CSL - 603
LAB - 1 REPORT
DECISION TREES AND FORESTS

Eeshaan Sharma
2015CSB1011

September 2017

1 INTRODUCTION

The goal of this lab is to gain a better understanding of Decision Tree based inductive classifier by implementing ID3 Decision Tree Algorithm in order to predict sentiment of a movie review. This goal is realized by constructing a Decision Tree on a Training Dataset which is a small subset of the given Large Movie Review Dataset. The accuracy of the learned decision tree is then tested on a Test Dataset. As a part of a separate Experiment the performance of the decision tree is calculated by adding noise in the dataset. Since Decision Trees tend to overfit the Training Dataset, to improve their performance on the Test Dataset numerous experiments such as Pruning, making a Decision Forest etc are also performed.

2 EXPERIMENT 1 - PRE-PROCESSING

Experiment 1 involved generating a Data File consisting of 1000 (500 +ve and 500 -ve) Training, Validation and Test Examples from the large movie review dataset. It also involved selecting a small subset of 5000 features from the given list of roughly 89000 features.

2.1 GENERATING DATA FILE

The first 1000 lines of the Data File which make up the Training Set is formed by randomly sampling 1000 instances from the train directory of the aclImdb folder. The next 1000 lines involving the Validation Set is formed by again randomly sampling another completely different set of 1000 instances from the train directory. The last 1000 lines which make up the Test Set is formed by randomly sampling 1000 instances from the test directory of the aclImdb folder. All the 3 sets - Training, Validation and Test have 500 +ve and 500 -ve examples each.

2.2 SELECTING ATTRIBUTES

A list of 5000 attributes is selected based on the average polarity of the words given in the "imdbEr.txt" file. The first 2500 attributes consist of the top 2500 words with maximum positive average polarity and the last 2500 attributes consist of the words having least average polarity.

3 EXPERIMENT 2 - ID3 ALGORITHM

As a part of Experiment 2, ID3 Algorithm is applied to learn a decision tree for the training dataset. In order to split on the basis of a given attribute in the decision tree, a threshold is taken, according to which the set of examples are branched. Examples greater than the threshold for the attribute are branched to one side and examples less than the threshold are branched to the other side. The threshold is taken to be the average number of times the given attribute occurs in a review which is calculated by summing up all the occurrences of the given attribute and then dividing by the number of examples.

To study the statistics of the learned tree a stopping criteria of varying the probability of positive/negative samples i.e a threshold on the purity of a node is used. If at a given node the proportion of positive examples is greater than a given threshold, then the node is made a leaf node with positive label and same is the case when proportion of negative examples becomes greater than threshold. The effect of the stopping criteria is then seen on the following parameters.

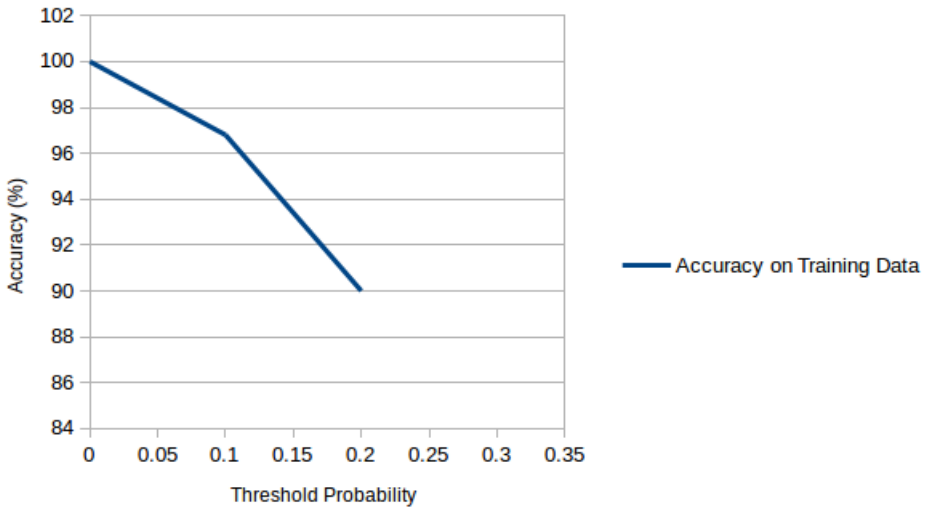
1. Accuracy on Training Dataset.
2. Accuracy on Validation Dataset.
3. Accuracy on Test Dataset.
4. Height of tree.
5. Number of Nodes in tree.
6. Number of Terminal Nodes in tree.
7. Attributes most frequently used as split functions in Internal Nodes of tree.

