Machine Learning – ITCS 6156

Assignment – II

Status: Complete

I had implemented all the modules required to build the 3 classifiers (Naive Bayes, Decision Tree and Logistic Regression). But I have shown my results on the subset of the 20 News Groups data. I tried at each level to see the efficiency of each classifier at least 70%. The only reason for me to implement on data subset is the running time. As running time bumps high I can’t do so many tests and analyze to improve the accuracy. But the code works fine for the main dataset also.

Analysis:

Naive Bayes Classifier: It is too fast and easy approach to solve a classification problem. It depends on a single parameter alpha which should bet set to 1, so that the classifier output will be at peak. The efficiency of this classifier might be effected because of missing values or contexts in the taken sample.

Decision Tree Classifier: It is dependent on two parameters. One parameter is the depth of the tree to be constructed, as depth increases the efficiency of the classifier also increases. The second parameter is the selection criteria we followed to choose a feature from all possible ones at the corresponding point. Basically, I used two of those criteria Information Gain and Gain Ratio. Among them Gain Ratio is better one. The efficiency of this classifier is also influenced due to the missing values in the sample. To handle this, we can replace missing value in feature ‘f’ with label ‘l’ with the most common value present in that feature with that label.

Logistic Regression Classifier: This classifier needs to be more tuned than the remaining two. One parameter is step value which is dependent on the data we choose. One major thing is in prior to build this model we need to frame an equation either linear or polynomial, based on the equation we choose the entire fitting of the model is decided. So choosing an equation would be more complicated for this classifier, apart from this it gives a high accuracy.

Challenges Faced:

The easiest part of the assignment is to build the Naive Bayesian Classifier because for writing the code as well as for implementing, it doesn’t take too much of time. The next easiest one is Logistic Regression Classifier where it is very easy to understand and code but it takes too much of time while executing to produce results. The one of the most challenging part of the assignment is to deal with the Decision Tree Classifier.

The first problem I faced was since I implemented the Decision Tree Classifier with normal matrix it gave me the error “Out of Memory”. To eradicate that problem, I changed entire implementation to sparse matrices.

The next problem is, without including cross validation when I executed the Decision Tree Classifier on entire dataset, it is too poor in its efficiency. In order to improve the problem, I thought to replace the attribute selection criteria (Information Gain) with Gain Ratio and it is able to increase its accuracy.

The next problem is with the time of execution, when I implemented Classifier with the cross validation concept it is taking too much of time to produce a result. So for that I selected the 5000 features at first using Gain Ratio and then executed the classifier which gave too low accuracy. To this issue I just removed the selection of features which is very poor. I think feature selection works when the missed attribute values are handled. Even I tried to handle those missing values but it's too time taking. So I thought of to implement it on the entire data set without going for feature selection.

Future Work:

We implemented by treating all features as binary features, I think this is also one of the reason for low performance. So, I will try to solve them by using multinomial approaches.