Assignment 2 – Part 2

Machine Learning Cs6375.501

Done By – Kinjal Basu (kxb170730) and Pooja Srinivasan (pxs176230)

1. Following are the assumptions we made –

* Boolean (0 or 1) attributes and Boolean (0 or 1) class values.
* no missing data or attributes
* the first row of the dataset will contain column names and each non-blank line after that will contain a new data instance
* If attribute is used in Left Subtree, we can still use it in Right Subtree for splitting, if it is the best attribute to split at that node.
* Left Child of a parent always has attribute (the attribute which is used to split the parent) value 0, and right child has attribute value 1

1. Following are the best results we could infer from this program –

The Decision Tree’s accuracy has been improved by randomly selecting nodes to prune. Below is the result of one such example -

Pre Prunned Accuracy

---------------------------

Total number of nodes in the tree = 265

Number of leaf nodes in the tree = 133

Number of training instances = 600

Number of training attributes = 20

Accuracy of the model on the training dataset = 100.00%

Number of validation instances = 2000

Number of validation attributes = 20

Accuracy of the model on the validation dataset = 74.30%

Number of testing instances = 2000

Number of testing attributes = 20

Accuracy of the model on the testing dataset = 75.15%

Post Prunned Accuracy

---------------------------

Total number of nodes in the tree = 211

Number of leaf nodes in the tree = 106

Number of training instances = 600

Number of training attributes = 20

Accuracy of the model on the training dataset = 95.00%

Number of validation instances = 2000

Number of validation attributes = 20

Accuracy of the model on the validation dataset = 76.95%

Number of testing instances = 2000

Number of testing attributes = 20

Accuracy of the model on the testing dataset = 76.65%

1. Accomplishment: We have improved the efficiency of the tree by randomly pruning the nodes and selecting the one with maximum accuracy
2. What have we learned?

Ans: Following are the things we have learned –

* Using ID3 algorithm to implement Decision Tree which is more efficient than using random attribute to split at every level.
* Choosing an attribute to split a node using Information Gain
* More the Information Gain, it is the attribute to split
* Pruning nodes randomly improved the accuracy of the tree, implying that there could be overfitting.