Project 3 - Assess Learners Report

GTID Student Name: Zhiyong Zhang

GT User ID: zzhang726

GT ID: 903370141

To evaluate whether overfitting occur with respect to leaf_size. The experiments are run with leaf_size from 1 to 60 using DTLearner and BagLearner with dataset istanbul.csv. RMSE is used as a metric for assessing overfitting. The results are plotted in Figure 1.

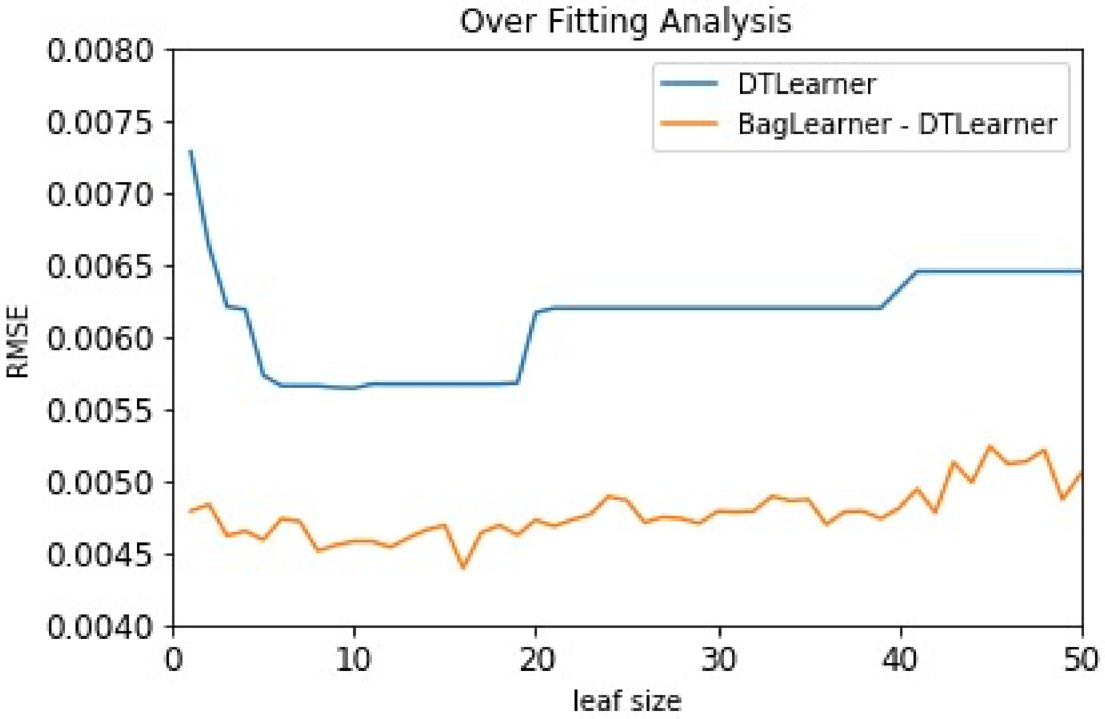

Figure 1

For DTLearner, RMSE would reduce when leaf_size is decreased from 50 to 20, then stays almost flat when leaf_size is further reduced to 5, then increases sharply if leaf_size is further reduced to 1. It shows overfitting occur when leaf_size is less then 5.

For BagLearner, bagging is used with bagsize of 20. Bagging would reduce overfitting significantly as it has much smaller RMSE compared to DTLearner as shown in Figure 1. And as leaf_size decreases, there is no obvious overfiiting.

To compare "classic" decision trees (DTLearner) versus random trees (RTLearner), The experiments are run with leaf_size from 1 to 60 using DTLearner and RTLearner with dataset istanbul.csv. Correlation is used as a metric for assessing overfitting. The results are plotted in Figure 2. DTLearner has more stable correlation and has better correlation when leaf_size is smaller than 20 as shown in Figure 2.

