Machine Learning (BITS F464)
Assignment 1
(Decision Tree, Random Forest, Boosting Techniques)

Report

Team Members:

1. Arram Saichand 2014A7PS046H
2. B.Siva Naga Sasank 2014A7PS050H
3. Kampara Sriteja 2014A7PS149H
4. Eda Amos William Prasada Rao 2014A8PS467H

Language used: Java

Data Set used:

- Adult dataset from UCI Machine Learning Repository.

Pre-processing applied on the dataset:

- The instances from the training data were considered as a stringbuffer from which the spaces after ‘,’ were removed and the remaining string was tokenized on ‘,’. These tokens were stored in an arraylist for easy access.

- The instances from the testing data were also processed in the same way as training data mentioned above. But the first line of the testing data which contained irrelevant data was removed. Some of the test instances had a fullstop at the end in the file and they were removed.

ID3(Decision Tree Learning):

- ID3 uses a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct the decision tree.
- For continuous valued attributes the concept of Gini index was used for splitting them into two categories.
- Accuracy achieved on training data – 80.55%
- Accuracy achieved on testing data – 79.23%
- Learning time on training data – 94s

Random Forest:

- The Random Forest was constructed with 100 decision trees at training time. And the output for each instance was decided based on a majority classification from the trees.
- System time was used as a seed for the random number generator function.
- The below are the statistics for one particular iteration.
- The result may be different each time because the trees in the random forest are generated randomly.
- No of features used – 4
- Number of trees used – 100
- Accuracy achieved on testing data – 76.37%
- Learning time on training data – 133s

Adaboosting:

- The Adaboosting technique was applied to the weaklearner(Random Forest) to convert it to a strong learner. 100 trees were chosen at random from the forest and were boosted using the Adaboosting technique.
- The below are the statistics for one particular iteration.
- The result may be different each time because the trees in the random forest are generated randomly.
- Number of classifiers used – 100
- Accuracy achieved on testing data – 82.279%
- Learning time on training data – 102s

Comparision of ID3, Random Forest and Adaboosting:

- Based on the accuracy obtained on the test data it can be concluded that Adaboosting performed better when compared to other two algorithms. This is because ID3 algorithm is generally prone to overfitting over the training data, and random forest, though it generated the trees randomly. Is a weak classifier. Boosting increased the performance of the random forest. Thus, Adaboosting is less prone to overfitting when compared to ID3 and has boosted performance over random forest algorithm.
- The random forest algorithm took longer time to run than ID3 algorithm because of the time taken to construct trees for the forest.