ABSTRACT

The decision tree method works by repeatedly dividing the location of features into imaginary limb regions so that each imaginary location provides a basis for making a different approximation. The decision tree system in existence so far applies to various future tasks such as classification and regression. These methods are popular in the field of data science with various benefits. This is due to limitations such as instability of predictions before slight changes in data, and this leads to a major change in the structure of the decision-making tree and has detrimental effects in terms of forecasting. On the other hand, to improve the prediction accuracy of a single base classifier or regressor, multiple decision trees are given parallel training for forecasting purposes and are known as random forests. The random forest technique is an ensemble method, it comprises of several decision tree which are trained on the subset of data or with the feature subspace, once all the tree are trained, their results are combined together for the purpose of prediction. As random forest is more stable than a decision tree it become more popular in the field of data science and machine learning. In this paper, we had provided an detailed introduction of the decision tree methods and random forest method. Also, how they works and for which type of problem they are suitable.

KEY WORDS: SUPERVISED LEARNING, DECISION TREE, RANDOM FOREST, CLASSIFICATION, REGRESSION

INTRODUCTION

The decision tree actually derived from the statistics field and then it is improved by a various researchers for the prediction purpose in the field of data mining, pattern recognition and machine learning. As decision tree comes up with several benefits like human readability but it is having several drawbacks like it suffers a lot when there is a little change in the data i.e. instability. To overcome the drawback of decision tree and to fully utilise its capability one can build random forest for this purpose. As per the name, Random Forest, consist of 'n', number of decision trees which are trained using the subspace of the training data and the result of all decision tree is collectively used for the final prediction of the 'y' label.

Due to large number of tree used for the prediction of 'y', the error of a single tree get overcome or compensated by other decision tree, available in the forest. In recent study, comparison of all the present classifier is done on 121 dataset and it is found that the result for random forest is far much more better than the other forecasting methods. The reason for this great prediction performance is the presence of the base classifier in the random forest. The performance of the random forest is completely dependent upon its base classifier and its training. Also, if subspace is provided in a more structured way to a individual decision tree, it gives more accurate result compare to the random training techniques. The complexity of Random Forest is comparatively low even though it comprises of several decision tree which are trained in parallel.
As far as literature is concerned, several studies discussed about the ensemble learning and its performance, but these studies are generalised and had little focused on Random Forest methods. Also, these studies given an superb feedback on various ensemble learning methods, but they are not that much sufficient. Apart from the previous studies, several researchers proposed a study which completely emphasised on the performance of random forest method and its related methods. The popularity of random forest in the field of machine learning, pattern recognition and data science is due to its superior performance compared to its peers, while there are methods which are present for the construction of random forests. In the recent book publication by criminisi and shotton, they pitched a aggregate model of random forest for handling various prediction learning jobs. In this book, they presented detailed theoretical description of random forest algorithm and its implementation in various prediction tasks. Nevertheless, they failed to explain the various issues related to random forest like pruning methods for handling big data and various processing methods.

The goal of this research paper is to provide an reader to give an introduction to Random Forest algorithm and its variants. In this paper, the detailed working of decision tree is presented and its advantages and disadvantages is discussed. Also, detailed study on construction of Random Forest algorithm is mentioned along with is advantages and disadvantages and suitability of the algorithm with respect to the data is discussed. The article is organised as follows: Section 2 gives an introduction of decision tree which plays a important role in the construction of random forest and its prediction accuracy. In this, we discussed in detail how the accuracy of single base classifier impacts the final accuracy of an ensemble algorithm. In section 3, detailed description of Random Forest algorithm is presented in terms of it working and its applicability to various domains and gives an overview of the Random Forest construction methods. Section 5, gives an general description on the fusing technique and detailed study of present popular techniques related to Random Forest algorithm is shown and finally in the section 6 conclusion is made.

2. The Decision Tree - The Building Block of RandomForest:
The decision tree method works by repeatedly dividing the location of features into imaginary limb regions so that each imaginary location provides a basis for making a different approximation. It consist of nodes and edges. The decision tree building process starts from the root node and it is also known as first split point. This split decides the divisions of the entire dataset. The split is done on the basis of calculation of entropy and information gain or gini index. The splitting process continues from top to bottom until no longer partitioning is required. The leaves present at the end of the decision tree represent the last partitions. In the decision tree, the root node do not contain any incoming edge but the test nodes must contain one edges coming from the upper node and it may have one or more outgoing edges. The decision is made by the decision node also known as leaves node which contains the predicted output.

A tree which consist of only root node (also known as parent node) and leaves and do not contain any test node in it then it is called as decision stump. For the classification type of problems, decision tree is widely used and the goal is to predict the object on the basis of their features. For the purpose of demonstration of working of decision tree we took an dataset sample of Iris flowers. Here, the goal is to predict the Iris flower species based on the the features of the Iris flower like petal length, petal width, sepal length and sepal width. This problem belongs to classification category, here the goal is to predict the type of iris flower. Here the tree will be called as classification tree and the child node will hold the predicted object.