3.1 Accuracy on Training Dataset

The following table summarizes the performance of varying the stopping criteria on the accuracy of the decision tree on the training dataset. For linearity of graph threshold probability is taken to 1 - (Relaxation in Stopping Criteria). Thus on x-axis of graph value of 0 indicating threshold value of 1

Threshold	1	0.9	0.8	0.7
Accuracy	100	95.8	88.8	82.6

The above observations clearly show that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the accuracy of our decision tree on the training



dataset decreases.

This is because with stopping criteria as 1 the decision tree correctly classifies all the examples in the training set but as we relax the stopping criteria, it might be the case that due to majority of examples of one category present at an internal node a training example is wrongly classified with same label as that of the majority examples.

3.2 Accuracy on Validation Dataset

The following table summarizes the performance of varying the stopping criteria on the accuracy of the decision tree on the validation dataset.

Threshold	1	0.9	0.8	0.7
Accuracy	66.1	66.6	68.8	69.9

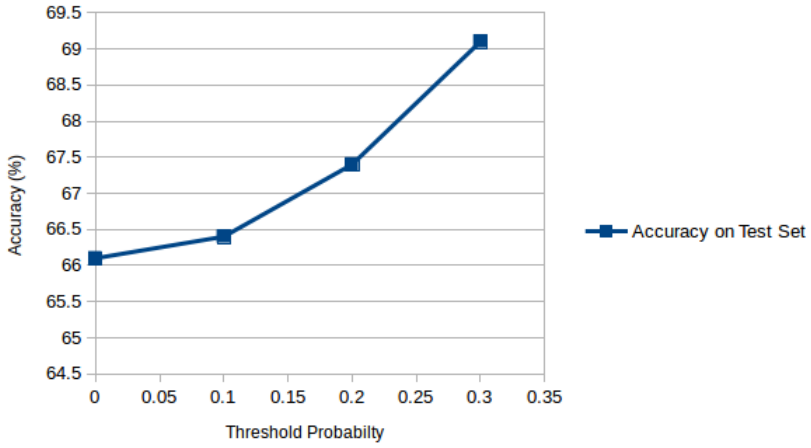
The above observations clearly shows that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the accuracy of our decision tree on the validation dataset increases. This is because as we relax the stopping criteria, the decision tree which overfits the training data, is able to give a better accuracy on unseen instances as degree of overfitting decreases as we relax the stopping criteria.

3.3 Accuracy on Test Dataset

The following table summarizes the performance of varying the stopping criteria on the accuracy of the decision tree on the test dataset.

Threshold	1	0.9	0.8	0.7
Accuracy	65.9	66.4	67.4	69.1

The above observations clearly shows that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the accuracy of our decision tree on the test dataset increases. This observation is due to reasons similar to the case of increase in accuracy on Validation Dataset as we relax the stopping criteria. For linearity of graph threshold probability is taken to 1 - (Relaxation in Stopping Criteria).



3.4 Height of Tree

The following table summarizes the variation in height of the Decision Tree learned by varying the stopping criteria.

