CS573_assignment4_XiaofengOu

Xiaofeng Ou

March 2019

1 Preprocessing

2 Implement Decision Trees, Bagging and Random Forests

2.1 Decision tree

```
Training accuracy DT:  0.77
Testing accuracy DT:  0.72
```

2.2 Bagging

```
Training accuracy BAGGING:  0.74
Testing accuracy BAGGING:  0.73
```

2.3 Random Forest

```
Training accuracy RF:  0.70
Testing accuracy RF:  0.69
```

3 The Influence of Tree Depth on Classifier Performance

1. See the graph.

2. **Hypothesis**: n this data set, the depth of the tree has no significant impact on the performance of decision tree, bagging and random forest. The t-test does not reject the hypothesis.

   ```
   p value for the t test at depth 3: 0.351231961844341
   p value for the t test at depth 5: 0.7759841693287619
   p value for the t test at depth 7: 0.8255814504559291
   p value for the t test at depth 9: 0.8173038344301737
   ```

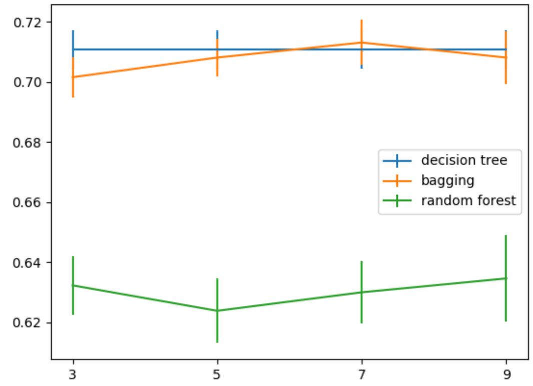

Figure 1: Performance v.s tree depth

4 Compare Performance of Different Models

1. See the graph.

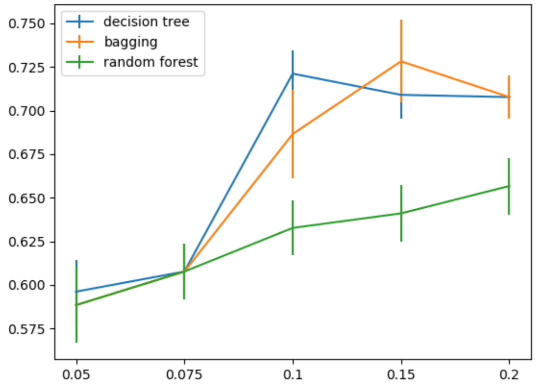

Figure 2: Performance of Different Models

2. **Hypothesis**: On this data set, model performance: bagging > decision tree > random forest
   The observed data showed that the performance of both decision tree and

bagging are better than random forest. But for the comparison between decision tree and bagging, we might need more experiments to confirm. We do not reject the hypothesis.

```
p value for the t test at t_frac 0.05: 0.8002635230569439
p value for the t test at t_frac 0.075: 1.0
p value for the t test at t_frac 0.1: 0.2673914520493188
p value for the t test at t_frac 0.15: 0.5116384016621487
p value for the t test at t_frac 0.2: 1.0
```

5 The Influence of Number of Trees on Classifier Performance

1. See the graph.

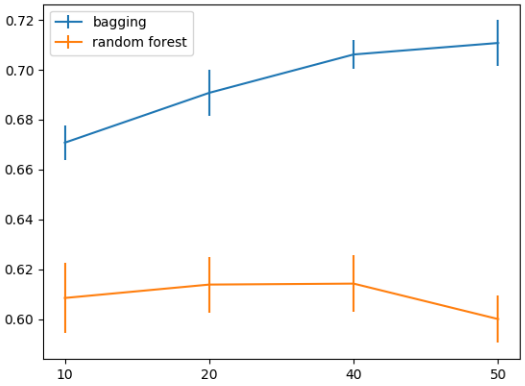

Figure 3: Performance v.s number of trees

2. **Hypothesis**: On this data set, as the number of tree increases, the performance of bagging gets better , while the performance of random forest does not differ significantly.
   The t-test rejects the hypothesis, which means number of tree has a significant impact on the performance of the models.

```
p value for the t test at tree number 10:  0.0013328664490904808
p value for the t test at tree number 20:  9.452711976889715e-05
p value for the t test at tree number 40:  1.890767173260147e-06
```

```
p value for the t test at tree number 50:  2.687081210239488e-07
```