CSL 603 Machine Learning(Lab 1)

Sentiment classification of movie reviews using decision trees and forests

Prinshu Kumar
2016csb1051

This assignment produces a decision tree for the given input and then predicts a decision tree model on the randomly generated train data. Random forest has also been created on 10 trees with 4000 features. The major parts of the assignment are as following :

1. **Creating Dataset** : Large Movie Review Dataset from Stanford has been used for the experiment. Train and Test data were created using randomly taking the BoW representation of the reviews provided with the dataset. Any review which is greater than or equal to seven has been given negative one value and any dataset which is less than or equal to four has been given positive one value. Equal number of positive as well as negative reviews has been taken for the train as well as test purpose. A huge number of words with their polarity is also provided with the dataset. 2500 words with the highest polarity and 2500 words with the negative polarity has been taken as the attributes for constructing the decision tree.

2. **Decision tree construction** : A decision tree has been constructed on the train data. The accuracy of the tree was 72% on Train data. Various analysis has been done in the below table 1.
   Early Stopping : The accuracy has increased in the case of early stopping for the test data whereas the accuracy has decreased in the case of the train data. This was due to the fact that the data was then less prone to overfitting.

	Full Tree	**Early Stopping**
Train	87%	85%
Test	72%	73%

Table 1 : Train and Test accuracy on the Full tree as well as with the early stopping

3. **Noise Addition** : 0.5, 1, 5 and 10% noise were added to the data and the accuracy decreased less at first and then it decreased very fastly. The accuracy on the test data is given in the below table 2.

Noise	0.5%	1%	5%	10%
Accuracy	72%	52%	50%	50%

Table 2 : Accuracy on the test data when various percentage of noise is added

The above decrease in accuracy is due to the fact that when the error increase then the tree start to learn the noise in the data.

4. **Pruning** : The tree has been pruned and as a result the accuracy has improved. If by removing any node the accuracy is increasing then that node is removed from the tree. Here we can see that after removal of some node the accuracy is decreased to as low as 46% and for some node the accuracy is 73%.
5. **Random Forest** : A random forest on ten trees was implemented and the observation is that the individual accuracy for each tree was very less. However the final accuracy was greater than the individual accuracy of the tree in the forest. To get the final output the maximum of the predictions of each tree was taken. The accuracy for each tree and the final accuracy is given in the below table 3.

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**Final**
Accu.	65%	61%	62%	62%	54%	56%	59%	53%	50%	50%	69%

Table 3 : Accuracy on the test data for the 10 trees of the random forest and the final result.

One thing that is to be observed is that the final accuracy is greater than the individual accuracy of any tree. This is due to the fact that there are some noise in the data which got trained in the tree. Once we take the maximum(higher number of prediction) of the prediction the noise in one tree is removed by the real training in the other tree.