International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

Random Forest Using R

Dr. G.Kishor Kumar
Rajeev Gandhi Memorial College of Engineering & Technology
Mr. K.Nageswara Reddy
Rajeev Gandhi Memorial College of Engineering & Technology
Mr.R.Raja Kumar
Rajeev Gandhi Memorial College of Engineering & Technology

ABSTRACT
Random forest is an ensemble of unpruned classification or regression trees created by using bootstrapped samples of the training samples of the training data and random feature selection in tree induction. Prediction is made by aggregating the predictions of the ensemble. This work investigates the Random forest for classification of data instances using R.
Keywords(Random Forest, Decision Tree, R)

INTRODUCTION

Random Forest is an ensemble learning based classification and regression technique. It is one of the commonly used predictive modelling and machine learning technique. In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. An error estimate is made for the cases which were not used while building the tree. That is called an OOB (Out-of-bag) error estimate which is mentioned as a percentage.

In a normal decision tree, one decision tree is built and in a random forest algorithm number of decision trees are built during the process. A vote from each of the decision trees is considered in deciding the final class of a case or an object, this is called ensemble process. This is a democratic process. Since, many decision trees are built and used in a process of Random Forest algorithm, it is called a forest.

Now, why is it “random”? A data frame or dataset has two dimensions - observations (or rows) and variables (or columns). For a building a decision tree, samples of a data frame are selected with replacement along with selecting a subset of variables for each of the decision tree. Both sampling of data frame and selection of subset of the variables are done randomly, so first word “random” is arrived.

KEY ADVANTAGES OF USING RANDOM FOREST

- Reduce chances of over-fitting
- Higher model performance or accuracy

Random Forest uses Gini Index based impurity measures for building decision tree. Gini Index is also used for building Classification and Regression Tree (CART). In earlier blogs we have explained working of CART Decision Tree and a worked out example of Gini Index calculation.Random Forest algorithm can be used for both classification and regression applications.

BOOTSTRAP

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

DESCRIPTION OF THE TECHNIQUE

Given a standard training set D of size n, bagging generates m new training sets {D_{i}}, each of size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each {D_{i}}. If n'=n, then for large n the set {D_{i}}is expected to have the fraction (1 - 1/e) (˜63.2%) of the unique examples of D, the rest being duplicates. This kind of sample is known as a bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit. They are two process:

MODEL CONSTRUCTION:Describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction: training set .The model is represented as classification rules, decision trees and mathematical formulas.

MODEL USAGE:for classifying future or unknown objects estimate accuracy of the model The known label of test sample is compared with the classified result from the model Accuracy rate is the percentage of test set samples that are correctly classified by the model Test set is independent of training set, otherwise over-fitting will occurClassification consists of assigning a class label to a set of unclassified cases.

SUPERVISED CLASSIFICATION Set of possible classes is not known. After classification we can try to assign a name to that class. Supervised Classification. The input data, also called the training set, consists of multiple records each having multiple attributes or features.

REGRESSION

Regression is similar to classification except that the targeted attribute's values are numeric, rather than categorical. The order or magnitude of the value is significant in some way.To reuse the credit card example, if you wanted to know what threshold of debt new customers are likely to accumulate on their credit card, you would use a regression model. Once run on the new customers, the regression model will match attribute values with predicted maximum debt levels and assign the predictions to each customer accordingly.This could be used to predict the age of customers with demographic and purchasing data, or to predict the frequency of insurance claims.

RANDOM FOREST TECHNIQUE Random forest is a way of averaging multiple deep decision trees,trained on different parts of the same training set,with the goal of overcoming overfitting problem of individual decision tree.In other words, random forest are an ensemble learning method for classification and regression thatoperate by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

DECISION TREE Decision tree represents a procedure for classifying categorical data on their attributes.

- The construction of decision tree does not require any domain knowledge or parameter setting.
- Decide which attribute (splitting point) to test at node N by determining the "best" way to separate or partition the tuples in D into individual classes.
- The splitting criterion is determined so that, ideally, the resulting partitions at each branch are "pure" as possible.
- Partition is pure if all of the tuples in it belong to the same class.
- To construct a decision tree first we need to find the Information gain for each attribute.
- Information gain is used as an attribute selection measure.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

- Pick the attribute which has highest Information gain.

C4.5 DECISION TREE C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan.C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

IMPLEMENTATIONC4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S={s_{1},s_{2},...} of already classified samples. Each sample s_{i}consists of a p-dimensional vector x_{1,i},x_{2,i},...,x_{p,i}), where the x_{j} represent attribute values or features of the sample, as well as the class in which s_{i} falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists.