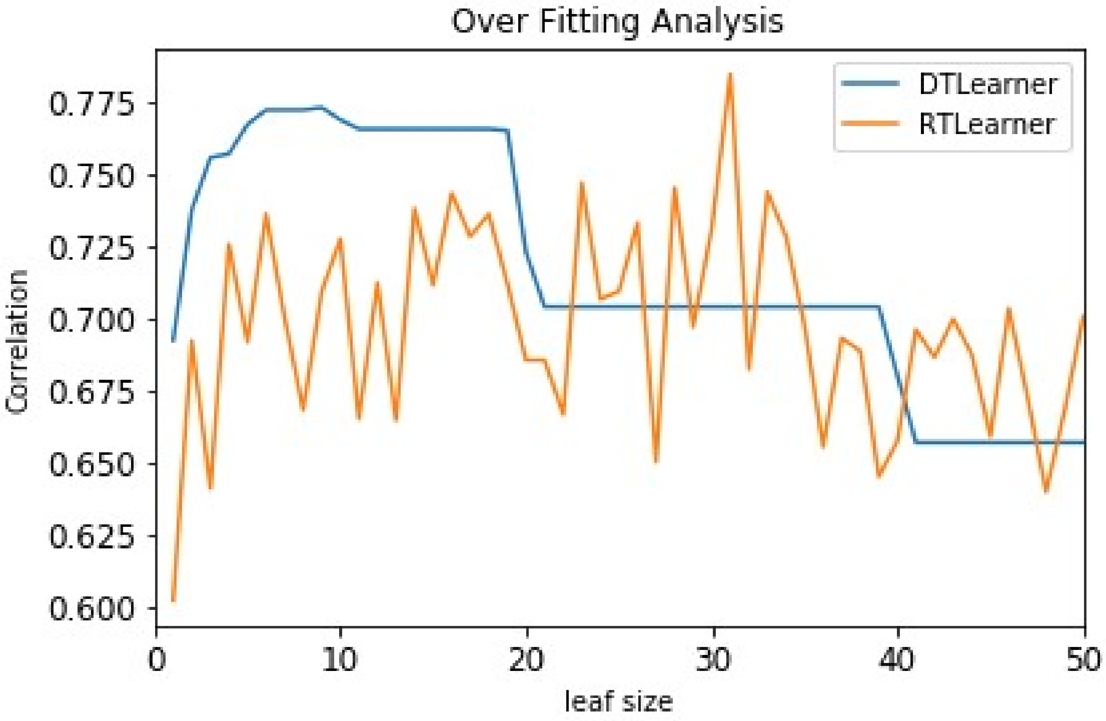

Figure 2.

The performance of training and querying measured as time elapsed using DTLearner and RTLearner with leaf_size from 1 to 60 are plotted in Figures 3-4. The DTLearner takes longer to train as it needs to select a variable to split. However, the query time are pretty similar to both learners.