The figure represent the classification tree for the Iris flower dataset, in this, the test node consist of ‘n’ number of object. The first split is based on petal length, which shows the importance of petal length parameter in identification of species, and this split is calculated on the basis of highest value of information gain or gini index. The petal length is called as root node. The petal width node and petal lengths are the internal node or test node, whereas node 2, node 5, node 6 and node7 are the leaves node. Decision trees divided the covariate space into unequal parts. One way to better understand the decision tree is to envisage the covariate space with selected partition boundaries. The below figure presents a two-dimensional projection skeleton that uses selected features to examine the tree. Indivisible decision-making tree divide the space by axis-parallel hyperplanes. The scatterplot shows the corresponding partition of the feature space by adding the boundary of the partition. In this case, the plot consists of four hyperplanes (boxes). Each box classification corresponds to one leaf of a tree. Based on the examples in it, we can visualise the square distribution of each frame.

Figure 1: A decision tree also known as classification tree for Iris_Flower dataset

Benefits of the Decision Tree Method: The reason behind the popularity of decision tree technique is due to the many benefits it offers which are stated below:

a) The structure of decision tree is clear and easy to understand when compared to the other present classifier. That is, if the decision tree has the correct
number of test nodes, it can be easily understood by non-commercial users. In addition, decision trees can be easily transformed into a set of post-rules, so this type of representation is considered understandable. In addition, it provides a selection of built-in functions which makes it more attractive than the rest.

b) Decisions will be made in a way other than parametric, i.e. there is no predefined hypothesis regarding the feature space and its distributions. Also, the decision tree can work with numeric data, data presented in text format and nominal data. The decision tree is capable enough to handle large data and it is scalable. It can provide a fast result with respect to the other classification technique and gives high prediction performance. Due to its popularity among the researcher, it is available in various open source platforms.

**Drawbacks of the Decision Tree Method:** The decision trees also have many shortcomings. As we will see later, some of these deficiencies can be mitigated with a decision to grow the forest. Here are the main disadvantages:

a) The decision tree suffers from the replication issue i.e. the performance of the tree degrades when it comes to intricate interactions between the features. As this technique uses the divide and conquer approach in the training phase, it gives superior performance but when it comes to dealing with highly co-relate feature the performance of the model decreases gradually.

b) It suffers a lot with the change in the data and shows myopic behaviour. Also, during the fragmentation process the overall insight on view of the data get decreases. It also suffers from the dimensionality i.e. if the data is getting fragmented on equal part on every split then the testing of the data on the features is getting reduced.

The Random Forest and Ensemble Learning Methods: The researcher proposed an ensemble’s methods to overcome the drawbacks of a single prediction model. The ensemble comprises of more than one model which are trained by the subset of the data belong to the same datasets. The objective is to achieve higher accuracy compare to the single prediction model. The reason for getting better accuracy, is because of multiple models are involved in decision making. The origin of ensemble was done on seventies when several researchers, proposed an concept of forecasting model where they combined several models together for the better prediction. The first ensemble which was built by Tukey, comprises of two regression models. Later, the authors, proposed an idea of fragmenting the input with two or more model. The author proposed an algorithm Adaptive Boosting in the year 1990, since then random forest received an attention from the data scientist and machine learning experts.

The number of times people proved that the combination of models gives better prediction than that of a single model. The ensemble approach gives superior performance when we combine several decision tree model works together as a base model. The amalgamation of decision tree methods along with the concept of ensemble learning generates a model known as Random Forests. The random forest models was applied to various domains like pattern matching, information retrieval, forecasting of stock indices and medicine because of superior performance. The authors, given an several reasons in the favour of random forest with respect to its performance which are as follows: a) The random forest works very well even though the size of the dataset is very low for the training purpose and also the hypothesis space increases which gives a better fit for the decision tree. b) The construction of the decision tree model is termed as Np-Hard problem, and problem with the decision tree model is that it get stuck in local minima, so to avoid local minima, combination of more than one tree, reduces the risk.

**Construction of Random Forest Algorithm:** The following steps need to be followed for the generation of Random Forests model, first, diversity plays a vital role for better prediction of the model. If diversity present in the dataset is more than each decision tree classifier will get mixture of the dataset and this results in good training of the base classifier. The performance of individual trees matters a lot for the overall accuracy of the Random Forest model. In order to improve the accuracy of the RF model, we need to make sure that each base learner get trained on the subset of data with better diversity. The holistic approach for the manipulation of input data, is to use the data required for the purpose of training with replacement. For the manipulation purpose, the random forest algorithm get trained with different subset of features rather than selecting best feature present in the dataset. This leads to randomness which helps in achieving the good accuracy.

Also, several researcher proposed the concept for increasing the performance of RF algorithm with the help of alteration in parameters. It is found that algorithm tends to give better accuracy when there is a change in the algorithm parameters. Splitting is the initial division that applies to a small teaching process. Another tree was created on each subdivision. There are two obvious ways to split original data: partitioning and partitioning. In
the horizontal section - the original dataset is divided into several datasets with identifiable information from the original dataset, each with a list of original examples. In the vertical section, the vertical dataset is divided into several databases containing the same content as the original dataset, each with a primary identifier. The author proposed a method for the creation of a subspace of attributes by doing the combination of larger base classifiers.

**CONCLUSION**

In this paper, we discussed various ways of construction of decision tree and random forest model and how base classifier accuracy is important for the ensemble model performance. The random forest is to improve the generalization capacity of a single decision tree by combining the production of multiple trees. This overview article highlights the main decisions and techniques used in random forests, in particular the three main factors involved in the use of random forests were explored in detail.

References:


