International Journal of Electrical Electronics & Computer Science Engineering
Special Issue - NCSCT-2018 | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
March, 2018 | Available Online at www.ijeecse.com

A Study on the Classification Using Random Forest

Ketan Agarwal[1], Krutibash Nayak[2]
[1]UG Scholar, [2]Assistant Professor, CSE Department, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan
[1]2014pietcsketan@poornima.org,[2]kruti@poornima.org

***Abstract:** This review paper tells about the use of random forest algorithm which is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Random forest algorithm can use both for classification and the regression kind of problems.*

***Keywords:** Meta Random Forest, Relief Random Forest, Dynamic Integration of Random Forest, Small Random Forest, Forest-RK, BAGA Algorithm, Dynamic construction of Random Forest.*

I. INTRODUCTION

Random Forest or random decision forest are a gathering learning technique for arrangement, regression and different undertakings, that work by building a huge number of decision trees at preparing time and yielding the class that is the method of the classes(classification) or mean expectation (regression) of the individual trees. Forest of trees splitting with oblique hyperplanes, if randomly restricted to be insensitive to only selected feature dimensions, can gain accuracy as they grow without suffering from overtraining.[5]

Random Forest Algorithm- Random Forest Algorithm is a supervised classification algorithm. It makes the forest with various trees. When all is said in done, the more trees in the forest the more robust the forest looks like. Similarly, in the forest classifier, the higher the quantity of trees in the forest gives the high exactness comes about. Its classifier will handle the missing values. When we have more trees in the forest, random forest classifier won't overfit the model and it can also model the random forest classifier for categorical values.

Decision tree idea is more to the rule based framework. On the off chance that the dataset is given with targets and features, at that point the decision tree algorithm will come up with some set of rules. A similar arrangement of rules can be utilized to play out the forecast on the test dataset. [5]

Some application of random forest algorithm are as follows:

1. Banking
2. Medicine
3. Stock Market
4. E-Commerce

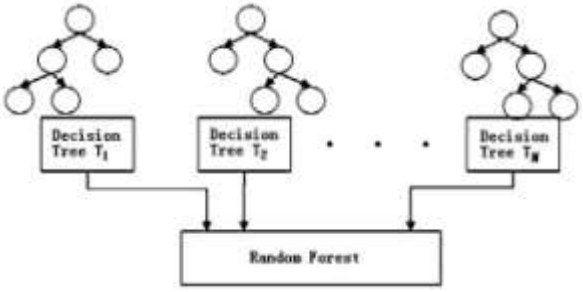

Fig. 1. Random Forest Algorithm from Decision Tree

II. LITERATURE SURVEY

S.N.	Year	Title	Author's Name	Methodology	Identified Problem
1	2001	Random Forests	Leo Breiman	Working for test set errors and mean squared test set error with different data sets like diabetes data etc	Because of the law of large numbers, they do not overfit
2	2009	Research on Machine Learning Framework Based on Random Forest Algorithm	Qiong Ren, Hui Cheng and Hai Han	It has explained the working of random forest framework for machine learning using flow charts	Random forest models are black boxes that are very hard to interpret
3	2012	Analysis of a Random Forests Model	Gerard Biau	The mathematical properties of Random forest	Despite growing interest and practical use, there has been little exploration of the statistical properties of random forests

4	2016	Meta Analysis of Research in Random Forests for Classification	Arnu Pretorius, Surette Bierman and Sarel J. Steel	Data Analysis & Statistical Operations	Random Forests are very fast to train, but quite slow to create predictions once trained
5	2017	Random Forest	Eesha Goel, Er. Abhilasha	Description of usage of Random Forest in various fields like Medicine, Agriculture, Astronomy, etc.	Description of usage of Random Forest in various fields like Medicine, Agriculture, Astronomy, etc.

