CSL 503-603

Report Assignment 1

Sentiment Analysis of movie reviews

By: Abhinav Jindal
2016csb1026
Aug 28, 2018

Overview

This program is used to predict the sentiment given the string of a movie review using decision tree. For normal decision tree, I have taken 5000 features and 1000 instances. Several different cases have been observed as shown below:

Early Stopping

The height of the tree is fixed in this method and whenever a node reaches at that level its sentiment value is decided on the frequency of more positive or negative reviews at that point.

Height Ratio To Original Tree	Total Number of Nodes	Total leaves	Most used indexes for splitting (index : frequency)	Accuracy on training set (%)	Accuracy on test set (%)
1	819	410	422:7,328:7,622:5,584:4	92.2	70.199
3/4	711	356	422:7,328:7,622:5,584:4	92.2	70.199
1/2	533	267	328:7,422:6,622:4,584:4	91.2	69.699
1/4	233	112	422:5,328:4,344:4,373:3	79.0	73.9
1/8	3	2	240:1	55.9	56.399

Observations: The accuracy(on test data) increases at a certain level (here h/4),

otherwise it decreases . Therefore the reduction should be done till a certain level. As expected, the accuracy on training data decreases with increase in height reduction.

Noise addition

Noise is added in the training data and the tree is trained with that data.

Percent of noise added	Total Number of Nodes	Total leaves	Height of tree	Accuracy on training set (%)	Accuracy on test set (%)
0	819	410	215	92.2	70.199
0.5	821	411	215	92.0	69.6
1	825	413	220	91.6	71.1
5	835	418	211	90.6	70.0
10	839	420	229	89.4	69.699

Observations: The accuracy(on test data) is not affected much by noise addition. But it decreases on training data with increase in noise. The height and consequently, the nodes and leaves show an increase in number with increase in noise.

Pruning

A validation set is created and the decision tree is pruned until the accuracy increases on this validation set. This reduces the tree height and nodes.

	Total Number of Nodes	Total leaves	Height of tree	Accuracy on training set (%)	Accuracy on test set (%)
normal tree	819	410	215	92.2	70.199
after pruning	225	113	89	86.4	74.7

Observations: Pruning shows a significant increase in accuracy on test data. It is fairly fast as compared to random forest and shows good results. Accuracy on training data decreases here too.

Random Forest (Feature Bagging)

(for 2000 features)
A random set of 2000 features is chosen from the 5000 features and the given number of trees are made from those sets. The result from max number of trees is chosen as the sentiment value.

Number of trees in a forest	**Accuracy on training set (%)**	**Accuracy on test set (%)**
1	79.3	68.7
2	75.1	64.1
5	84.399	71.399
10	82.1	73.1
20	84.2	72.5
50	85.0	72.1

Observations: Accuracy mostly increases with increase in number of trees. With number of trees more (here more than 20) it exceeds the accuracy of normal decision tree with 5000 features. But this process is very slow as compared to pruning.