Decision Tree Sentiment Predictor of Movie

Machine Learning (CSL 603) - Lab1

Piyush Pilaniya (2016csb1049)

Abstract- A decision tree is decision support tool that uses tree like graph. The following report shows consequences of various experiments done to produce the tree and random forest. The increment and decrement in the accuracy over test data is discussed in details.

Index Terms- Machine Learning, Decision Tree, Random Forest, Pruning

I. Introduction

This report present the observations and results obtained by completing the task of movie sentiment classification using decision tree and performing various experiments on it. The problem was at first stage tackled by ID3 algorithm of decision tree which was further improved by various techniques like early stopping, pruning, random forest .

II. Experiment 1:Preprocessing

We have been given with a dataset which contains 25k examples each for training and testing and around 89k features of words. So the first task is to sample these features and training data.

For the sampling of training data, I have picked at random 1000 instances both from train and test data and stored them in 'train_data.txt' and 'test_data.txt' respectively. Out of the instances picked 500 are positive and 500 are negative.

For the case of attribute we have been given with the polarity of the 89k words of the dictionary, out of which we have to select 5000 words as attributes. The selection criteria was to choose the most positive and the most negative polarity words. For which the data was first sorted and 2500 words from each side were selected.

III. Experiment 2: ID3 Algorithm

ID3 algorithm is a greedy algorithm to implement a decision tree. In this algo, **feature selection** is one of the major benefit, at every node we choose the feature for which the information is maximum out of all available features . Information gain is measured in terms of entropy. If the reduction in the entropy is higher then Information gain will also be higher.

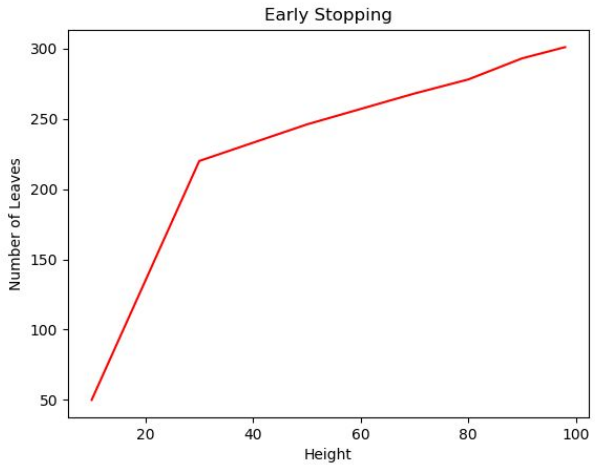
Now once the feature which is best with the given examples and attribute is found then we need to split the examples further. Those examples which are negative for that particular attribute goes to left subtree and those which are positive goes to right subtree and the above algorithm is recursively followed on both left and right subtree.

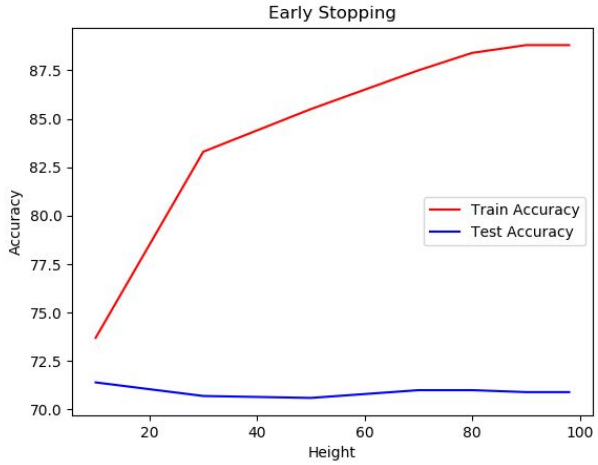
Observation of Decision Tree without early stopping	
Accuracy	70.9
Height	98
Number of Leaves	301
Total Nodes	601

Early Stopping: This is a type of regularization method used to avoid **overfitting** at an early stage by setting some threshold.

Here I have set a threshold on the height of the tree. If the tree height exceeds that particular threshold value then it is stopped at that point because creating further might lead to overfitting of the train data leading to higher train accuracy and ultimately lower test accuracy.

Below are the result of varying the height of early stopping:



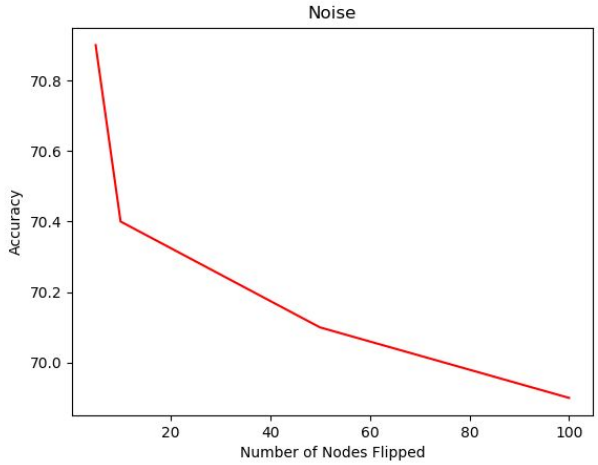


We can see through the graph that as height of the tree is decreasing meaning as the threshold for early stopping is reduced , the number of leaves is also reducing . Whereas in the case of accuracy the training accuracy is getting reduced and the testing accuracy first decreases and then increases and becomes constant.

Attributes which are **most frequently** used as **split function** can be founded by traversing the tree built using ID3 and then at every point updating the list maintained to store the number of occurrences of the attribute.

IV. Experiment 3:Introducing Noise

Noise refers to irrelevant information or randomness in the data. We can introduce noise in the data by flipping the labels of some of the data. For this experiment we are required to randomly select 0.5%,1%,5% and 10% of the data.



We can see that as the number of nodes which are flipped increases the testing accuracy decreases because of the randomness in the data.

V. Experiment 4:Pruning

Pruning is widely used technique to prevent overfitting by reducing the complexity of the classifier and hence improve the accuracy on the test data set. In this we make attempt by removing the particular section of the tree and checking if the accuracy is getting improved or not. In this I have started from the top most root and then traversed the tree in level order fashion using queue and then remove the tree in which accuracy was getting improved. After pruning the size of the tree got reduced to 60% approximately. Unfortunately in the last hour code got some error and i submitted without that code and without the graph.

VI. Experiment 5: Random Forest

Random Forest are ensemble learning method that is operated by constructing a variety of decision tree by any method : either feature bagging or instance bagging. Here I have used feature bagging by selecting 2k feature out of 5000 features randomly. This is done for various number of trees , from 1 to 20. It was observed that by increasing the number of trees , the accuracy was also increasing.

Observation regarding Random Forest	
Accuracy on test set with 20 trees	71.7