**Data Mining Project Report**

**Siddharth Sujir Mohan (1207653769)**

**Kranthi K Rekala (1207663258)**

**Introduction:**

A classification project involves classifying a set of unclassified records by building a model using a set of training records. The training set is used to build a model by using the desired classifier algorithm. The choice of the classifier algorithm depends on the type of the data that we want to classify. For the project, we were given a data set consisting of 5000 attributes which were of numeric type. The class was of nominal type (1 and -1). The training set had 6500 records and the test set had 1000 records. The class labels were given on a separate file. For simplicity, the class labels were merged with the training records before constructing the model. We also used cross fold validation while building the model.

**Submission-1 Details:**

For the classification task, we used Weka tool to build the model and for predicting the class labels of the test record.

The following tools and algorithms were used for the classification project

1. Weka 3.6.12
2. libsvm

Algorithms tested during the development:

1. Random forest
2. SVM
3. Naïve Bayes

During the first submission, we used 2000 trees to construct the model and 200-folds for cross folds validation. The model accuracy that we obtained was 93.36% and the prediction percentage that was obtained was 96.7%

**Submission-3 Details:**

For the 3rd phase of submission, we considered the fact that using 2000 trees and 200 folds for cross fold validation may give rise to **model over-fitting** which in turn may reduce the accuracy of the final prediction. Hence we tried to reduce the number of trees used by the random forest algorithm and also reduced the number of folds in cross fold validation to 10. The model accuracy that we obtained dropped to 92.6%.

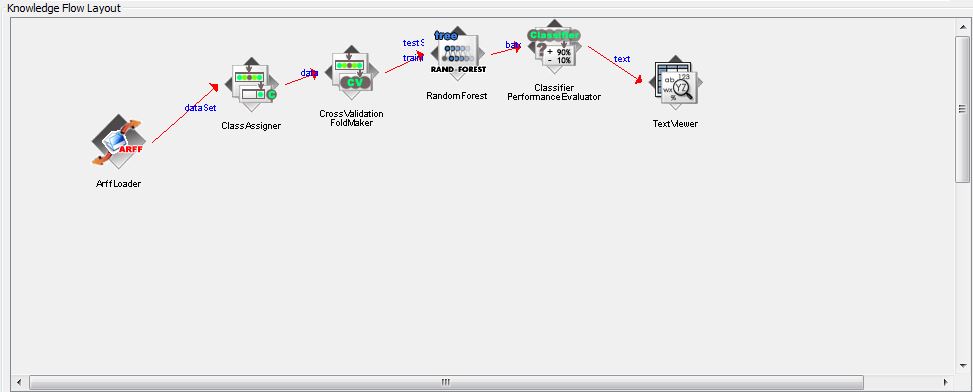
We assume that the model over-fitting problem would have been overcome and the accuracy of the final prediction would go up by few percentage.

|  |  |
| --- | --- |
| **Classification Approach** | **cross validation score** |
| Decision Tree | 88.34% |
| Random Forest (200 folds) | 93.36% |
| Random Forest (9 folds) | 92.6% |
| svm | 65% |
| Naïve Bayes | 85.5% |

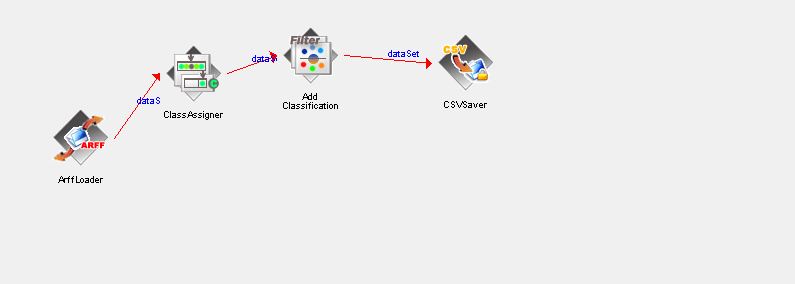
From the above table we obtained the maximum accuracy for the model that used Random forest with 200 folds and with a model accuracy 93.36%.

**Code:**

**Knowledge flow layout for building the model**



**Knowledge flow layout for predicting the labels for the test set**

****

**Knowledge Flow to Build Model**

folds: 10

**Classsifier Used:**

Random Forest

options:

numTrees 100

maxDepth 0

numFeatures 0

seed 1

**Reference:**

**[1] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomForest.html**