Report: Project 3

CS7646 Machine Learning for Trading, 2019 Spring

Hui Xia (hxia40)
903459648
Georgia Institute of Technology

1. **Question 1**: Does overfitting occur with respect to leaf_size? Use the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).

1.1. Experimental Methods

Prior to other handling steps, the data in `Istanbul.csv` was randomly shuffled once (random seed is 123) for improved analysis result. Then, a Decision Tree Learner (DTL) was trained using the first 60% of the data in `Istanbul.csv` and tested using the other 40% of the data.

Among the *n* columns of the data, the last column is referred as the *Y* column, which represents the result that to be decided by the rest of the columns (i.e. the X_1, X_2 … X_{n-1} columns). The *Y* column that was generated by passing the *X* columns through the DTL was compared with their original counterparts before DTL training, and then used to calculate the in-sample (using the training data) and out-of-sample (using the testing data) Root Mean Square Error (RMSE) values.

The in- and out-of-sample RMSE values were calculated (by code) when the DTL use a range of leaf size values (1-100). The RMSE values were then plotted against the leaf size vales.

1.2. Result

The plotted data is demonstrated in **Figure 1**. In general, when the value of leaf size is larger than 5, both in- and out-of-sample RMSE show an ascending trend when the leaf size increases. However, when the leaf size is lower than 5, the out-of-sample RMSE decreases when the leaf size increases, as shown in **Figure1a**. On the other hand, the in-sample RMSE increases in this range. When leaf size is smaller than 20, the in-sample RMSE is lower than the out-of-sample RMSE. When leaf size is larger than 80, the in-sample RMSE is higher than the out-of-sample RMSE. When leaf size is in between 20 and 80, the in-sample RMSE and the out-of-sample RMSE have similar value.

1.3. Discussion

In this experiment, the data was randomly shuffled once before it was used to train the DTL. However, while handling financial data, this 'shuffle before training' method in in general being suggested against. This is because that the data used for training should always be generated earlier than the data used for testing, to avoid the 'predict the future' effect.

In general, the RMSE value (for both in- and out-of-sample) increases when the leaf size increases. This is expected, as for a certain data set, training using larger leaf size will result in a 'rougher' decision tree, and thus more inaccurate prediction. The trend deviation between the in- and out-of-sample RMSE values suggest that when the leaf size is too small (<5 in this case), the DTL describe the samples in an overly detailed manner. Thus, the DTL overfits the data when leaf size < 5.

It is also noticeable that when the leaf size is using a larger value (> 80 in this experiment, as demonstrated in **Figure 1d**) the RMSE of out-of-sample become lower compared with the in-sample. Although this case is not expected, as the training data should fit better with the DTL compared with other data. However, it is reasonable to consider that a certain number of leaves should be needed to generate enough 'cases' to the learner to study, and thus to keep the DTL reliable. As we have discussed above, if the leaf size increases while the size of the training data cannot change accordingly (and this is a common scenario, as data is precious), the generated number of leaves will decrease accordingly. For an extreme case, if the leaf size equal to the size of the data, the trained DTL will only have one leaf, and is not necessarily useful anymore. Thus, there should be a trade-off between leaf size and leaf numbers exist, which defines the upper-limit of the leaf size.

1.4. Conclusion

When the leaf size is too small (<5 in this experiment), over fitting does occur in the DTL.

