CSL 603 - Machine Learning : Assignment 1

By - Komal Chugh
2016csb1124

Objective : The objective of this assignment was to classify a movie review as positive or negative. We used ID3 algorithm to build decision tree from training dataset. Further we tried to improve the test accuracy by performing experiments like early stopping, pruning and decision forest using feature bagging.

Pre - Processing : We need reviews and attributes which will be used for building decision tree. 1000 random reviews are selected from train dataset ensuring that exactly 500 have positive label and 500 have negative label. Similar thing is done for test dataset. And also, 2500 attributes having most positive sentiment and 2500 having most negative sentiment have been selected and are used as features in decision tree.

ID3 Algorithm : We have constructed a binary decision tree and the splitting criteria used is the presence of attribute in a review. If the attribute is present, review goes to left child, otherwise it goes to right child. The attribute which gives highest information gain is used as node and left and right child are selected in the similar manner. If there is no increase in the information gain ie if it remains zero, a leaf is made with label equal to maximum occurring label in the examples.

Experiment 2 : **(Early Stopping and most frequently used attributes)**

- Early Stopping → The criteria used for early stopping was the number of examples reaching a node or the maximum depth of a tree. Different experiments were conducted with different stopping criteria. The first one was the minimum number of examples at a node ie, if the number of examples is less than the threshold then make it a leaf. The second criteria was based on maximum depth of tree. The effect of early stopping on train and test accuracies is discussed below :

Stopping Criteria	Train Accuracy	Test Accuracy	Number of leaves
Without early stopping	92.2	69.2	426
Number of examples < 50	84.9	74.5	225

Number of examples < 100	84.8	75.1	215
Maximum depth = 50	88.2	69.7	252

Observation : Train accuracy decreases with stopping criteria whereas test accuracy increases. And also the number of leaves (terminal nodes) decreases. This is because decision tree is becoming less biased towards the training set and hence it reduces overfitting. But if we increase the stopping criteria beyond a limit, for eg number of examples < 250 then both the test and train accuracies would decrease. Because in this case many examples will get wrong label.

- Most frequently used attribute → Without early stopping, the top 10 most frequently used attributes are as follows :

Attribute	Frequency
crap	8
ridiculous	7
pathetic	7
terrible	6
horrible	6
badly	5
idiotic	4
redeeming	4
garbage	4
rubbish	4

With early stopping criteria as number of examples < 50 , the distribution changes to :

Attribute	Frequency
crap	2
stupid	2
disgusting	2
avoid	2
sucks	2
dud	2
poor	2
awful	2
terrible	2
mullets	1

Experiment 3 : **(Addition of noise in training dataset)**

Noise is added by randomly switching the labels of training dataset. Following table shows effect of noise on the number of nodes in the tree.

%age noise	Number of nodes	Test Accuracy
0	425	69.2
0.5	427	69.1
1	417	69.2
5	432	68.6
10	451	68.9
20	463	68.0

Observation : Noisy data causes overfitting and the decision tree tries to achieve maximum accuracy on the training data and hence the complexity of the tree increases. As can be seen, the number of nodes in the tree grow with the addition of noise. The prediction accuracy on the test dataset generally decreases because the decision tree is overfitted. And hence shorter trees are preferred over complex trees.

Experiment 4 : **(Pruning)**

In this experiment, test data is used as validation set and hence pruning results in increase in test accuracy. For pruning, subtree that gives maximum increase in accuracy is removed and a leaf node is made whose label is decided by maximum occurring label in the examples at that node. The following table shows the transition from the initial tree to final tree:

Tree	Test Accuracy	Train Accuracy	Number of nodes
Initial tree	69.2	92.2	425
Intermediate tree 1	72.4	92.0	389
Intermediate tree 2	73.3	91.2	365
Intermediate tree 3	74.0	90.3	338
Intermediate tree 4	74.6	89.4	198
Intermediate tree 5	74.9	89.2	195
Intermediate tree 6	75.1	89.2	194
Intermediate tree 7	75.3	88.7	175
Intermediate tree 8	75.4	88.6	174
Final tree	75.5	88.2	162

Observation : Number of nodes will obviously decrease since we are removing subtrees that result in increase in accuracy. Test accuracy increases because our algorithm is designed in such a way that if there is no increase in test accuracy, we stop

pruning further. On the other hand, train accuracy decreases because removal of a subtree will lead to increase in misclassification of labels.

Experiment 5 : **(Decision Forest using feature bagging)**

In this experiment, various number of trees were added to the forest and their accuracy was recorded. In each tree, there are 2000 features which have been chosen randomly from 5000 features. For predicting label of a review, majority voting has been used ie from each tree label was extracted and then the maximum occurring label is the output of the forest. The following table shows the effect of number of trees on test and train accuracy :

Number of trees in forest	Train Accuracy	Test Accuracy
1	76.6	67.1
3	82.3	70.1
5	82.1	71.4
10	83.6	71.1
15	84.1	71.1
20	85.1	72.1
25	84.9	72.3

Observation : Test accuracy increases on increasing the number of trees in the forest because now predicted label is chosen from an increasing number of trees (majority voting) and hence overfitting is reduced. In the similar manner train accuracy increases.