II. METHODOLOGY

The classification and regression of information relies upon the accuracy and performance of Random Forest. A Random Forest is a troupe method, Experiments are performed with its base classifier to enhance its exactness and performance. [5]

In order to have a good ensemble model, base classifier ought to be differing and exact. Some of the improvement are explained as follows:

A. Meta Random Forest:

Meta learning techniques are applies to the random forest. It is based on the concept that random forest is made as a base classifier. The performance of this model is tested and compared with the random forest algorithm. Meta Random Forest is generated by using bagging and boosting approach. In case of bagging, random forest is taken as a base classifier with the bagging approach and in case of boosting, random forest as a base classifier is implemented with the boosting approach. By comparing all approaches Bagged random had shown excellent research. The boosting random forest fails due to the presence of deterioration in generalizing performances. This is because the complexity of the classifier becomes complex and it is unable to improve its performance on datasets. [5]

B. ReliefF Random Forest:

Through test it is watched that Gini index can't have the capacity to distinguish strong conditional dependencies among the characteristics. This is because it measures the impurity of the class value distribution before and after the split on evaluated attribute. The same behavior is observed from other measures such as Gain ratio, DKM, MDL, j-measure. This problem is solved by ReliefF [Robnik and Sikonja (2004)].

ReliefF in random forest was utilized to assess traits in the pre-processing step and quality appraisals are utilized as weight for choosing subsamples of properties at each level of tree. This decrease the correlation between the attributes while maintaining the strength.[5]

C. Dynamic Integration of Random Forest:

The first Random forest calculation was using combination techniques and selection techniques that are listed below:

1. Combination Technique known as Weighted Voting is utilized where every node will have weight which is relative to the assessed generalization performance of the corresponding classification.

2. Selection Technique known as Cross-Validation Majority (CVM) is utilized where the highest cross-validation accuracy is selected.

These above methodologies are static. They select just a single model or the blend of models consistently. Dynamic combination of random forest was proposed by [Tsymbal et al. (2006)] thinking about each new occurrence. They utilize three methodologies that are recorded underneath:

1. Dynamic Selection (DC)
2. Dynamic Voting (DV)
3. Dynamic Voting and Selection (DVS)

The procedure is utilized for above methods are portrayed underneath:

- First of all, errors of each base classifier on each instance of the training set is evaluated by using cross validation method. Then, K-nearest neighbour of each new instance is estimated.

Lastly, the weighted nearest neighbor is used to evaluate the performance of base classifier.

- DS selects the classifier having least local error.
- In DV, the weight received by each base classifier is proportional to the estimation of local performance.
- In DVS, the base classifiers with the errors are discarded that falls under upper half of the error interval. Remaining classifiers are operated using DV approach.[5]

D. Forest-RK.:

In this hyper -parameter i.e. k number of features are selected randomly at each node during the tree induction process. The value of k is set arbitrary without any theoretical or practical approach. Therefore, a method

for arbitrary setting of value of k is introduced [Bernard et al. (2008)]. This method is known as Forest-RK. In this method, the value of k is not a hyper-parameter which means that k does not play an important role for growing accurate RF classifiers. Rather, this method provides at least statically accurate results as the accurate results are provided by Forest-RI method with default settings.

E. Small Random Forest:

With a specific end goal to build up the random forest, expansive number of trees are created to make the model more steady and less inclined to the predication errors. However, because of the huge number of trees it ends up hard to interpret the forest. To determine this issue [Zhang and Wang (2009)] two targets were proposed to shrink the forest that are recorded underneath:

- To maintain a similar level of accuracy for prediction.
- To reduce the number of trees to intercept easily.

Three measures are considered with a specific end goal to decide the negligible size of the forest.

- A tree will be removed if it has lesser impact on the accuracy of prediction. The procedure is as: first the prediction accuracy of the forest is resolved. Second, for each tree T, the prediction accuracy is resolved of the forest that rejects T trees. The difference from the first forest and the forest excluded T trees are figured. These are least importance and can be expelled from the forest. This technique is known as "by prediction".
- Other remaining techniques are based on similarity of trees. It implies that trees are expelled from the forest that are like other trees in the forest. [5]

