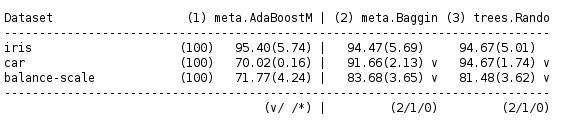
Compsci 361 Assignment 2

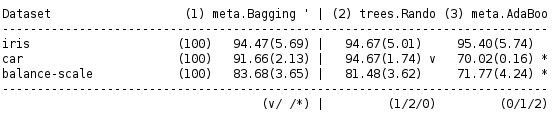
Hasnain Cheena  
190411106  
hche737

## Part A

Screenshot of the 3 runs:

## 





**Class label split**

Iris – 3 classes 50 instances in each class

Car – unacc=1210, acc=384, good=69, vgood=65

Balance-scale L=288, R=288, B=49

## Part B - A paragraph about each of the 3 datasets saying why you think a particular algorithm did the best on each dataset. What is it about this particular dataset that made a particular algorithm do better or worse? Use your knowledge of the algorithms, to hypothesis what must be true about the datasets.

**Iris dataset**

We have no evidence (at a 5% significance level) of a difference in performance between the AdaBoost, RandomForest and BaggingTree algorithm on the iris dataset.

This indicates that the iris dataset must have low target label noise as the AdaBoost algorithm was able to correct true errors and not overfit to class errors/noise.

Further, the iris dataset must have a equal distribution of target labels and all/some of the attributes are of similar predictive strength. This is because if their was a large discrepency in the distribution of target labels, the re-sampled training sets would be appear similar and thus the tree’s would be correlated. However, the tree’s are different from one another thus the RandomForest algorithm performs well.

Along with the equal distribution of target labels the BaggingTree algorithm works well because the attributes in the iris dataset are of similar predictive strength. The BaggingTree algorithm considers all attributes avaliable at each split in the tree. If the iris dataset had a discrepency in predictive strength between the attributes then the tree’s produced would have ben using the same strong predictors at the top and hence be highly correlated.

**Car dataset**

We have evidence (at a 5% significance level) that the RandomForest performed the best, followed by BaggingTree and then AdaBoost.

The AdaBoost algorithm did not perform well because there must be significant target label noise in the underlying data. This means that the AdaBoost algorithm was overfitting, correcting for the errors in the dataset.

RandomForest performed the best among all three algorithms. This is because the RandomForest’s characteristic to randomly select a subset of the predictors as split candidates at each split in the tree. Thus each tree in the ensemble created by the RandomForest was different with no/low correlation.

In comparision the BaggingTree algorithm was affected by the underlying discrepency between the strength of predictors in the dataset. The car dataset must contain a few attributes with very strong prediction power (relative to the other attributes). These strong predictors will be used at the top of each tree, thus each tree is similar and correlated.

**Balance-Scale dataset**

We have evidence (at a 5% signficance level) that AdaBoost performed worse than both RandomForst and BaggingTree algorithms. We have no evidence to differentiate performance between RandomForest and BaggingTree on the Baalnce-Scale dataset.

The AdaBoost algorithm did not perform well because there must be significant target label noise in the underlying data. This means that the AdaBoost algorithm was overfitting, correcting for the errors in the dataset.

The RandomForest and BaggingTree performed similarly indicating tree’s created by both algorithms were different enough from one another with no/low correlation between them. These algorithms performed well because the bagging mechanism within both created re-sampled training sets that were different enough from one another. Thus the underlying data must have relatively equal distribution of class labels and attributes of similar predictive strength.

## Part C -

## A paragraph explaining why you get different results, about which algorithms are statistically significantly different from each other, depending on which algorithm you run first.

## Explain why the results differ depending on the order the algorithms are run in

The results (shown in the tables in Part A) differ depending on which order the algorithms are run in. This is because in each run, a different algorithm has been used as the baseline for the paired t-test. The experimental accuracies of other two algorithms are then compared to the baseline algorithm to determine whether there is a significant difference in accuracy and subsequently which algorithm is better. Thus to evaluate the rankings of all three algorithms, three runs of the experimenter are needed.

## Part D - A final paragraph saying which algorithm you think is more reliable and why.

I believe the RandomForest algorithm is the more reliable than AdaBoost and BaggingTree algorithms.

RandomForest is more reliable than AdaBoost because of AdaBoost’s underlying boosting. Boosting is highly sensitive to target label noise/errors. All data sources will contain errors and so this is a huge disadvantage when using AdaBoost.

RandomForest is more reliable than BaggingTree. Both BaggingTree and RandomForest perform bagging but RandomForest additionally performs feature bagging. Thus the tree’s created by the RandomForest algorithm are typically more different than the tree’s created by the BaggingTree algorithm. This means that generally the ensemble created by the RandomForest will be more diverse and thus better than the BaggingTree.