Bagging, Random Forest , Boosting, and two combined methods.

Yuwei Zheng

I. Abstract

This report provides an overlook on three well-known tree decision methods, namely: Bagging, Random Forest, and Boosting, and two recently suggested methods called the Boosted Random Forest and One-Step Boosted Forest introduced by Y. Mishina, M. Tsuchiya, and H. Fujiyoshi, and I. Ghosal, and G. Hooker.

II. Introduction

The decision tree is a well-used method for data analysis, especially for capturing complicated interactions. Unfortunately, instability and high variance are downsides associated with it. Bagging, Random Forest, and Boosting are methods created to improve its stability and reduce uncertainty.

III. Bagging

Bagging is an abbreviation of bootstrap aggregation, a procedure for reducing the variance of decision trees. Given a set of data points $X = (x_1, x_2, ..., x_n)$, B subsets each contains k data points picked from the training data set X_{tr}, we conduct following steps:

1) Using bootstrap method to pick up k data points with replacement.
2) Repeat previous step for B times, so we have B subsets.
3) For each subset created, a tree model is generated. Therefore, we will have B trees in total.
4) Lastly, we average all B trees to obtain the mean coefficient and a reduced variance:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f^b(x)$$

$$\hat{\beta} = \frac{1}{B}\sum_{b=1}^{B} \beta$$

$$\hat{\sigma}^2(\beta) = \frac{\sigma^2(\beta)}{B}$$

Bagging has many advantages, it is better than a piece-wise constant plus a forward step-wising method as it can reduce uncertainty. However, as we can see from the diagram shown below, the splits of different trees are similar to each other. This is because that k data points of each tree are highly correlated to another tree and trees are generated by entire variables, thus it can be improved.

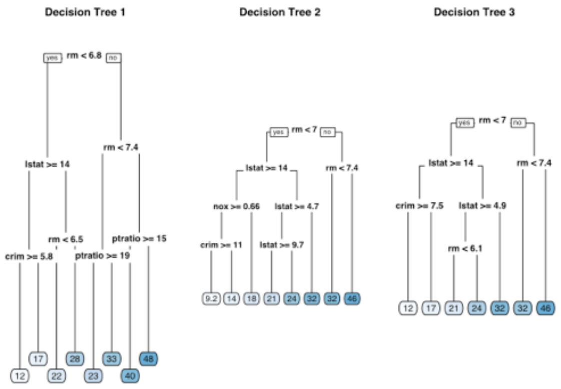

Fig. 1. Each decision tree is been generated by one model with one subset of training data.

IV. Random Forest

Random Forest can be seen as an extended form of bagging. Apart from picking treatment data with replacement. For each tree generated, only partial variables are used: selecting m features randomly from p features $m < p$. Because the features included in each model are different, trees are less correlated. Then, repeat bagging procedures and average all trees.

Random Forest improves the bagging method by using random features to mitigate correlations between trees, which therefore reduces uncertainty. Much empirical evidence has already proved that the Random Forest is a better method.

Bagging and Random Forest are two ensemble methods, many predictions are combined, which requires a lot of memory for calculation.

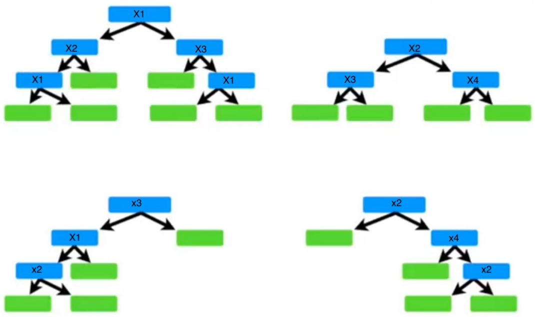

Fig. 2. Decision trees are modelled with different features so they are less correlated.

V. Boosting

Boosting is a weak learner method (Bagging and Random Forests are strong learners); it fits a simple model to the training data X_{tr}sequentially to construct a more complex model. planting a tree based on the residual of the previous one.

1) Plant tree T_1 with all treatment data points X_{tr}.
2) Set initial weight distribution to be uniform : $w_i = \frac{1}{N}$
3) Find residual of T_1 and increases the weight w_i for the misclassified data points.
4) Replant a tree T_2 with re-weighted X'_{tr}data points.
5) Repeat step 1) and 2) until meeting a pre-designed limit.
6) Integrate all trees. [$(y_i, G(x_i))$ captures the residual]:

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(y_i, G(x_i))$$

While adding all trees at the last step, Boosting gives higher weight to more accurate models and lower weight to less accurate models thus gradually improves the model. It is an efficient method as it increases the accuracy of the prediction without using a large memory. However, since the second tree is planted upon the first one, boosting can take a long time to complete.

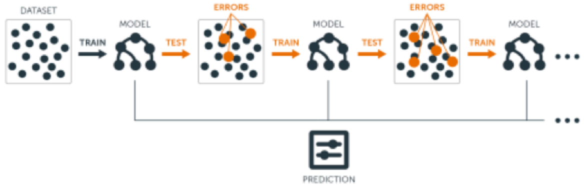

Fig. 3. Each new tree is planted based on the residual of the previous one.