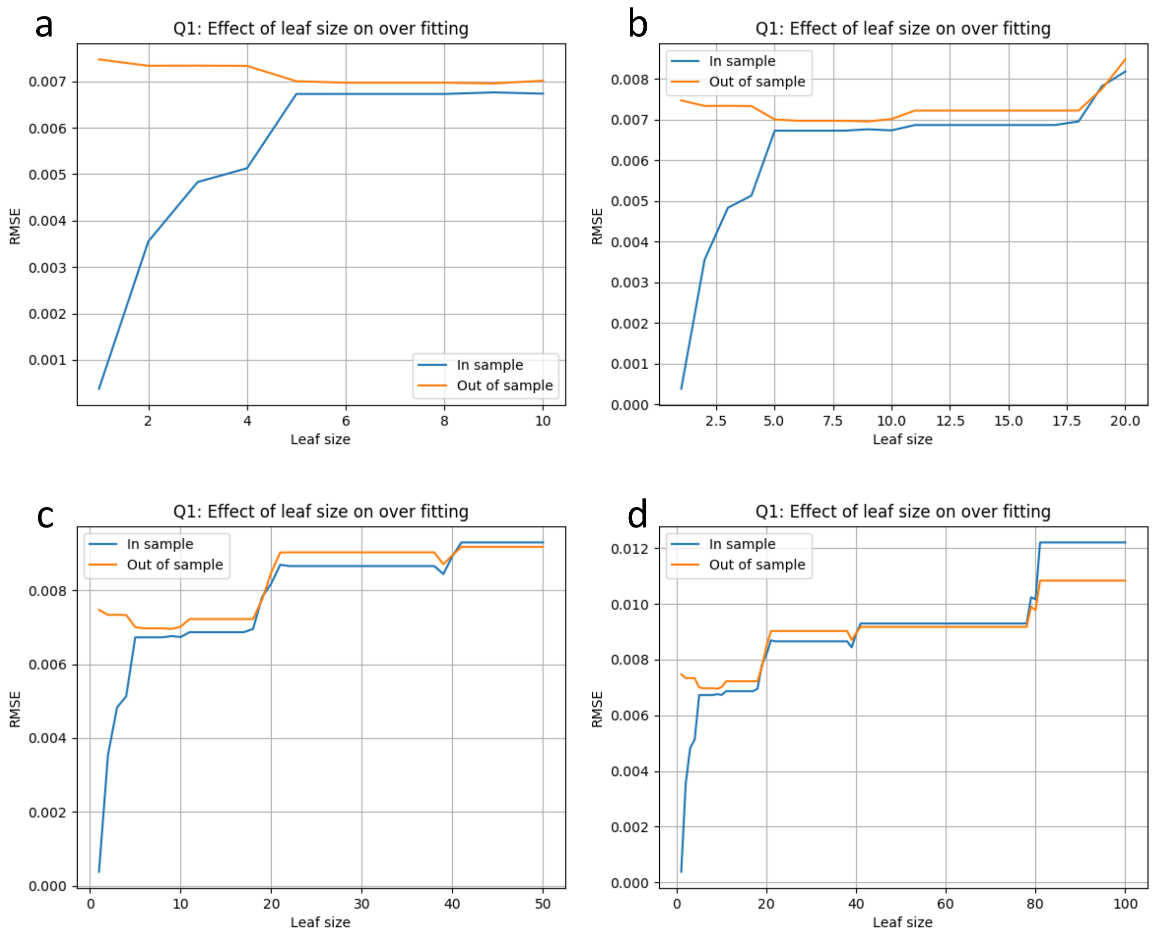

Figure 1: Plotting in sample and out of sample RMSE values versus the leaf size using DTL. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100.

2. **Question 2**: Can bagging reduce or eliminate overfitting with respect to leaf_size? Again use the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.

2.1. Experimental Methods

Prior to other handling steps, the data in `Istanbul.csv` was randomly shuffled once as described in the section 1.1. Then, a Bagging Learner (BL) that contains 20

DTL was trained using the first 60% of the data in `Istanbul.csv` and tested using the other 40% of the data. To obtain smoother lines and analyze the data more accurately, another set of experiment is performed, which performs 20 iterations under each leaf size.

Similar as in section 1.1, for both experiments (i.e. one with single iteration per leaf size, another with 20 iterations per leaf size), the in- and out-of-sample RMSE values under a range of leaf size (1-100) were calculated, which were then plotted versus the leaf size.

2.2. Result

The in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL, with single iteration per leaf size is shown in **Figure 2**, and the figure with 20 iterations per leaf size is shown in **Figure 3**.

Both **Figure 2** and **Figure 3**, suggest that the in- and out-of-sample RMSE have an ascending trend when the leaf size increases. Unlike the DTL shown in **Figure 1**, no over fitting is observed.

