such as the maximum likelihood classifier, neural network, support vector machine (SVM), and Random Forest and compare their results.

Kalmegh et al. 2015 In this paper, the basic way of interacting with these methods is by invoking them from the command line. However, convenient interactive graphical user interfaces are provided for data exploration, for setting up largescale experiments on distributed computing platforms, and for designing configurations for streamed data processing. These interfaces constitute an advanced environment for experimental data mining. Classification is an important data mining technique with broad applications. It classifies data of various kinds. This paper has been carried out to make a performance evaluation of REPTree, Simple Cart and RandomTree classification algorithm. The paper sets out to make comparative evaluation of classifiers REPTree, Simple Cart and Random Tree in the context of dataset of Indian news to maximize true positive rate and minimize false positive rate. For processing Weka API were used.

Gracia Jacob et al. 2015 It works by incorporating best feature selection algorithm with the Random Forest to gives better accuracy. Correlation based Feature Subset Selection algorithm selects the optimal subset of features. The optimal features are fed as a part of Random Forest classification to give better accuracy in software defect prediction. The six optimal subset of features were selected for PC1 dataset. The features are selected by the CFS and utilized by Random Forest to improve the accuracy of existing Random Forest. The experiments were carried on publicNASA datasets of PROMISE repository

III. IMPLEMENTATION WORK

Data mining is the process of discovering interesting patterns, pattern evaluation and knowledge presentation that allow the users to analyze data from the different dimension, categorize it and summarize the relationships which identified during the data mining process. We will carry out research in following steps:

- Use Python Script Tool
- Take Student Dataset
- Select the processor from Data Set
- Implement Random Forest & Tree Algorithm on dataset
- Display the predictions from input Data Set.
- Cross validation on Implemented result
- Our implementation work show experimentations which includes different input output values in student data set. Different outputs are generated for each case when inputs are changes different data sets.

1. Firstly, we open orange 2.7 tool . Then ,click on file option for taking new file with file name. Then ,click on OK button .

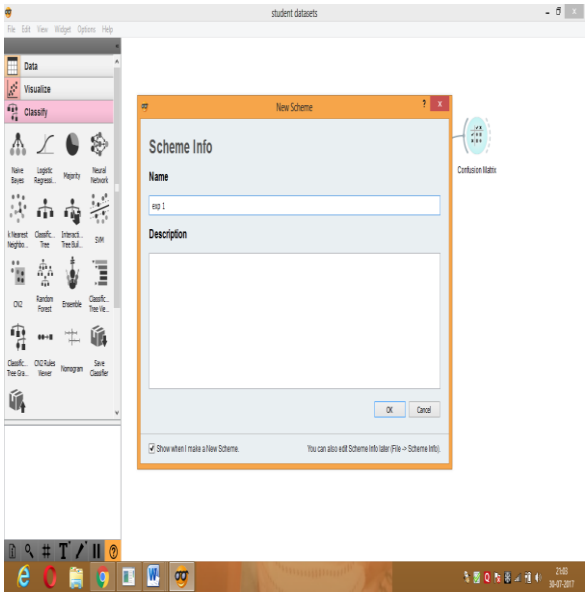

Fig 1: Selection of File

2. After create a file, to make scenario on which we work on it. The scenario makes easily because all terms are drag-drop on it .

