Ensemble Learning and Random Forest

medium.datadriveninvestor.com/ensemble-learning-and-random-forest-7430ebf3da7e

January 22, 2019

Photo by Dennis Kummer on Unsplash

This is an adaptive learning methodology to combine several algorithms to extract better results than the individual performances.

The main motto of ensemble methods is as follows:

1. *To decrease the variance(bagging).*
2. *To decrease the bias(boosting)*
3. *To improve predictions(stacking)*

The Ensemble method can be applied in two ways:

1. **Sequential***: In this method, the dependency of base learners are exploited. This involves the evaluation and reweighing of the unimportant examples. e.g. Adaboost.*
2. **Parallel:** *In this method, the independence of the base learners are exploited. This involves the evaluation by simply averaging the outputs of the base learner. e.g. Random forest.*

The ensemble can be of two types :

1. **Homogeneous ensemble:** *In this, we have single base learning algorithm like random forest which only uses the decision tree algorithm*.
2. **Heterogeneous ensemble :** *In this, we have different base estimators algorithms.*

Techniques used in the ensemble learning:

1. **Bagging :** *Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For aggregating the outputs of base learners, bagging uses* 'voting for classification' *and* 'averaging for regression*'.*
2. **Boosting:***Boosting refers to a family of algorithms that are able to convert weak learners to strong learners. The predictions are then combined through a weighted majority vote (classification) or a weighted sum (regression) to produce the final prediction.*
3. **Stacking:** *Stacking is an ensemble learning technique that combines multiple classifications or regression models via a meta-classifier or a meta-regression. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features.*

Random forest Algorithms :

Random Forest algorithm is a supervised learning algorithm. There is a direct relationship between the number of trees in the forest and the results it can get. It uses a number of decision trees and predicts the more accurate result by averaging in case of regression and voting in case of classification.

How Random Forest algorithm works :

1. Randomly select "**K**" features from total "**m**" features where **k << m**
2. Among the "**K**" features, calculate the node "**d**" using the best split point
3. Split the node into n**odes** using the **best split**
4. Repeat the 1 **to 3** steps until "l" number of nodes has been reached
5. Build forest by repeating steps 1 **to 4** for "n" number times to create **"n" number of trees**

In the next stage, with the random forest classifier created, we will make the prediction.

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm

Advantages of Random Forest algorithm :

Compared with other classification techniques, there are three advantages :

1. For applications in classification problems, Random Forest algorithm will avoid the overfitting problem.
2. For both classification and regression task, the same random forest algorithm can be used.
3. The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.