Random Forest

Epoch IIT Hyderabad

Arin Aggarwal
MA22BTECH11006

Contents

1 Introduction

Random Forest is a supervised classification algorithm used for Regression and Classification. It is an extension of Decision Trees involving bagging method which eliminates the overfitting encountered in normal Decision Trees.

2 Bagging

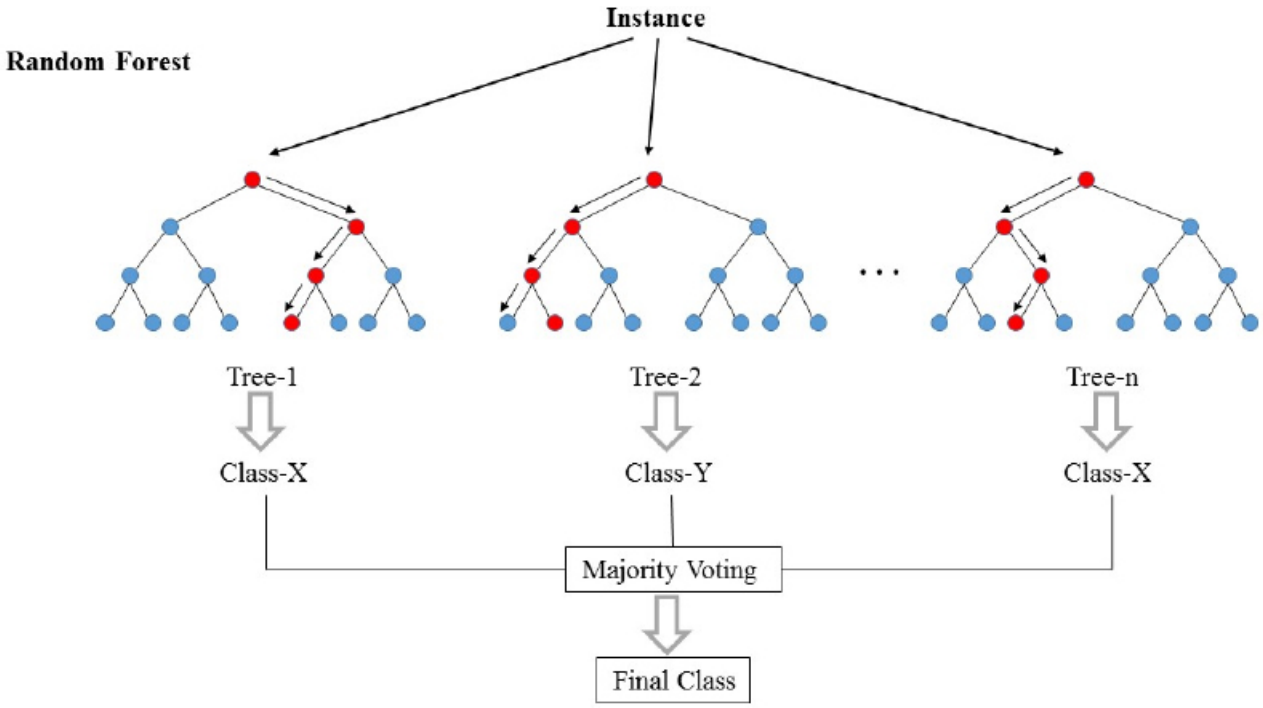
Bagging means algorithm involving **B**ootstrapping and **Agg**regation.

1) **Bootstrapping**: Bootstrapping is used to create random variation among each tree. As the very first step, we sample from the training data with replacement. Each sample has the same data points as the original training data. We build a tree with each sample. Theoretically, sampling with replacement helps keep the variance of Random Forest low and prevents overfitting. The leftover data for each Tree is called Out-of-bag Datasets or **OOB Data**. They usually amount to about 1/3rd of the dataset. Randomly selecting part of features for each node of each Decision Tree. This again increases the random variation between trees and prevents overfitting.

2) **Aggregation**: Once trees are built, we input a data point through each tree, and take a majority vote for classification problems or compute the average of outputs of each tree for regression problems. This is called aggregation.

3 Working

1) Select the data (bootstrapping) to make a decision tree.
2) At every step of the decision tree, decide the features to be used and make the tree.
3) Decide a number and repeat steps 1 and 2 to make that many decision trees.
4) For new given datapoint, Run it through all the trees and the class that it is assigned majorly will be the result (Aggregation).



4 Fine Tuning

- To set the number of features to be chosen in bootstrapping, we usually start with square-root of number of features and work for an optimal value using Cross-Validation.
- What is the optimal number of decision trees to be made? For this, it recommended to make large amount of trees initially (say 1000 trees) and then select the well performing trees out of them. Generally 64-128 is given as a good range for number of trees in the forest.

5 Error Calculation

For Error calculation in Random Forest, we make use of the OOB data. Each data is run through the forest and ratio of wrongly classified data to correctly classified is called OOB Error. This error can be compared for different values of number of trees and number of features used for trees and best values for both can be chosen. The OOB data (effectively used for testing) eliminates splitting of training-testing data and hence the whole data set can be used for training.