Random Forest

Luis Cabezas Suarez
dept. Electronic Engineering
Hochschule Hamm Lippstadt
luis-david.cabezas-suarez@stud.hshl.de

***Abstract*—Physicians, academics, have traditionally used particular types of regression approaches, Methods that need more difficult mathematics have become more frequent as computing power has become more widely dispersed. Survival analysis has broadened in scope, especially in this era of "big data" and machine learning. The goal of this research is to look into a technique known as Random Forest. The Random Forest technique is a regression tree technique that achieves excellent predicted accuracy by using bootstrap aggregation and randomization of parameters.**

I. Introduction

Before approaching to this well known algorithm. it is considered that it is based on the idea of Breimans's baggin idea. Even though being a great method that can be used for classification as well as regression. also the accuracy can be measured as well. it is a great algorithm where you can combine plenty of tree decisions making some certain of categories or a way of continuing a pattern. [1]. it is basically the idea of having a lot of trees and letting them choose for the class the most important. in this paper we would go in the general idea of this random forest algorithm ans his approach. as well as to understand most of his features in the idea of classification and regression. nevertheless to also explain how fast it goes to determinate errors and how importance some variables can be.

II. Random forest

As mentioned before or in a way that we could understand from his name "forest" is the idea of having a bunch of trees also known as decision trees collecting a lot of variables but each tree depending on the specific variables. the idea came basically by collecting data and creating this decision trees but they have a particular problem and it was that they were not easy when you need it to make new examples in other words not flexible enough to new samples.

This is the part where the random forest take role by using more flexible ways for the decision trees and as a result the random forest gives a a result with more accuracy. Normally you used a regular dataset for the classic decision tree where you have different variables, however for the building of the random forest is meant that we need to use a data called boostrappdataset that consists in the same approach as taking random samples from the data set with the idea that we are free to actually choose the same sample more than once for the purpose of accuracy. then it is really important to classify them by selecting just random subsets of columns at each steps.

III. The concept of classification and regression

Random Forests' trees are based on the binary recursive partitioning trees described in the monograph. These trees use a series of binary partitions ("splits") on specific variables to partition the predictor space. The tree's "root" node encompasses the entire predictor space. The nodes that aren't split are known as "terminal nodes," and they make up the predictor space's last partition. Depending on the value of one of the predictor variables, each nonterminal node splits into two descendant nodes, one on the left and one on the right. [1]

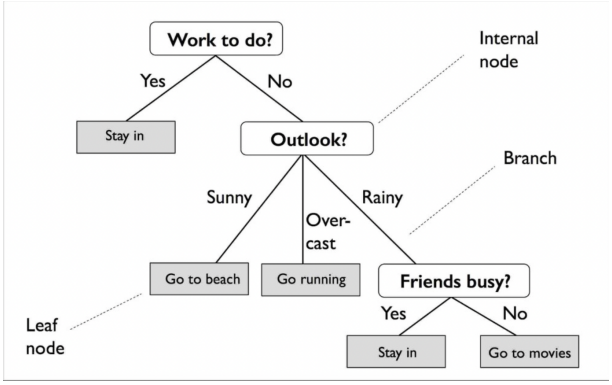

Fig. 1. Classification and Regression analysis with decision trees [5]

IV. Estimate of error by using Out-of-bag data

Tibshirani (1996) and Wolpert and Macready (1996) recommended using out-of-bag estimates in generalization error estimates. Wolpert and Macready worked on regression issues and proposed several approaches for assessing bagged predictors' generalization error. Tibshirani estimated generalization error for arbitrary classifiers using out-of-bag estimations of variance. Breiman (1996b) shows that the out-of-bag estimate is as accurate as using a test set of the same size as the training set in his study of error estimates for bagged classifiers. Using the out-of-bag error estimate eliminates the need for a separate test set. [2]

The idea of this is basically the fact that we you start making the boostrapped dataset since you selection is based randomly

and as well some of the samples are repeating more than once. of course there are some exmaples that are not selected and these are called out of bag data. in short explanation it is the data that was not used to created the bootstrap dataset.therefore with this we could actually measure how accurate it is our random forest by the quantity of out of bags that were correctly classified and the rest that were not classfied are known as the error out of bag.

V. Importance of using Random forest

Random forests are a useful technique for forecasting. They do not overfit due to the Law of Large Numbers. They become accurate classifiers and regressors when the correct kind of randomness is injected. Furthermore, the framework provides insight into the random forest's ability to forecast in terms of the strength of individual predictors and their associations. The use of out-of-bag estimate makes theoretical strength and correlation values concrete. [2]

For a long time, it was assumed that forests couldn't compete with arcing algorithms in terms of accuracy. Our data disprove this notion, yet lead to interesting concerns. Boosting and arcing algorithms can help reduce both bias and variation (Schapire et al., 1998). The adaptive bagging approach (Breiman, 1999) in regression was created to eliminate bias and works well in both classification and regression. It does, however, vary the training set as it progresses, similar to arcing. Forests produce comparable outcomes to boosting and adaptive bagging, but do not update the training set in a progressive manner. Their accuracy suggests that they work to eliminate bias. [2]

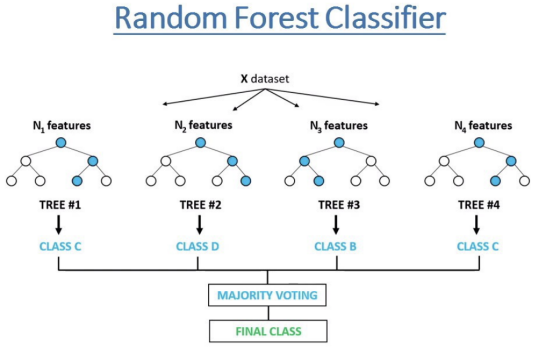

Fig. 2. Random Forest classifier [6]

VI. Conclusion

References

[1] Cha Zhang • Yunqian Ma Editors Ensemble Machine Learning Methods and Applications

[2] Random Forests LEO BREIMAN Machine Learning, 45, 5–32, 2001 Kluwer Academic Publishers. Manufactured in The Netherlands. Statistics Department, University of California, Berkeley, CA 94720 Editor: Robert E. Schapire

[3] Multivariate random forests Mark Segal and Yuanyuan Xiao

[4] Random Forests and Decision Trees Jehad Ali1, Rehanullah Khan2, Nasir Ahmad3, Imran Maqsood4 1 Computer Systems Engineering, UET Peshawar, Pakistan 2 Sarhad University of Science and Information Technology, Peshawar, Pakistan

[5] https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

[6] https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6