INFORMATION GAIN

Information gain is used as an attribute selection measure

$$G \quad (S, A) = E \qquad (s) - \sum_{v} \frac{|s_v|}{|s|} E \qquad (s_v)$$

- The entropy is a measure of the uncertainty associated with a random variable.
- E(S) = - p(P)log2 p(P) – p(N)log2 p(N)

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected

HOW TO FINE TUNE RANDOM FOREST

Two parameters are important in the random forest algorithm:

- Number of trees used in the forest (ntree)
- Number of random variables used in each tree (mtry).

HOW RANDOM FOREST WORKS

Each tree is grown as follows:

1. Random Record Selection: Each tree is trained on roughly 2/3rd of the total training data (exactly 63.2%) . Cases are drawn at random with replacement from the original data. This sample will be the training set for growing the tree.

2. Random Variable Selection: Some predictor variables (say, m) are selected at random out of all the predictor variables and the best split on these m is used to split the node. By default, m is square root of the total number of all predictors for classification. For regression, m is the total number of all predictors divided by 3.The value of m is held constant during the forest growing.

3. For each tree, using the leftover (36.8%) data, calculate the misclassification rate - out of bag (OOB) error rate. Aggregate error from all trees to determine over all OOB error rate for the classification.

4. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes over all the trees in the forest.

VARIABLE IMPORTANCE

- Importance feature in Random Forest is variable importance.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

- Random forests can be used to rank the importance of variable in a regression or classification problem.

SHORTCOMINGS OF RANDOM FOREST

- Random forest are not good at generalizing to the completely new data . For example, if I tell you that 1 chocolate costs $1, 2 chocolates cost $2, and 3 chocolates cost $3, how much do 10 chocolates cost? A linear regression can easily figure this out, while a Random Forest has no way of finding the answer.
- If a variable is a categorical variable with multiple levels, random forests are biased towards the variable having multiple levels.

MEASURES/TOOLS OF RANDOM FOREST

- rpart, rpart.plot package are used for constructing the Decision tree.
- randomForest package is used for constructing Random Forest which contain different number of trees.
- Rweka package is used for constructing c4.5decisiontree.

TO BUILD A DECISION TREE

```
d<-rpart(class.variable~.,data,method="class")
```

- rpart.plot(d , type=2,extra=109)

TO BUILD A RANDOMFOREST

```
r<-randomForest(class.variable~.,data,ntree=10)
```

TO PREDICT THE ERROR RATE

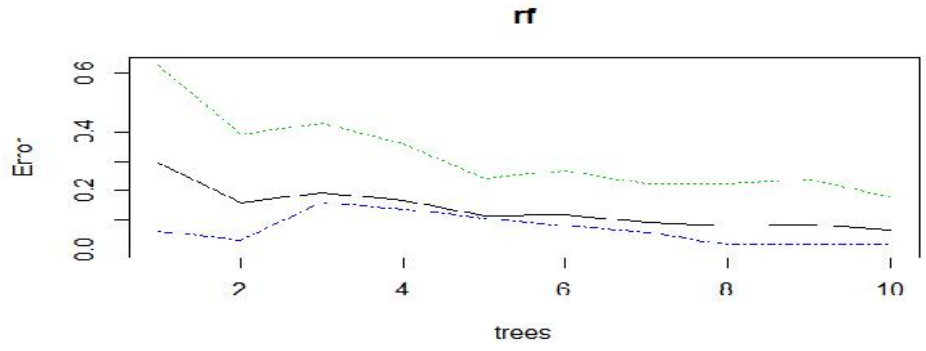
```
P<-predict(r ,testdata)
table(testdata [,5],p)
```

IRIS DATASET

For experimental work we used Iris data set. Iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.The iris dataset (included with R) contains four measurements for 150 flowers representing three species of iris (Iris setosa, versicolor and virginica). On this page there are photos of the three species, and some notes on classification based on sepal area versus petal area.

Error rates for Random Forest

The experiment results over Iris data set is shown in the following figure. In this figure the graph denotes the error rate of Iris data set respect to class label. It is clearly shown that as the no of trees in the Random Forest increases the error rate is decreased. Therefore the Random Forest works better than a single decision tree for classification.



International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

CONCLUSION

Random forest is the best method for classification and regression. Random forest is very effective in eliminating noise in the model input data. Because Random forest builds many trees using a subset of the available input variables and their values, it inherently contains some underlying decision trees that omit the noise generating variables. Random forest solve the over fitting problem. Random forest is best suited classification method for datasets having categorical, real and integer values. The experimental results shown that the Random forest gave better classification results than the C4.5 decision tree.

REFERENCES

[1] http://www.Uci machine repository.com

[2] http://www.w3schools.com

[3] C4.5 Programs for Machine Learning

[4] Data Mining: Han and Kamber