CS573 – HW4

Wen-Ling Chi

Email: chi64@purdue.edu

1 Preprocessing

2. Implement Decision Trees, Bagging and Random Forests

Decision Tree

Training Accuracy: 0.78

Testing Accuracy: 0.70

Bagging

Training Accuracy: 0.80

Testing Accuracy: 0.75

Random Forest

Training Accuracy: 0.73

Testing Accuracy: 0.70

```
C:\Users\wenne\OneDrive\□ □ \course\data_mining\hw\hmw4>python trees.py trainingSet.csv testSet.csv 1
Training Accuracy DT: 0.78
Testing Accuracy DT: 0.70

C:\Users\wenne\OneDrive\□ □ \course\data_mining\hw\hmw4>python trees.py trainingSet.csv testSet.csv 2
Training Accuracy BT: 0.80
Testing Accuracy BT: 0.75

C:\Users\wenne\OneDrive\□ □ \course\data_mining\hw\hmw4>python trees.py trainingSet.csv testSet.csv 3
Training Accuracy RT: 0.73
Testing Accuracy RT: 0.70
```

3 The Influence of Tree Depth on Classifier Performance

(a) Draw a plot to compare decision tree, bagging, random forest with 10-fold cv and with different tree depth

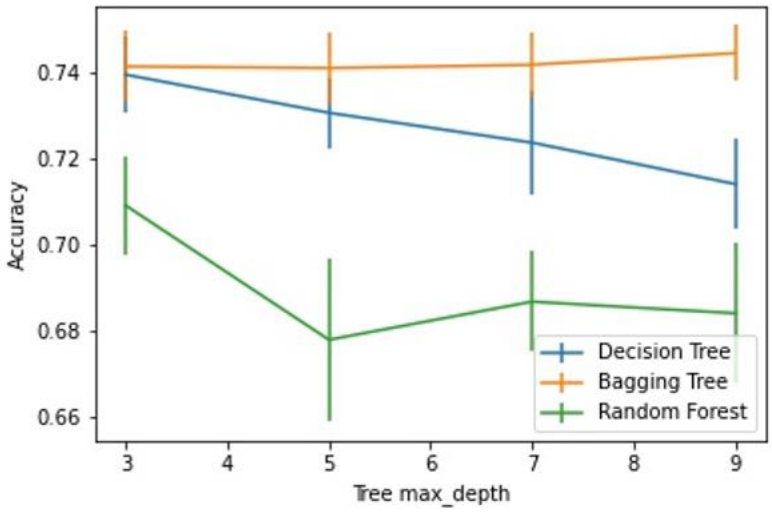

Figure 1: Comparison of different max depth

(b) Formulate a hypothesis about the performance difference (of any two models of your choice) you observe as the depth limit of trees change. Discuss whether the observed data support the hypothesis or not (i.e., are the observed differences significant?). Use significance level (alpha value) 0.05

Hypothesis: Bagging performs better than Random Forest.

Here I used tailed t-test to compare Bagging and Random Forest. The following results are the comparison of Bagging and Random Forest with different max depth.

- Ttest_relResult(statistic=-2.366525713184916, pvalue=0.04214825348592433)
- Ttest_relResult(statistic=-3.948501894699269, pvalue=0.0033623360267848873)
- Ttest_relResult(statistic=-2.3370894934721016, pvalue=0.04423061681282805)
- Ttest_relResult(statistic=-3.8470179358650642, pvalue=0.003924169319596173)

Considering that all statistics are negative, the significance level is 0.05. Because all p-values are less than 0.05, it supports my hypothesis Bagging performs better than Random Forest.

4 Compare Performance of Different Models

(a) Plot the learning curves for the three models (in the same plot), with the average accuracy of the 10 trials on y-axis, and the training fraction on x-axis. Include error bars to indicate ±1 standard error (see descriptions in Assignment 3, Q3(ii)(b) for how to compute standard error).

