Decision Tree ID3

Arunaksha Talukdar

Inrtroduction

Data_form.py is used to Create Training_Data_Selected and Test_Data_Selected each containing 1000 reviews 500 positive , 500 negative from test , training data picked using Random Index Generator , Chosen_Attr_index is formed by taking highest and lowest valued 5000 attributes in IMDB list.

Observation 1

Early Stopping can be achieved by visiting until a certain Tree depth is Reached and Stopping there , making the decision by seeing the label of the node ie Positive Dominant or Negative Dominant .

Features Considered = 1000 out of 5000 For Time Complexity.

Tree Depth	Tree Nodes	Training Data Accuracy(%)	Test Data Accuracy(%)
4	21	64.3	63.3
5	31	64.8	63.4
6	43	65.4	63.4
7	53	65.7	63.4

8	65	65.9	63.4
9	75	66	63.4
10	83	66.3	63.5
15	142	72.1	65

The analysis can be made that with Increasing the Depth of Decision Tree can Help us to Reach A very high Accuracy in predicting the Trainig_data but Accuracy for new Test_data seem not to increase that Rapidly .

Observation 2

Noise is added by changing the labels of positive and negative samples of the Data Set and the tree is Checked for the accuracy of New noise added Training Data .

NOISE_ADDED_TREE_NODES (For Nodes = 123)	NOISE RATIO
127	0.5%
127	1%
125	5%
132	10%

With Increasing noise ratio the number of nodes in the tree increases indicating the complexity of the decision tree formed is increased due to addition of noisy labeled data.

KEY NOTE : Also if number of nodes allowed is decreased by limiting the depth then the effect of noise ratio addition diminishes.

Obervation 3

We used Pruning technique on our Decision Tree in hope of decreasing it’s complexity

By avoiding or reducing overfitting of Data.

In pruning , the less contributing / power nodes are removed by making them leaf node and checking that accuracy of prediction remains constant or increases .

If Accuracy increase or stays close to initial after removing low powering nodes then pruning is successful.

Accuracy-Before-Pruning : 66.3	Nodes:125
Accuracy-During-Pruning: 66.1	Reduced-Nodes:7
Accuracy-During-Pruning: 65.9	Reduced-Nodes:11
Accuracy-During-Pruning: 66.3	Reduced-Nodes:3
Accuracy-After-Pruning: 66.3	Nodes:81

As we can analyze the Pruning Done on our Tree is Successful as Maintaining the Accuracy we were able to Reduce the Number of nodes in the Decision tree , reducing the overall complexity of the Decision tree.

Observation 4

The Forest is used to take the average analysis over data sentiment prediction which in this case is using arbitrary random sequence of Attributes to form 5 , 10 , 15 tress and taking the output of accuracy of prediction from them rather than taking from one tree.

Forest_size	Accuracy_of_prediction
5	55.9%

10	55.5%
20	61.1%

The result is calculated on depth of 5 for time complexity , accuracy can be incread by increasing max depth at the cost of extra-time.