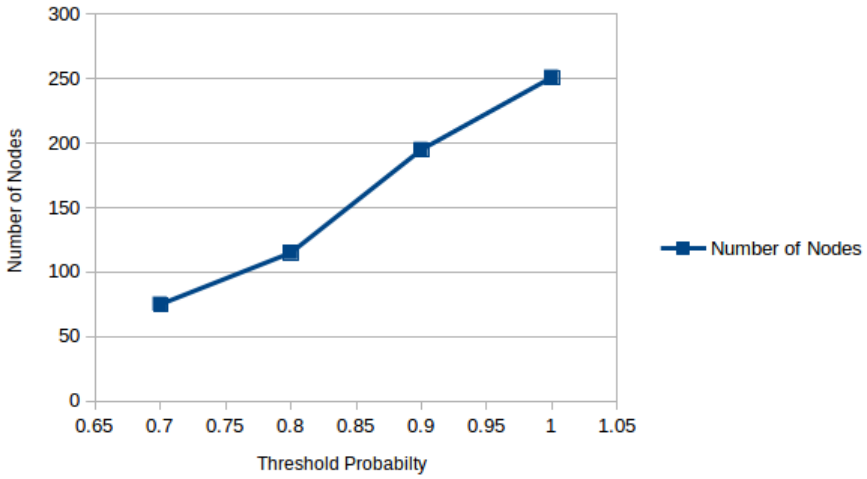
Threshold	1	0.9	0.8	0.7
Height	44	43	37	35

The above observations show that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the height of the learned decision tree decreases. This is because due to adding an early stopping criteria an internal node which could have grown further is simply assigned to be a leaf node of the label of majority examples, which in some sense restricts the tree from growing.

3.5 Number of Nodes in Tree

The following table summarizes the variation in the number of nodes in the Decision Tree learned by varying the stopping criteria.

Threshold	1	0.9	0.8	0.7
Nodes	251	195	115	75



The above observations show that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the number of nodes in the learned decision tree decreases. Its explanation is similar to the trend shown in the case of Height of learned decision tree, as early stopping restricts the tree from growing.

3.6 Number of Terminal Nodes in Tree

The following table summarizes the variation in the number of terminal nodes in the Decision Tree learned by varying the stopping criteria.

Threshold	1	0.9	0.8	0.7
Nodes	126	98	58	38

The above observations show that as we relax the stopping criteria i.e decrease it from 1 to 0.7 the number of terminal nodes in the learned decision tree decreases. Although using early stopping we convert an internal node to a leaf node with majority label but that internal node could have grown further and added several leaf nodes to the tree, which in turn just adds 1 leaf due to early stopping criteria.

3.7 Attributes Used Frequently for Splitting

After summarizing the statistics for all the different early stopping criteria, an analysis is made by traversing through all the 4 decision trees learned on the number of times a given attribute is used for splitting as an Internal Node. The most frequently used attributes are then reported.