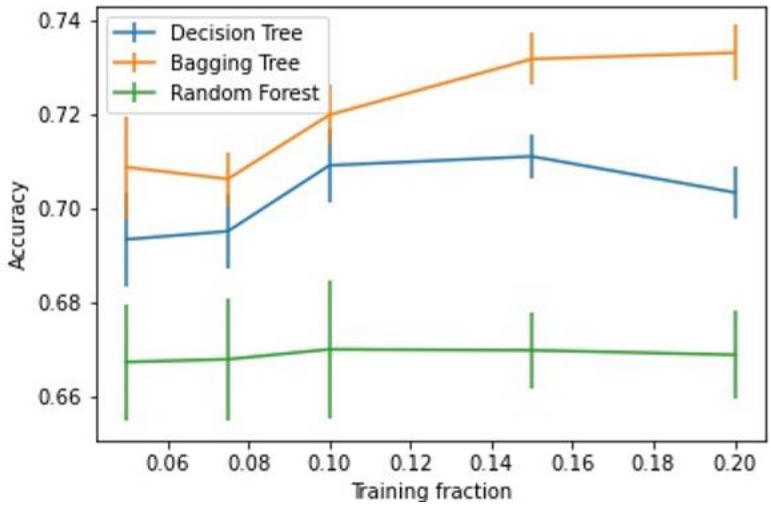

Figure 2: Comparison of different t_frac

(b) Formulate a hypothesis about the performance difference you observe between the decision tree and any one of the ensemble methods. Discuss whether the observed data support the hypothesis or not (i.e., are the observed differences significant?). Use significance level (alpha value) 0.05.

Hypothesis: Bagging performs better than Random Forest.

Here I used tailed t-test to compare Bagging and Random Forest. The following results are the comparison of Bagging and Random Forest with different t_frac.

- Ttest_relResult(statistic=-4.111158735339499, pvalue=0.002632452758859007)
- Ttest_relResult(statistic=-3.0682610614845958, pvalue=0.013391579331970055)
- Ttest_relResult(statistic=-3.6794090197085207, pvalue=0.00508012522265386)
- Ttest_relResult(statistic=-6.254737798444848, pvalue=0.0001487904048228116)
- Ttest_relResult(statistic=-5.815469765951343, pvalue=0.0002544978035289578)

Considering that all statistics are negative, the significance level is 0.05. Because all p-values are less than 0.05, it supports my hypothesis Bagging performs better than Random Forest.

5 The Influence of Number of Trees on Classifier Performance

(a) Plot the average accuracy for 10-fold cross validation on y-axis, and number of trees on x-axis. Include error bars that indicate ±1 standard error. Please include the curves for the two models in one figure.

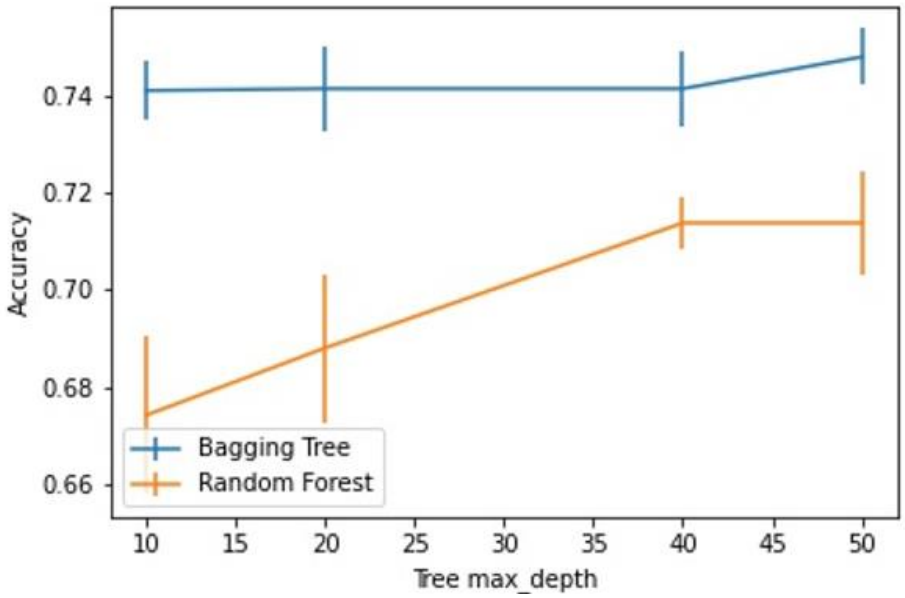

Figure 3: Comparison of different numbers of trees

(b) Formulate a hypothesis about the performance difference you observe between the two ensemble models. Discuss how the observed data support the hypothesis (i.e., are the observed differences significant?). Use significance level (alpha value) 0.05.

Hypothesis: Bagging performs better than Random Forest.

Here I used tailed t-test to compare Bagging and Random Forest. The following results are the comparison of Bagging and Random Forest with different numbers of trees.

- Ttest_relResult(statistic=-3.843642140686407, pvalue=0.0039444850064678265)
- Ttest_relResult(statistic=-3.086857059108836, pvalue=0.0129952513203266)
- Ttest_relResult(statistic=-3.5244467990901205, pvalue=0.00647010233846323)
- Ttest_relResult(statistic=-3.013728303560220 5, pvalue=0.014627202638258494)

Considering that all statistics are negative, the significance level is 0.05. Because all p-values are less than 0.05, it supports my hypothesis Bagging performs better than Random Forest.