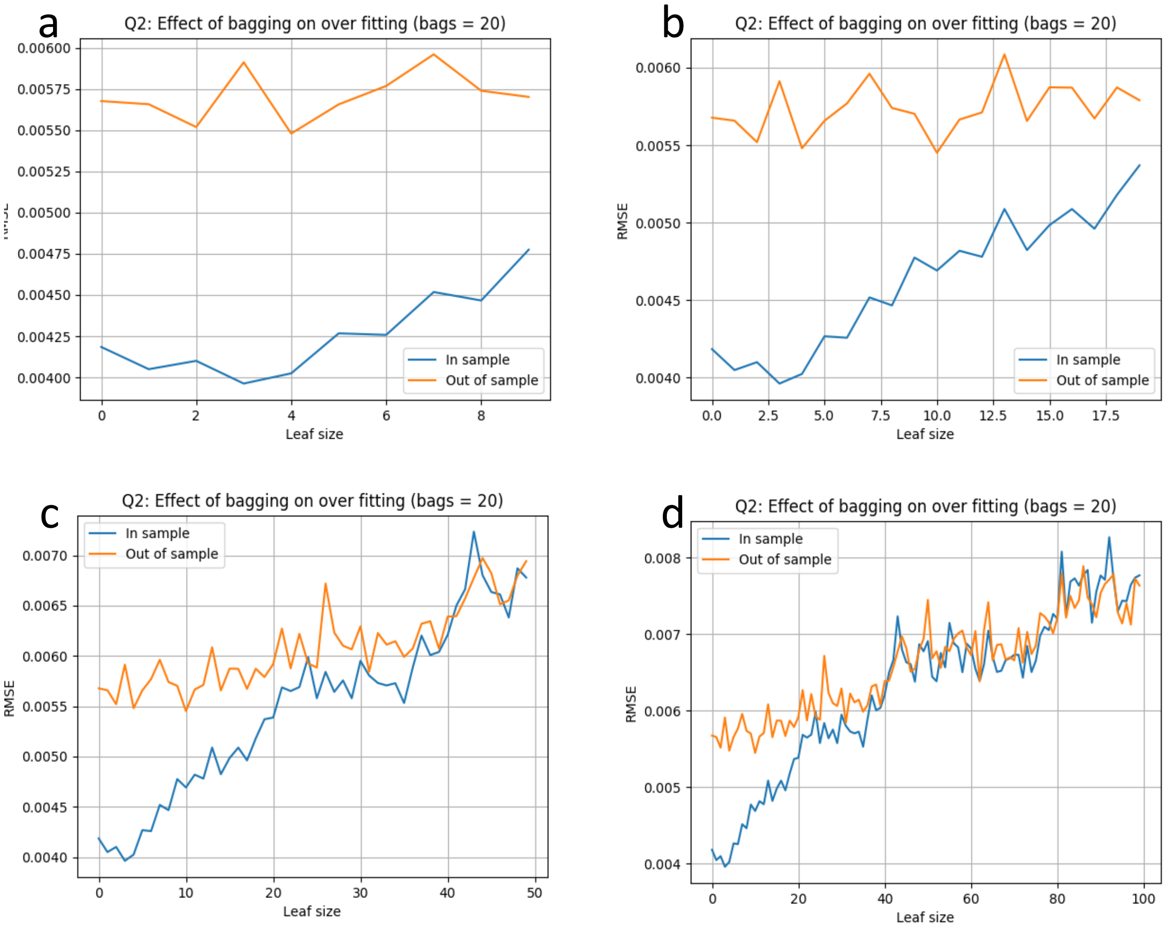

Figure 2: Plotting in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100.

Compared with the DTL, the RMSE values in the BL is in general (when the leaf size is larger than 1) lower. When leaf size = 100, the in- and out-of-sample RMSE generated by the BL are between 0.007 and 0.008, while the counterpart that were generated by the DTL are above 0.01. However, the in-sample RMSE generated by DTL is lower than 0.001, which is lower than the counterpart generated by the BL.

2.3. Discussion

Similar to the DTL experiment, larger leaf size in the BL will result in a higher RMSE value, for both the in- and out-of-sample data. Compared with the DTL, using the same set of data, the BL nearly always grant a lower RMSE. The only scenario that the DTL has a lower RMSE value happens when the leaf size is 1, for the in-sample data. However, this scenario could be considered as useless, as it basically means to 'predicting' the data while the data analyzer already has it.