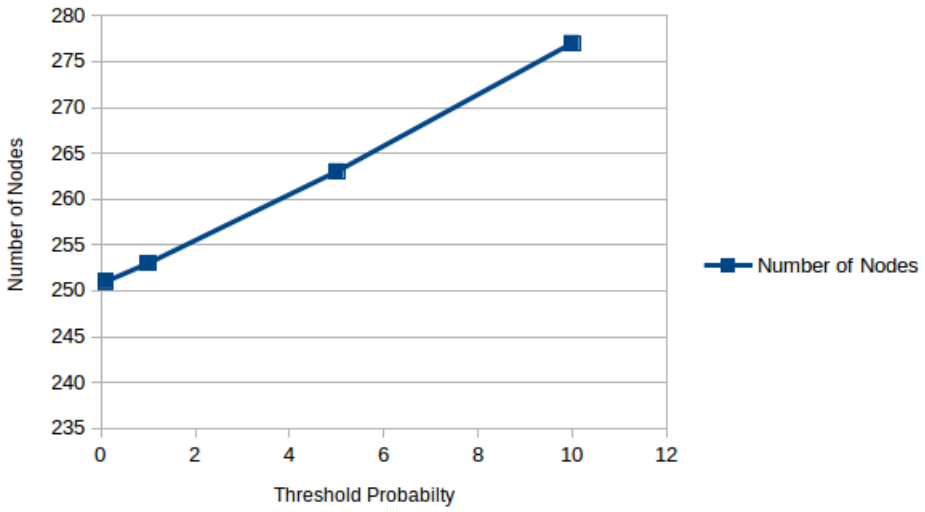
4 EXPERIMENT 3 - ADDING NOISE

In this experiment the task was to add noise to the dataset, by randomly switching the label of a proportion of data points in the training dataset. Different percentages of noise were randomly added to the dataset and then an analysis was made on the variation in the Number of Nodes in the Tree.

The following table summarizes the variation in the number of nodes in the Decision Tree learned by varying the percentage of noise in the training dataset.

Noise %	0.1	1	5	10
Nodes	251	253	263	277

The above observations show that as we increase the percentage of noise in the training dataset, the number of nodes in the Decision Tree increase. Decision trees always try to completely fit all the training examples. This leads to the problem of overfitting. Adding noise to the dataset by randomly switching the label of a proportion of data points can have a serious effect as it can confuse the learning algorithm to produce a long and complex model. This is why sometimes it was also observed that the number of nodes in the tree increased upto 10000. Similar to the trend of number of nodes in the tree is the trend of change in height of the tree upon adding noise.



The trends for accuracy on the Test and Validation Set tend to be completely random with the percentage of noise, as the Decision Tree is only concerned about completely learning the training examples. But more often than not adding noise leads to the formation of complex decision trees which in turn

generally degrades the performance of decision trees in prdicting accurately for the examples present in the Validation and Test Dataset

5 EXPERIMENT 4 - PRUNING DECISION TREE

In their attempt to completely learn the training data, decision trees face the problem of overfitting and hence their performance is sub-standard with respect to accurately predicting the examples in the test data. In order to avoid overfitting, Experiment 4 involved pruning the constructed decision tree. The idea used was that of rule post prunning, in which a decision tree was first allowed to overfit the given training data and then the impact of pruning each possible sub-tree was evaluated on a given Validation Set. Sub-trees that resulted in maximum improvement on the validation data were greedily removed.

The impact of pruning is analysed on the following features -

1. Accuracy on the Training Dataset
2. Accuracy on the Validation Dataset
3. Accuracy on the Test Dataset
4. Number of Nodes.

5.1 Accuracy on Training Dataset

Since pruning prevents overfitting, thus it doesn't allow the decision tree to completely learn the training data set. Hence it is observed the accuracy on the Training Set decreases as we prune the decision tree. Following observations were made -

Accuracy on Training Data before pruning = 100%

Accuracy on Training Data after pruning = 69.1%

5.2 Accuracy on Validation Dataset

Since pruning prevents overfitting by greedily removing sub-trees that increase accuracy over the validation set, so it is obeserved that accuracy on the Validation Dataset increases after prunning.

Accuracy on Validation Data before pruning = 66.1%

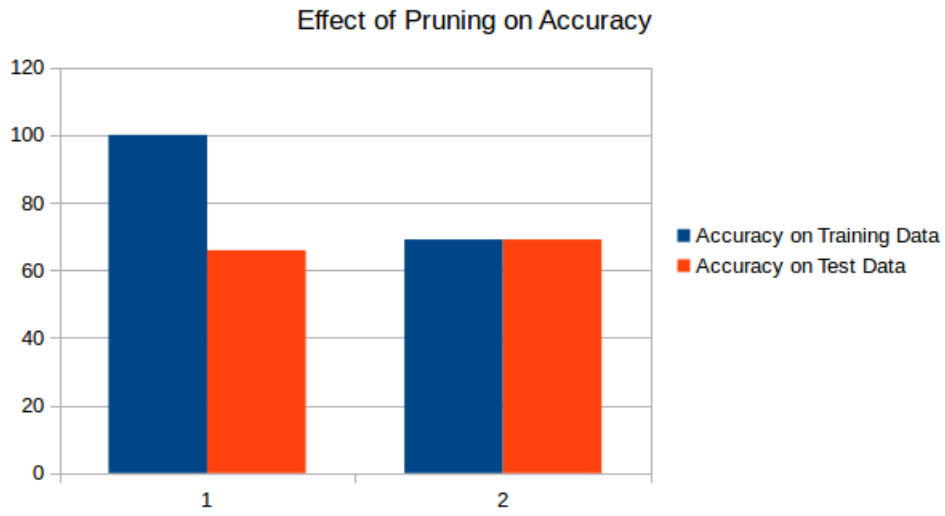
Accuracy on Validation Data after pruning = 72.5%

