Omar Alayoubi (oalayo2)
CS 412 Fall 2018
Homework 4

Overall Accuracy:

Dataset	DecisionTree	RandomForest (Average 5 runs)
balance.scale	0.711	0.755
nursery	0.972	0.916
led	0.861	0.832
synthetic.social	0.495	0.54

Brief Introduction on the classification framework:

I have implemented two classification frameworks. The first framework uses a single Decision Tree Classifier, and the second framework uses a Random Forest (RI – Random Input Selection). The two frameworks are written in python 2.7

Decision Tree Classifier

I used the gini-index as the attribute selection measure. The Decision Tree will pick the best attribute that gives the largest impurity reduction. In addition, the decision tree has a heuristic method to pick the best path when it sees a new value in the testing dataset that haven't seen before in the training dataset. The heuristic method works by finding the closest value that it has seen before to the new value given. For example, if the training set contains attribute 1 with values 1, 2, 5 & 7, and the testing dataset has a row with attribute 1 equals 4. The heuristic method will treat the 4 as a 5 because 5 is the closest value to 4.

Random Forest Classifier (RI – Random Input Selection):

The Random Forest Classifier has ten Decision Tree Classifier. Each Decision Tree Classifier get trained on a subset of the training dataset. The subset of training dataset is picked randomly with replacement. In addition, each decision tree classifier has its own subset of features. All the decision tree classifiers have the same subset size of the training dataset and the same subset size of features as well.

Model Evaluation:

Decision Tree Classifier Overall Accuracy

Dataset	Train	Test
balance.scale	1.0	0.711
nursery	1.0	0.972
led	0.859	0.861
synthetic.social	1.0	0.495

Decision Tree Classifier F-1 Score:

balance.scale

Class	Train	Test
1	1	0
2	1	0.823
3	1	0.744

nursery

Class	Train	Test
1	1	0.958
2	1	0.684
3	1	0.983
4	1	1.0
5	1	0

led

Class	Train	Test
1	0.766	0.773
2	0.899	0.900

synthetic.social

Class	Train	Test
1	1	0.486
2	1	0.478
3	1	0.508
4	1	0.507

Random Forest Overall Accuracy

Note: These values are taken by running the program one time. Thus, someone may not find the same exact values, however, the values should be close enough.

Dataset	Train	Test
balance.scale	0.87	0.773
nursery	0.939	0.924
led	0.830	0.829
synthetic.social	0.851	0.523

Random Forest Classifier F-1 Score:

balance.scale

Class	Train	Test
1	0	0
2	0.886	0.785
3	0.894	0.8

nursery

Class	Train	Test
1	0.889	0.849
2	0.712	0.422
3	0.922	0.890
4	0.981	0.976
5	0	0

led

Class	Train	Test
1	0.659	0.623
2	0.884	0.874

synthetic.social

Class	Train	Test
1	0.837	0.477
2	0.833	0.479
3	0.853	0.551
4	0.846	0.606

Parameters:

For Random Forest Classifier, I used the following parameters to determine the appropriate F value and split the size of the training dataset.

```
params = {
        "nursery.train" : [1.2, 3.0],
        "balance.scale.train" : [1.25, 2.5],
        "led.train" : [1.25, 2.0],
        "synthetic.social.train" : [2.0, 3.0]
}
```

The first number “Divider” in the list represent how much percentage we are selecting of features for each decision tree from all the possible features. For example, if we have 10 features in our training dataset and the divider equals to 2.0. Then we are selecting 10/2.0 = 5 features. The

second number “DataSplitter” in the list represent the size of the subset we are selecting from the training dataset. For example, if we have a training dataset with 100 data points and the DataSplitter is set to 3.0 then we are approximately selecting 33 data points to train the decision trees in the Random Forest. The number of decision trees is always set to 10.

There is really no explanation on why I picked these parameters. I simply ran the model with different parameters and I found that these parameters give the best accuracy average for the given dataset.

Conclusion:

The results are mixed. Sometimes the Decision Tree Classifier outperformed the Random Forest Classifier and other times the Random Forest Classifier outperformed the Decision Tree Classifier. However, I think the random forest classifier is more flexible than the Decision Tree Classifier when it comes to outliers because random forest uses the majority vote of multiple decision trees.