Report 2: Assess Learners

Fanrui Yan
yanfanrui@gatech.edu

***Abstract*—**In this report, I would analyze relationship of overfitting and leaf size of decision tree, as well as the effect of bagging to overfitting. I will also briefly touch the difference of decision tree and random tree.

1 INTRODUCTION

I would analyze overfitting of decision tree with different number of leafs and different number of bags. The purpose is to see the effect of leaf size and bagging to overfitting. Smaller leaf size tends to lead overfitting and bagging could eliminate over fitting.

I have also briefly analyzed the difference of decision tree and random trees from speed and MAE. Random tree is much faster than decision tree in speed, and single decision tree performs about the same as single random tree, but random trees performs better than decision tree when apply bagging.

2 METHODS

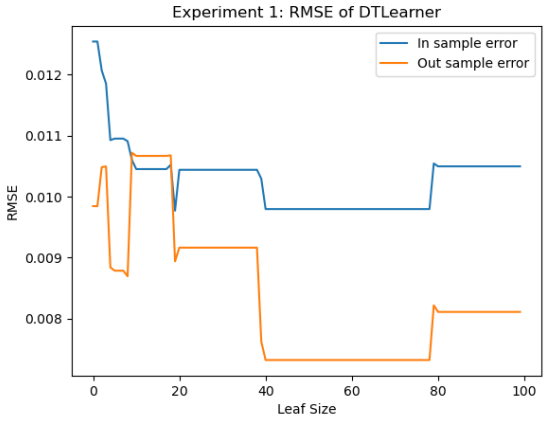
I did 3 experiments to complete the analysis, first one is to see how the root mean squared error (RMSE) changes while leaf size changes.Second one is to introduce bagging technique and see how the root mean squared error (RMSE) changes while bag number changes. The third experiment contains two parts, one part is to record the time for decision tree and random tree to perform a task and compare the times used (speed); another part is to compare their mean absolute error (MAE).

3 DISCUSSION

3.1 Experiment 1

Question: Research and discuss overfitting as observed in the experiment. (Use the dataset Istanbul.csv with DTLearner). Support your assertion with graphs/charts. (Do not use bagging in Experiment 1). Does overfitting occur with respect to leaf size? For which values of leaf size does overfitting occur? Indicate the starting point and the direction of overfitting. Support your answer in the discussion or analysis. Use RMSE as your metric for assessing overfitting.

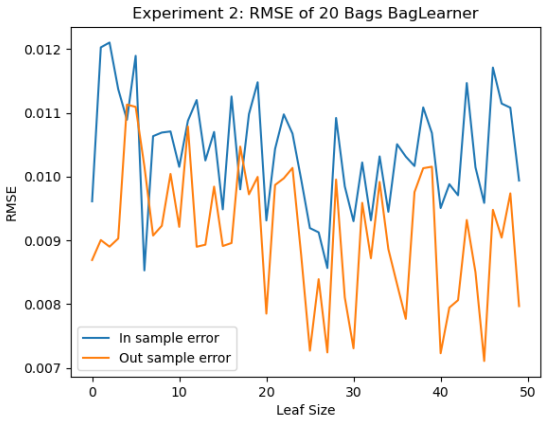
Answer: Yes, overfitting occurs with respect to leaf size. The smaller the leaf size, the more likely to overfit. When leaf size is larger than 5 there seems to be no overfitting.



3.2 Experiment 2

Question: Research and discuss the use of bagging and its effect on overfitting. (Again, use the dataset Istanbul.csv with DTLearner.) Provide charts to validate your conclusions. Use RMSE as your metric. Can bagging reduce overfitting with respect to leaf size? Can bagging eliminate overfitting with respect to leaf size?

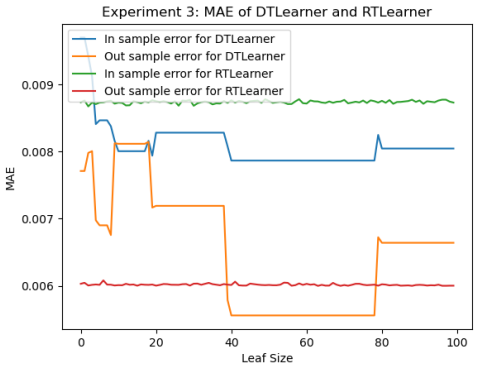
Answer: Yes, bagging should reduce or even eliminate overfitting. The lines are more towards the middle.

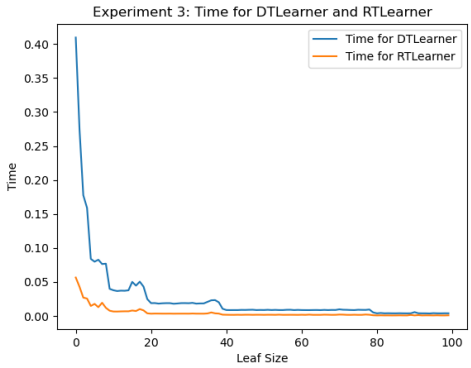


3.3 Experiment 3

Question: Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). For this part of the report, you must conduct new experiments; do not use the results of Experiment 1. Importantly, RMSE and correlation are not allowed as metrics for this experiment. Provide at least two new quantitative measures in the comparison. Provide charts to support your conclusions. In which ways is one method better than the other? Which learner had better performance (based on your selected measures) and why do you think that was the case? Is one learner likely to always be superior to another (why or why not)?

Answer: I'm comparing decision tree and random trees from 2 points of views - speed and MAE. Random tree is faster than decision tree to train, random tree is better from this point. MAE tells us the difference between predicted value and actual value. Random tree has higher MAE than decision tree which means decision tree is more accurate than random tree.





4 SUMMARY

In conclusion, smaller leaf size tends to lead overfitting and bagging could eliminate over fitting. Random tree is faster than decision tree, but single decision tree is more accurate than random tree.