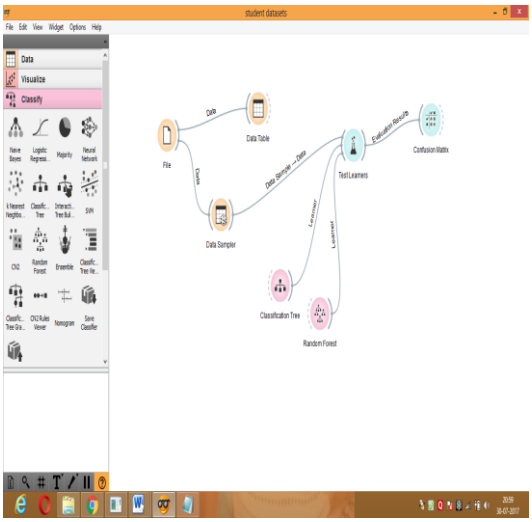

Fig 2: Creation of Scenario

 ISSN: 2454-4248
77 – 80

3. In third step, double click on file icon. After click,to select the file then click on reload option .

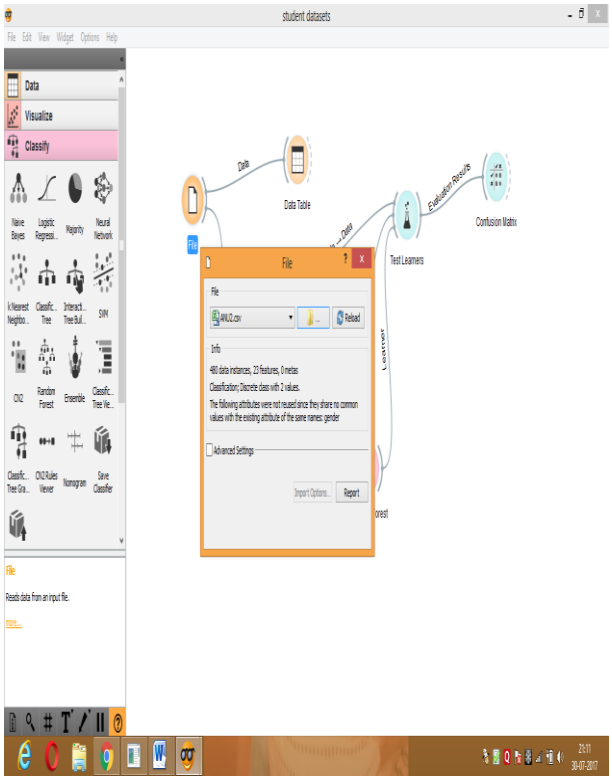

Fig 3: Extraction of file

4. After reload the file, to find the result . For finding the result, double –click on Test-learner icon. Then, results are shown with comparison between classification tree algorithm and random forest tree algorithm .

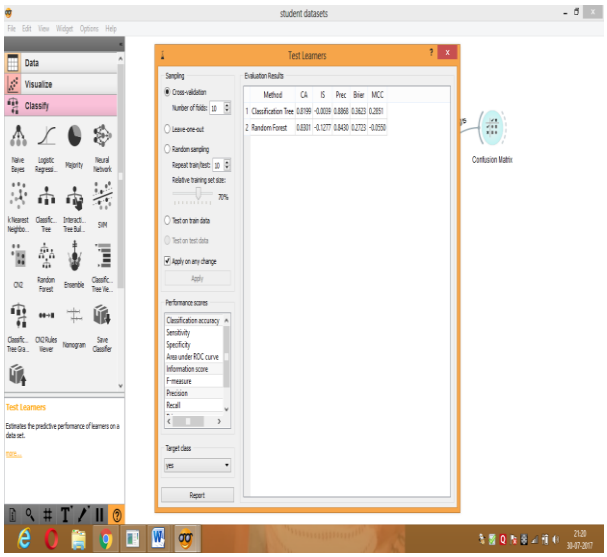

Fig 4: Find the results

5. For finding the prediction values, double-click on confusion–matrix. Then the first prediction of classification tree is shown then random tree prediction shown.

