BITS F464 - Machine Learning

Assignment 1

Implementation of ID3, Random Forest and AdaBoost Algorithms

G V Sandeep	2014A7PS106H
Snehal Wadhwani	2014A7PS430H
Tanmaya Dabral	2014A7PS138H
Kushagra Agarwal	2014AAPS334H

BITS Pilani
Hyderabad Campus

Pre-Processing

1. The data contains both numeric and string values. So the data was first read row by row and an ArrayList of DataSet objects was created. Each DataSet object contains the data as it is.
2. The rows containing missing values were ignored as they constitute only 7% of the total data.
3. Since there are some continuous attributes in the data, they were split into two categories based on the split which gives minimum entropy.
4. We then iterate through the ArrayList of DataSet objects and populate a matrix with the corresponding integer value. We get the corresponding numeric value from a class called DataRef where every attribute column is given a number and the possible values it can take are associated with an integer.
   For Eg : attribute[1][0] = "Private", majorRef[1] = "workClass" means that 1 st column of the matrix stores values corresponding to the attribute workClass and "Private" in work Class corresponds to 0.

Note: All those values which are less than the calculated split are given the value 0 and those greater than the split are given the value 1.

Comparing Results

	ID3	**Random Forest**	**AdaBoost**
Running Time	1.9 seconds	41.4 seconds	335 seconds
Accuracy	81.3%	83.5%	84.7%

Random Forest

1. No of trees made : 300
2. No of features considered for splitting : 4

AdaBoost

1. No of instances chosen to build weak-classifier (Decision tree with weak hypothesis) : 10
2. No of weak-classifiers constructed: 5000

Some Observations

1. The accuracy of the simple ID3 takes a dip from 81.32% to around 80% at half the training data whereas, methods like Random Forest and AdaBoost (at around 82.8 and 84.3 respectively) still maintains more or less the same accuracy. This shows that these techniques make the decision tree more robust.

2. Even at 1/4th of the training data, Random Forest and AdaBoost give an accuracy of around 82 and 83.5 % respectively whereas the simple ID3 gives an accuracy of around 79%.

3. Random Forest and AdaBoost do not give the exact same accuracy every time. They tend to vary within a small limit. In case of Random Forest, when we increase the number of trees, the limit becomes smaller. AdaBoost, on the other hand, gives the highest accuracy on test data at a particular value for number of weak classifiers used with the accuracy decreasing if the number of classifiers are increased or decreased.

4. Only 10 examples are considered for AdaBoost at one time because, AdaBoost works on weak - classifiers. AdaBoost does not effect the accuracy if the hypotheses used for boosting are already strong and hence only 10 examples are used to build the decision tree which results in a weak hypothesis.