VI. Combining Random Forest and Boosting Methods

Two groups of scholars have introduced two different methods both contain a combination of Random Forest and Boosting procedures. Boosted Random Forest, introduced by Mishina, Tsuchiya, and Fujiyoshi in 2014, is a model boosting each randomly planted decision tree. Whereas, One-Step Boosted Forest, introduced by Ghosal and Hookercan in 2018, is a model summarising two Random Forests where the second one is formed based on the residual of the first one. Further clarification will be demonstrated below:

A. Boosted Random Forest

1) Initialise sample weight $w_i = \frac{1}{N}$
2) Create a subset of k training data X_{tr} with replacement.
3) Choosing m features randomly.
4) Plant tree T_1 with m features, calculate it's residual and update the weight w_i.
5) Repeat step 2), 3), and replant tree T_2.
6) Finally combine all trees T_i with accuracy based weights.

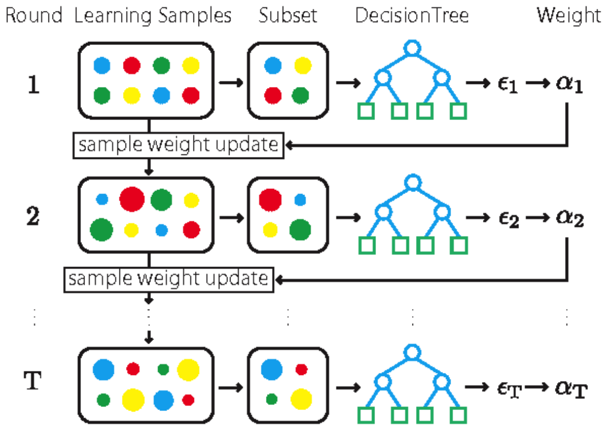
Round
Learning Samples
Subset
DecisionTree
Weight
1
$\epsilon_1 \rightarrow \alpha_1$
sample weight update
2
$\epsilon_2 \rightarrow \alpha_2$
sample weight update
T
$\epsilon_T \rightarrow \alpha_T$

Fig. 4. Residuals are calculated for each decision tree with different feature sets m.

B. One-Step Boosted Forest

1) Plant the first Random Forest (same procedures as IV).
2) Calculate the first estimation (T^b is the tree function):

$$\hat{f}(x_1) = \sum_{b=1}^{B} T^b(x)$$

3) Calculate it's residuals and create a new data set X'_{tr} based on the updated weights:

$$\epsilon_i = Y_i - \hat{f}(x_1)$$

$$X'_i = (e_i, X_i)_{i=1}^{n}$$

4) Plant another Random Forest and calculate the second estimation $\hat{f}(x_2)$.
5) Finally combine both estimations:

$$\hat{f}(x) = \hat{f}(x_1) + \hat{f}(x_2)$$

Ghosal and Hookercan later also compared Boosted Forest with Out-of bag residuals, In-bag residuals, and Bootstrapped resamples.

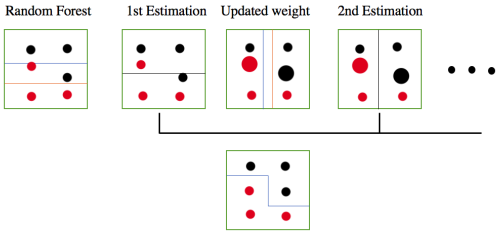

Fig. 5. Random Forest is modeled based on the residual of the previous one

Both groups of researchers have proved empirically that a model which combines Boosting and Random Forest methods would perform better, in terms of stability and accuracy, than only one of them. A combined model can also save at least 47% memory (Mishina, Tsuchiya, and Fujiyoshi,2014). Ghosal and Hookercan also proved that a Boosted Forest constructed with out-of bag residuals perform the best.

VII. Conclusion

This report includes 3 classical decision tree models: Bagging, Random Forest, and Boosting, and 2 modified models combined Random Forest and Boosting methods in different ways. It would be interesting to justify those 2 methods with more real data, although they have not been packaged in R yet.

[1] Y. Mishina, M.Tsuchiya, H.Fujiyoshi, 2014. Boosted Random Forest.

[2] I. Ghosal, G. Hooker, 2018. Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and its Varaicne Estimate.

[3]UC-r Github. 2019. Regression. [Online] Available at: https://uc-r.github.io/regression-trees. [Accessed 5 May 2019].

[4]UC-r Github. 2019. Gradient Boosting Machines. [ONLINE] Available at: https://uc-r.github.io/gbm-regression. [Accessed 5 May 2019].

[5]CNblog. 2019. Distinguish Boosting and Bagging. [ONLINE] Available at: https://www.cnblogs.com/dudumiaomiao/p/6361777.html. [Accessed 5 May 2019].