In both experiment performed using the BL (i.e. one iteration per leaf size and 20 iterations per leaf size), no overfitting could be observed.

2.4. Conclusion

Bagging does greatly reduce overfitting, if not completely eliminates it.

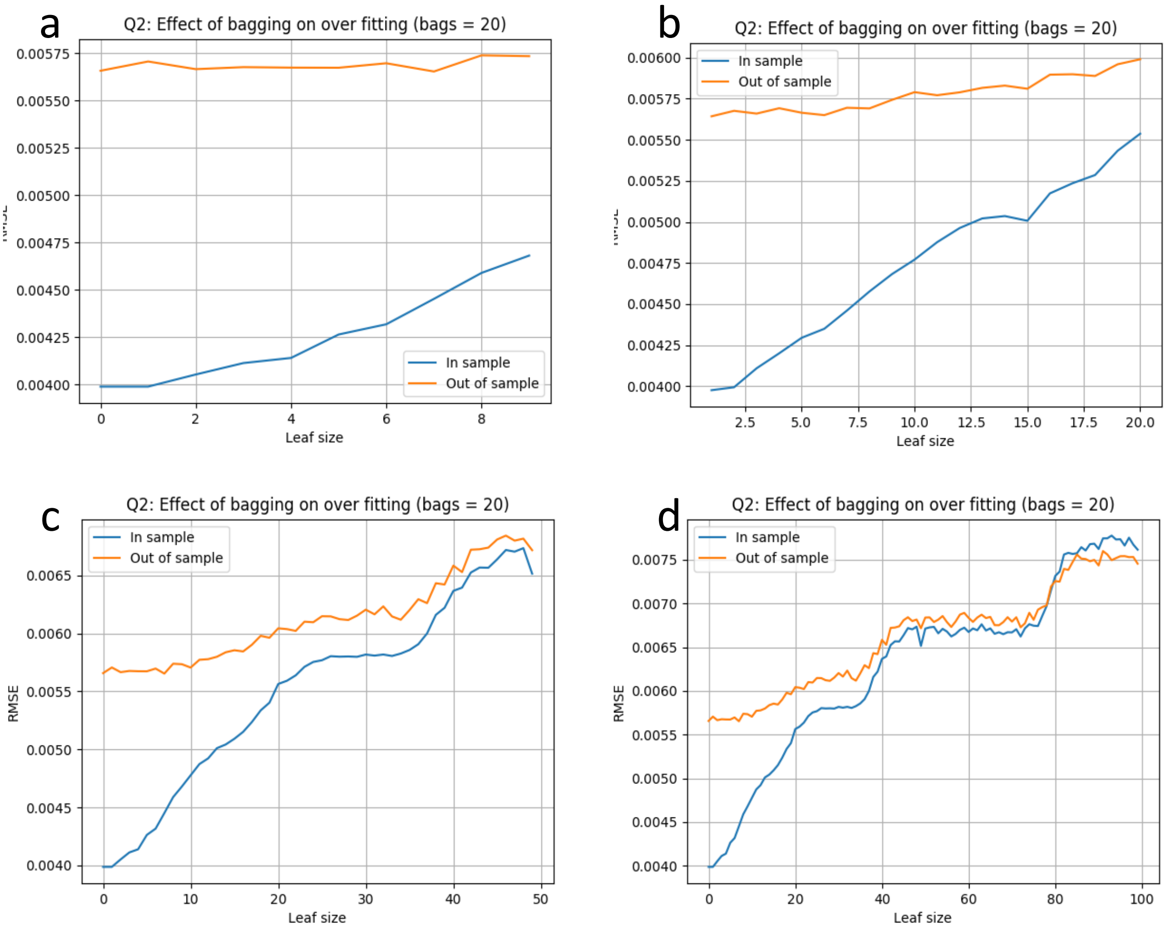

Figure 3: Plotting in-sample and out-of-sample RMSE values versus the leaf size using BL that contains 20 DTL, with 20 iterations for each leaf size. (a). leaf size between 1 and 10. (b). leaf size between 1 and 20. (c). leaf size between 1 and 50. (d). leaf size between 1 and 100

3. **Question 3**: Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, do not use two measures for accuracy.) Note for this part of the report you must conduct new experiments, don't use the results of the experiments above for this.