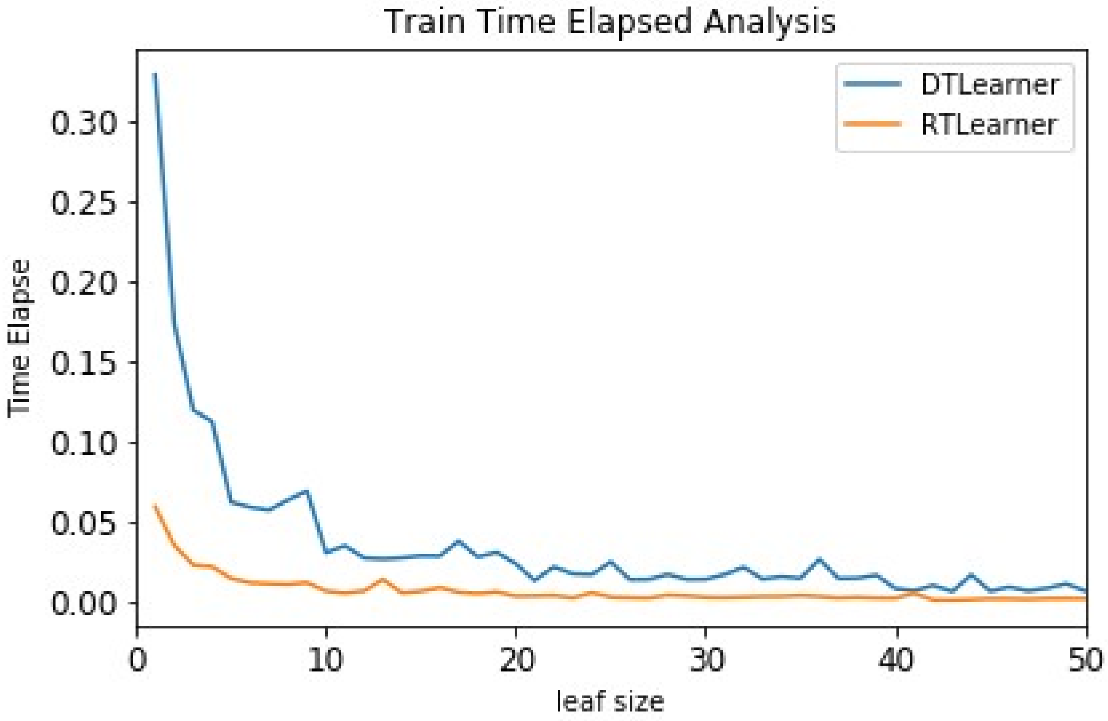

Figure 3

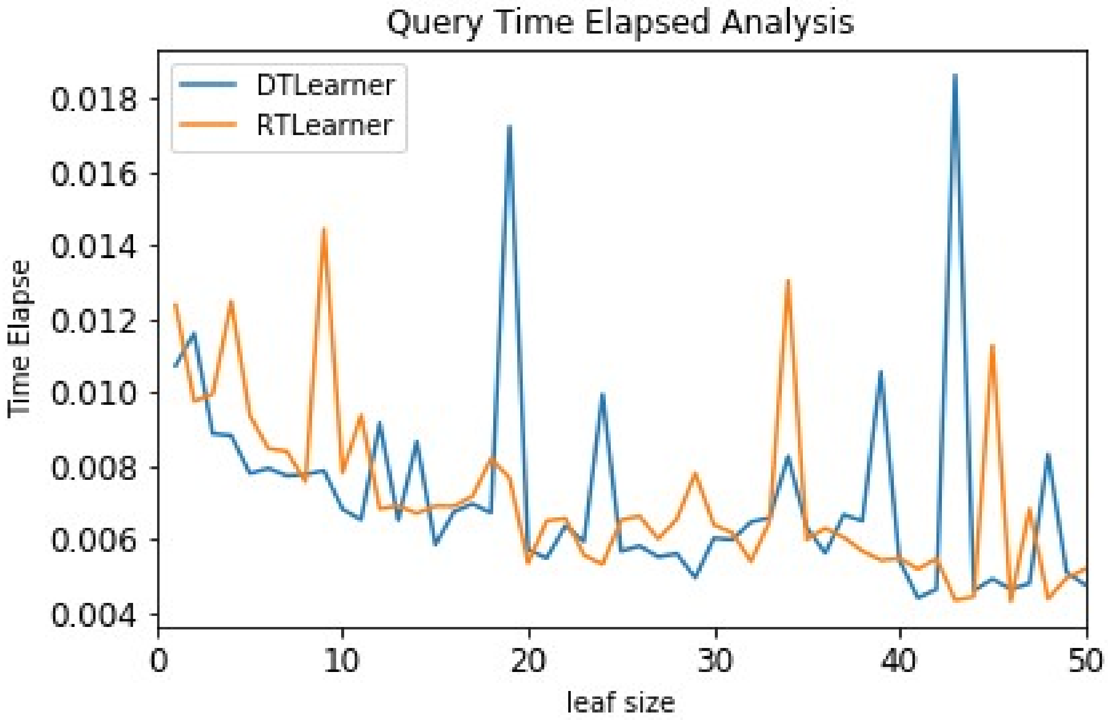

Figure 4