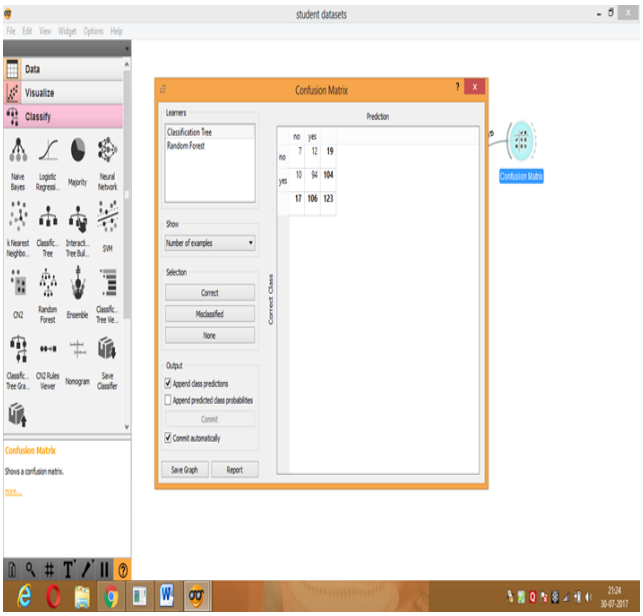

Fig 5: Prediction values for classification tree

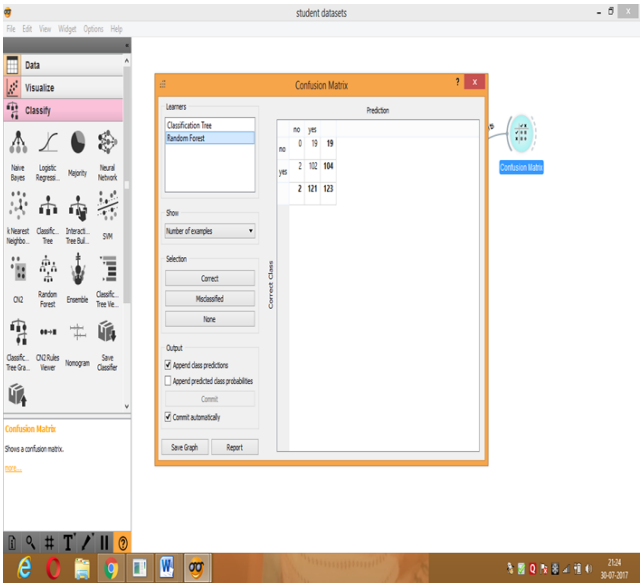

Fig 6: Prediction values for random forest tree

Conclusion

In previous research work, Classification tree algorithm is used but in this research work the Random Forest tree algorithm is used to enhance the performance of traditional algorithm. According to the comparison between Classification tree algorithm and Random Forest tree algorithm, the value of CA in Random tree algorithm is 0.96 rather than the Classification tree algorithm is 0.94. Similarly, the IS value in Random Forest tree algorithm is -0.33 rather than the Classification tree is -0.16. The prediction value of Random tree algorithm is 0.9624 rather than the Classification tree algorithm used. To conclude, the better performance can be achieved an implementing the Random Forest tree algorithm rather than the Classification tree algorithm.

 ISSN: 2454-4248
77 – 80

REFERENCES

[1]. Gyozo Gidofalvi (2007). Privacy-Preserving Data Mining on Moving Object Trajectories in International Conference on Mobile Data Management. 60-68.

[2]. Isaac Cano (2009). Generation of Synthetic Data by means of fuzzy c-Regression in FUZZ-IEEE. 20-24.

[3]. Haiyang Zheng et al., (2009). Predicting the power of a wind farm at different time scales IEEE. 20-24.

[4]. Isaac Canoand Susana Ladra (2010). Evaluation of Information Loss for Privacy Preserving Data Mining through comparison of Fuzzy Partitions inIEEE International Conference on Fuzzy SystemS. 1-10.

[5]. Sebastián et al., (2010) . Educational Data Mining: A Review of the State-of-the-Art inIEEE International Conference on Fuzzy SystemS. 1-10.

[6]. Weijia Yang (2010). A novel anonymization algorithm: Privacy protection and knowledge preservation in Expert Systems with Applications 37 (1). 756–766.

[7]. Jingjing Qiang (2011). Privacy-Preserving SVM of Horizontally Partitioned Data for Linear Classification in International Congress on Image and Signal Processing. 23-27.

[8]. Khaled Alotaibi (2012). Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification in ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust. 334-337.

[9]. Kshitij Pathak (2012). Privacy Preserving Association Rule Mining by Introducing Concept of Impact Factor in IEEE Conference on Industrial Electronics and Applications. 1458-1461.

[10]. Majid Bashir Malik (2012). Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects in Third International Conference on Computer and Communication Technology. 26-32.

[11]. Lei Xu (2014). Information Security in Big Data: Privacy and Data Mining in IEEE Publications 2 (1). 1149-1176.

[12]. P.Usha (2014). Sensitive Attribute based Non-Homogeneous Anonymization for Privacy Preserving Data Mining in International Conference on Information Communication and Embedded Systems. 1-5.

[13]. N P Nethravathi (2015). CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining in IEEE International WIE Conference on Electrical and Computer Engineering. 190-194.

[14]. Qi Jia (2016). Privacy-preserving Data Classification and Similarity Evaluation for Distributed Systems in IEEE 36th International Conference on Distributed Computing Systems. 690-699.

[15]. Sunil Dutt Jha. 5 data mining techniques for optimal results. http://www.computerweekly.com/tip/5-data-mining-techniques-for-optimal-results. Accessed on July, 2016.

[16]. Jehad Ali (2012). Random Forests and Decision Trees IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814.

[17]. Keshari Ratha (2015) Decision tree analysis on j48 and random forest algorithm for data mining using breast cancer microarray dataset International Journal of Advanced Technology in Engineering and Science Issues, Vol. 3, Issue 5, No 1, November2015.

[18]. Ahmed Zriqat (2016) A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016

[19]. Eesha Goel (2017) Random Forest: A Review International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 1, January 2017

[20]. Prajwala T R (2015) A Comparative Study on Decision Tree and Random Forest Using R Tool International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2015.