3.1. Experimental Methods

Prior to other handling steps, the data in `Istanbul.csv` was randomly shuffled once as described in the section 1.1. the DTL and Random Tree Learner (RTL) were trained using the first 60% of the data in `Istanbul.csv` and tested using the other 40% of the data.

For both DTL and RTL, the in- and out-of-sample Mean Absolute Error (MSE) and learner running time were calculated under a range of leaf size (1-100), which were then plotted versus the leaf size. MAE is a measure of difference between two continuous variables, which is calculated using the equation below:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |Y_{pred} - Y_{given}|}{n}$$

Where Y_{pred} and Y_{given} stand for the predicted *Y* values and the given *Y* values, respectively, and *n* is the total numbers of *Y* values. The MAE and learner running time were then respectively plotted versus the leaf size.

3.2. Result

Overfitting happens when using DTL with a smaller leaf size (leaf size < 5, **Figure 4a-b**), while it does not happen when using RTL (**Figure 4c-d**).

The MAE value generated by DTL is smaller compared with the ones generated by RTL. For example, at leaf size = 50, the in- and out-of-sample MAE generated by DTL are about 0.007, and the MAE generated by RTL are about 0.008. Moreover, the MAE generated by the RTL is more unstable (noisy) compared with their DTL counterparts.

As shown in **Figure 5**, for the same data set and all leaf sizes, RTL requires less running time compared with DTL. Especially, when the leaf size is small (<20 in this experiment), the time efficiency advantage of the RTL is more predominant. For example, DTL need 0.05 seconds to perform one run, but the time need by RTL is less than 0.01 seconds.

3.3. Discussion

Similar to the DTL, larger leaf size in the RTL will sacrifice accuracy (result in a higher MAE value), for both the in- and out-of-sample data. Compared with the DTL, using the same set of data, the RTL nearly always result a (about 10 -20%) higher MAE, or, is less accurate. Moreover, the MAE calculated from RTL predictions are less stable, which is originated from the internal randomness of the RTL.

Sacrificing the accuracy and prediction stability, RTL obtained an advantage of not overfitting, and performs better on running time compared with DTL. This advantage would be especially beneficial when performing machine learning over big data.

However, in the cases where the data source is limited, the accuracy advantage of DTL will grant it more preference compared with RTL.