5.3 Accuracy on Test Dataset

Similar to the trend shown by increase in accuracy on the Validation Set, pruning also improves performance by increasing the accuracy of decision trees in predicting unseen examples present in the Test Dataset.

Accuracy on Test Data before pruning = 65.9%

Accuracy on Test Data after pruning = 69.1%



5.4 Number of Nodes

As pruning greedily removes sub-trees for increasing accuracy on the validation dataset, the number of nodes and height of the tree after pruning decrease significantly.

Number of nodes in Decision Tree before pruning = 251

Number of nodes in Decision Tree after pruning = 91

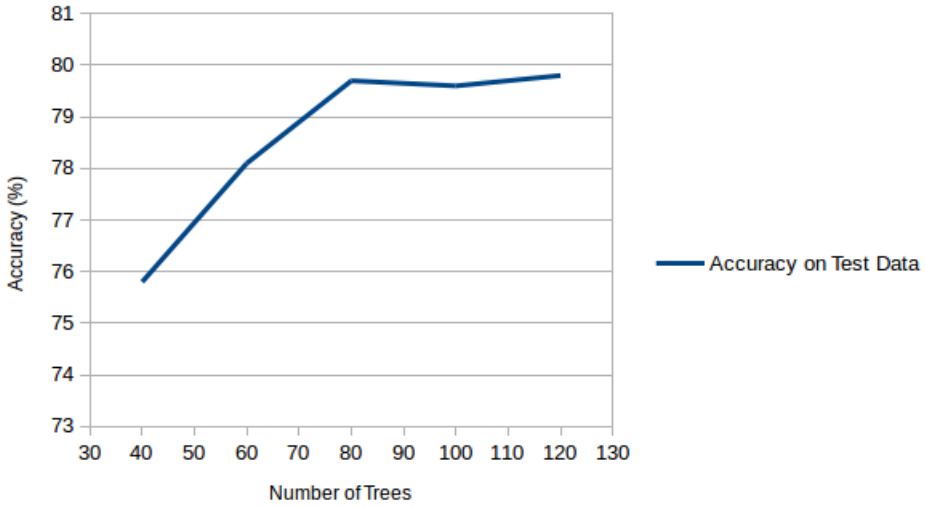
6 EXPERIMENT 5 - DECISION FOREST

The task in this experiment was to learn a decision forest for the given Training Dataset, that uses feature bagging. The accuracy on the Test Data was then calculated by evaluating the label for each example from each tree in the forest and then using majority voting. A random subset of 500 features was used to

construct the decision forest.

Constructing a Decision Forest is another preventive measure for overfitting and its performance with respect to the number of trees in the forest can be summarized using the following table.

Trees	40	60	80	100	120
Accuracy %	75.8	78.1	79.7	79.6	79.8



Increasing the number of trees in the decision forest in general leads to an increase in the accuracy of the Test Data, but when the number of trees in the forest become very large, too much randomness is seen in the performance. General observations have also been made which show that increasing the size of the forest leads to an increase in accuracy, as Decision Forests reduce overfitting by averaging the output over several trees. Feature bagging provides another advantage in terms of decorrelating the trees, which in turn reduces variance when we take average over all the trees.