F. BAGA Algorithm:

BAGA Algorithm is used for dataset which are too long. This algorithm is an expansion of "Overproduce and Choose" paradigm. BAGA algorithm will use combination of bagging and genetic algorithm in order to generate the component classifiers in proper execution time by considering the set of decision tree as an input. This algorithm can select the classifiers dynamically. First, the set of classifiers are created with decision trees and then genetic algorithm is applied to select the optimal set of classifiers for an ensemble. [5]

G. Dynamic Construction of Random Forest:

This technique conquers the disadvantage of choosing the classifier in advance. In this strategy, powerfully the base classifiers are chosen to form the forest. The development of forest is performed by including each tree. While adding the tree each opportunity to frame the forest the accuracy is recoded and online fitting procedure is applied on the curve to express the variations in accuracy. When the difference between the curve of accuracy and online fitting procedure satisfy certain specific criteria then the procedure will be terminated. [5]

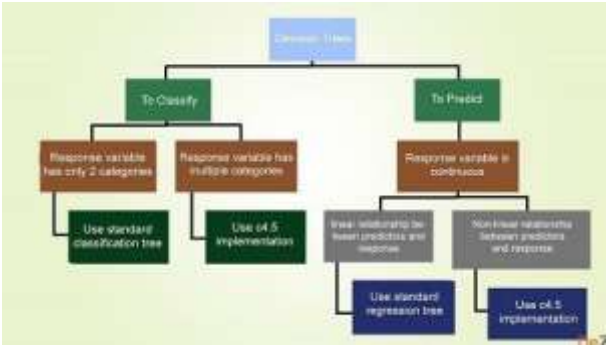

Fig. 2. Uses of Decision Tree

III. APPLICATION OF RANDOM FOREST

A. Banking:

In the banking sector, random forest algorithm widely used in two main application. These are for finding the reliable client and finding the extortion clients.

The dependable client implies not the client who pays well, but rather additionally the client whom can take the immense sum as advance and pays the credit premium appropriately to the bank. As the development of the bank absolutely relies upon the dependable clients. The bank customers data highly analyzed to find the pattern for the loyal customer based the customer details.

Similarly, there is have to recognize the client who are not productive for the bank, such as taking the credit and paying the advance premium appropriately or discover the exception clients. In the event that the bank can distinguish propositions sort of client before giving the advance the client. Bank will get an opportunity to not support the credit to these sorts of clients. For this situation, likewise irregular backforest calculation is utilized to recognize the clients who are not beneficial for the bank.

B. Medicine:

In medicine field, random forest algorithm is used identify the correct combination of the components to validate the medicine. Random forest algorithm is also helpful for identifying the disease by analyzing the patient's medical records.

C. Stock Market:

In the stock market, random forest algorithm used to identify the stock behaviour as well as the expected loss or profit by purchasing the particular stock.

D. E-Commerce:

In web based business, the random forest used only in the small segment of the recommendation engine for

identifying the likely hood of customer liking the recommend products base on the similar kinds of customers.

Running random forest calculation on extensive dataset requires top of the line GPU frameworks. In the event that you are not having any GPU framework. You can simply run the machine learning models in cloud facilitated work area. You can utilize cloud work area online stage to run top of the line machine taking in models from sitting any edge of the world.

Fig. 3. Application of Random Forest

IV. ADVANTAGES OF RANDOM FOREST

- The overfitting issue will never come when we utilize the random forest algorithm in any classification issue.
- The same random forest algorithm can be utilized for both classification and regression undertaking.
- The random forest algorithm can be used for feature engineering which means identifying the most important features out of the available features from the training dataset.

V. CONCLUSION.

This review paper on Random forest concludes the current ongoing work on Random forest and the applications of Random forest. Random forest is an ensemble method which generates accurate results but, on the other hand it is a time consuming method too as compared with the other techniques.

VII. REFERENCES.

[1] Leo Breiman, Random Forests, 2001.

[2] Qiong Ren, Hui Cheng and Hai Han, Research on Machine Learning Framework Based on Random Forest Algorithm, 2009.

[3] Gerard Biau, Analysis of a Random Forests Model, 2012.

[4] Arnu Pretorius, Surette Bierman and Sarel J. Steel, 2016.

[5] Eesha Goel ,Er. Abhilasha, Random Forest, 2017.