CSL 603 - Machine Learning (Lab 1)

Sentiment Classification Of Movie Review

Rishabh singh - 2016CSB1054

August 28, 2018

Introduction

In this lab I made a simple decision tree to classify the result of movie reviews as ‘Positive’ or ‘Negative’ on the basis of Imdb rating. I took randomly 500 positive reviews and 500 negative review create a training dataset and a Test Dataset. I also selected top 2500 attributes with positive polarity and top 2500 attributes with negative polarity. With this dataset I learned the Decision tree and the learned Tree shows following statistics.

1. Effect of Early Stopping

I used height of tree as Early stopping criteria for decision tree. I tried height of tree equal to 90,75,55,40,30 percentage of the actual height. The following table shows the observation for this cases.

S.no.	Height Percentage of Actual Decision Tree	Number of Terminal Nodes	Most frequently attribute used to split	Frequency of most frequently attribute to split	Accuracy on Training Dataset	Accuracy on Test Dataset
1.	100%	419	344	8	91.5	71.5
2.	90%	397	344	8	91.5	71.5
3.	75%	364	344	8	91.5	71.5
4.	55%	313	344	7	91.4	71.5
5.	40%	269	344	7	89.1	71.1
6.	30%	246	344	7	86.8	70.9

From the above table we observe that in our case early stopping is not helping much as it is only decreasing the accuracy on the the Test dataset. Moreover since we are reducing the height, which modifies the tree, so the accuracy on the Training dataset also decreases. The number of the terminal (leaf) node also decreases which is direct result of the reduction of height. Moreover we also observe that reducing the height doesn't change the node which is used as split maximum number of times.

2. Effect of adding Noise in Dataset

In this Experiment, I added noise to the Training dataset and then calculated accuracies of the tree. Since I randomly added the the noise the result are purely random. There may be an increase or decrease in the accuracies depending on the split. The result obtained from my dataset are listed below.

S.no	Noise (%) added in Training dataset	Height of Tree	Number of Terminal nodes	Total Number of nodes	Accuracy
1.	0.0%	222	419	837	71.5
2.	0.5%	225	421	841	71.4
3.	1.0%	224	421	841	71.2
4.	5.0%	229	428	855	71.2
5.	10.0%	240	442	883	71.5

3. Effect of Pruning on Decision Tree

In this Experiment, I used Reduced Error Pruning method to prune the Decision tree. Obviously it resulted in increase in accuracy of decision tree. The result obtained are listed below in the table.

S.no	Number of Nodes	Height of Tree	Accuracy
1.	837	22	71.5
2.	557	99	71.8
3.	553	99	72.1
4.	539	99	72.5
5.	517	99	73.0
6.	515	99	73.1
7.	481	99	74.0
8.	447	99	74.8
9.	387	99	75.2

Since pruning deletes a subtree the number of nodes decreases. It also decreases the height of the tree.

4. Random Forest Creation

In this experiment I created a Random Forest using Feature Bagging method. I used random 2000 attributes and created at max 40 trees in forest. The result obtained are listed below in table.

Dataset 1

S.no.	Number of Trees In Forest	Accuracy
1.	1	71.5
2.	3	69.6
3.	5	68.4
4.	10	74.4
5.	18	73.2
6.	25	72.1
7.	40	72.0

Dataset 2

S.no.	Number of Trees In Forest	Accuracy
1.	1	71.1
2.	3	71.9
3.	5	73.5
4.	10	74.4
5.	18	74.6
6.	25	76.6
7.	40	75.4

From the table I observe that on an average accuracy increases by applying Random forest algorithm.