Data Mining Course Project Final Report
Team ID -14
Members: Yashashvi Kampalli , Vipin Sai Siripurapu

Problem: Given a large classified data set, build an efficient classifier to predict the class labels for an untrained data.

Features of data:
1) Very large data set mostly sparse
2) Large number of repeating or empty attributes

Approach 1: (Naive Bayes classification by eliminating redundant and empty attributes)

We have identified that a large umber of attributes in the given training set were repeating or empty. We performed preprocessing on this set to identify and eliminate duplicate and empty columns. The procedure is as follows:

1) We compute the hash of the row id and data values for each column (attribute) starting from column 1 and constructed a tree map with column ids as nodes
2) For all further columns starting from 1, we compare the hash value of the current column to existing nodes in the tree (i.e., already processed columns) and if the hash is same as any of the existing value, it means that the current column is a duplicate.
3) We made note of all the unique and duplicate columns and eliminated the duplicates from the training set.
4) We then used this list of redundant and unique column ids list to modify the test test accordingly to maintain the correspondence between the training and test set.

Results: After the reduction of attributes we found attributes to be unique and the data set size reduced considerably.

We then converted the reduced training data to weka format(ARFF) and built a Naive Bayes classifier on it.

Validation: We performed a 10 fold cross validation on this classifier using weka with the reduced data set and the resulting accuracy was 75 %.

Accuracy: The accuracy of this model on the test set turned out to be 73% close to the cross validation results.

Approach 2: (Ada boosting of the Naive classifier)

We tried boosting the classifier by performing AdaBoosting on it using weka. We created a set of ten classifiers on the data without resampling. We then performed Adaboosting on this set with a weight threshold of 100.

Results: The accuracy of Adaboosting on 10 fold cross validation was 70 %
Inference: The Adaboosting on the reduced data set does not improve the accuracy as a lot of features were eliminated.

Approach 3: (Updateable Naive Bayes on the entire data set without feature reduction)

To reduce the detrimental effects of attribute reduction, we built an updateable naive bayes classifier on the entire data set. We did not discretize the numerical attributes.

Results: The accuracy of this classifier on a 10 fold cross validation over training set was 76%.

Conclusion: The elimination of redundant and empty features was reducing the classification accuracy. So the redundant attribute were in fact related but not the same attribute.

Approach 4: (Updateable Naive Bayes Multinomial)

Since the data set is sparse, using probability distribution functions to calculate the prior probabilities could improve the accuracy. With this motivation we built a NaiveBayes classifier with multinomial probability distribution function using weka.

Results: The accuracy of this model on a 10 fold cross validation was 78%.