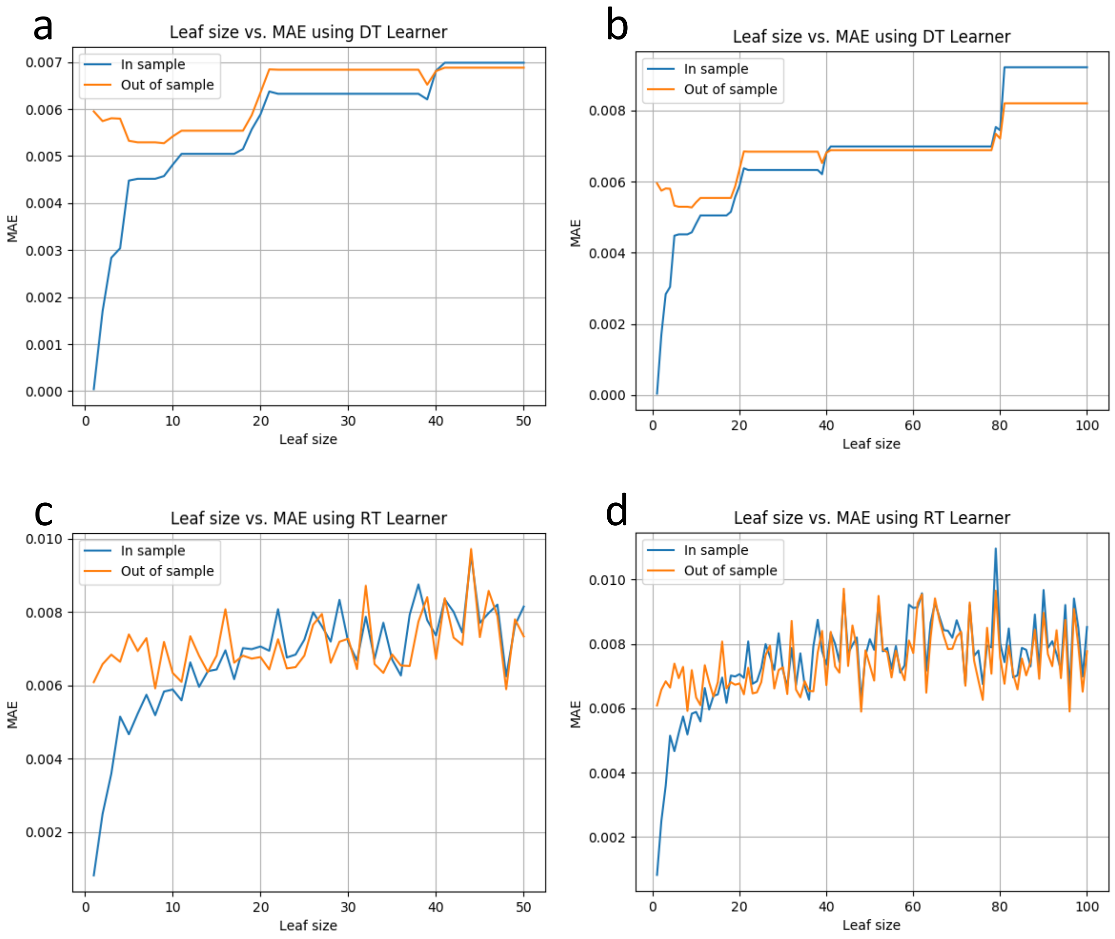

Figure 4: Plotting in-sample and out-of-sample MAE values versus the leaf size using DTL and RTL. (a). leaf size between 1 and 50 using DTL. (b). leaf size between 1 and 100 using DTL. (c). leaf size between 1 and 50 using RTL. (d). leaf size between 1 and 100 using RTL.

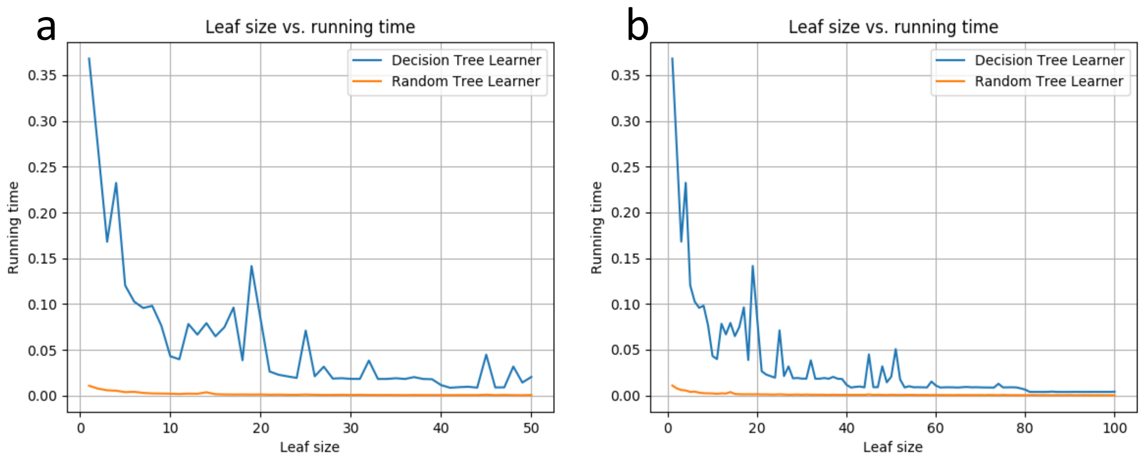

Figure 5: Plotting learner running time versus the leaf size using DTL and RTL. (a). leaf size between 1 and 50 (b). leaf size between 1 and 100.

3.4. Conclusion

Both RTL and DTL have their advantages. RTL is advantageous for it does not overfit, and quicker than DTL in running time. DTL is advantageous as it grants a more accurate prediction compared RTL (using same dataset and leaf size). The data analyzer need to apply them depend on other requirements (e.g. size of the dataset, running time requirement, leaf size requirement, and accuracy requirement).