Decision Tree, Random Forest, and Boosting Techniques

ASSIGNMENT-(1)

Machine Learning | BITS F464 | 10/11/2016

TEAM MEMBERS

- **JULURU VENKATA NAGA SWETHA** **2014A7PS128H**
- **WAJEEHA FATHIMA** **2014A7PS131H**
- **AYUSHI BEHL** **2014A7PS145H**
- **SOAMYA AGRAWAL** **2014A7PS185H**

The coding has been done in JAVA.

Details of training data:

No. of instances: 36251

No. of attributes: 15 (including output)

Details of training data:

No. of instances: 16281

No. of attributes: 15 (including output)

PRE-PROCESSING APPLIED FOR CALCUALTING THE CONTINUOUS VALUED ATTRIBUTES:

The attributes with continuous values are identified, and they are converted to discrete valued attributes (consisting 5 traits using 4 threshold values).These 4 threshold values are calculated such that the information gain is maximum. i.e., the number of instances with each of the 5 traits should be equal in the training data.To do so, all the row entries for the column containing the continuous valued attribute are stored in an array,which is then sorted.And the data at the index values of the threshold values are stored too.Then traversing through the entire column of the original data we change the continuous values to that of discrete if the values in the original array falls in the ranges of these threshold values.

PRE-PROCESSING APPLIED FOR CALCUALTING THE MISSING VALUES:

We identify the attributes that have missing values, and if the output of the instance having the missing value is positive, the missing value is replaced with the trait of the attribute with highest probability of positive outcome .Similar method is applied if the instance is negative.

COMPARISON

	ID3	Random Forest	Ada-Boosting
Accuracy	77.93133099932436	82.34752165100424	82.79589705792027
Learning Time (in milliseconds)	1093	1378	3415