Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
Vile Parle (W) Mumbai - 400 056.

Learning to Classify Imbalanced Data Streams for Binary Classifiers

Project Guide: Prof. Kiran Bhowmick

Megha Jakhotia (60004140033) Yash Kapadia (60004140040) Nikita Luthra (60004140047)

Abstract

Most of the Binary Classifiers predict accurately for balanced data sets. But when dealt with imbalanced data set, it intends to be biased towards majority class thus ignoring the minority class. However, misclassifying minority class data can result in heavy costs. Learning from imbalanced data has conventionally been conducted on stationary data sets, but when dealing with imbalanced data streams, it encompasses a greater challenge as the large input streams are accessible for a small amount of time. The task gets more difficult as the class labels in the data streams arrive with a delay. In this paper, we propose a model to handle these issues and classify the Imbalanced data streams.

Introduction

Classification is the process of assigning the input data sets into various groups. The training set on which a classification algorithm is trained might not contain a balance between the number of examples from the different classes and this raises the issue of Imbalanced data in classification problems.

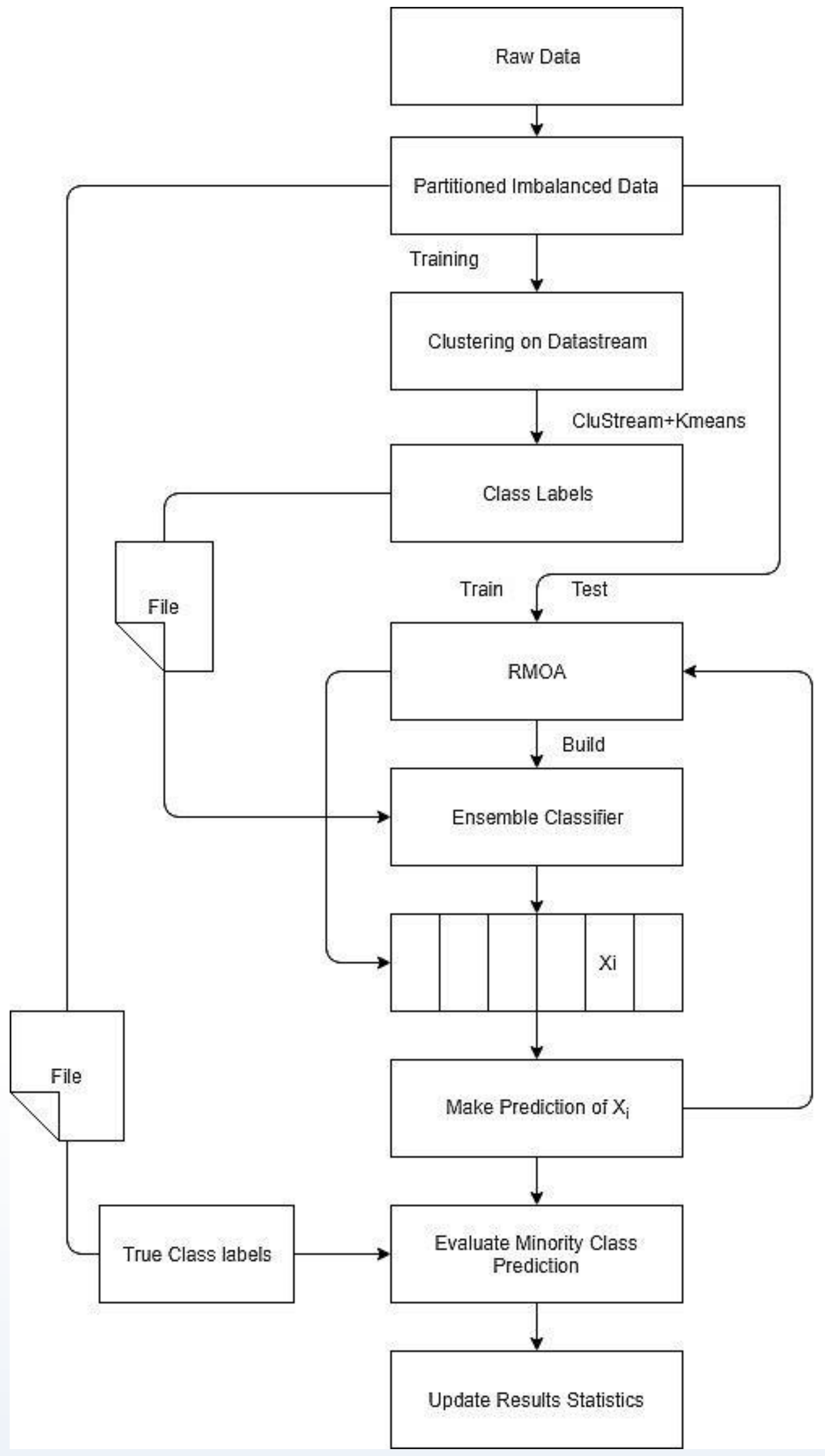
With the introduction of new technologies and applications, the data is generated at a rate much faster than the systems can handle. This data is made available in the form of data streams which is updated continuously. The data streams provide enormous amounts of data and for a brief period of time during which the algorithm has to perform the training and classification. Data streams characterized by these features suffer from skewed distribution. The classifier in such streams is biased towards the majority class and tends to misclassify the instances of the minority class.

The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of most standard learning algorithms.

The proposed method aims to classify the data accurately by increasing various performance measures such as recall, precision, F-score and not only rely on the accuracy of the algorithm.

Architecture

The data set has been divided into two parts. The first part will be used to train the model and the second part will be used as real-time streams input to the model for prediction. The first part of the data is split into three sets: training, cross-validation and test. The 3 sampled data sets are transformed to their respective data streams using the stream package in R.



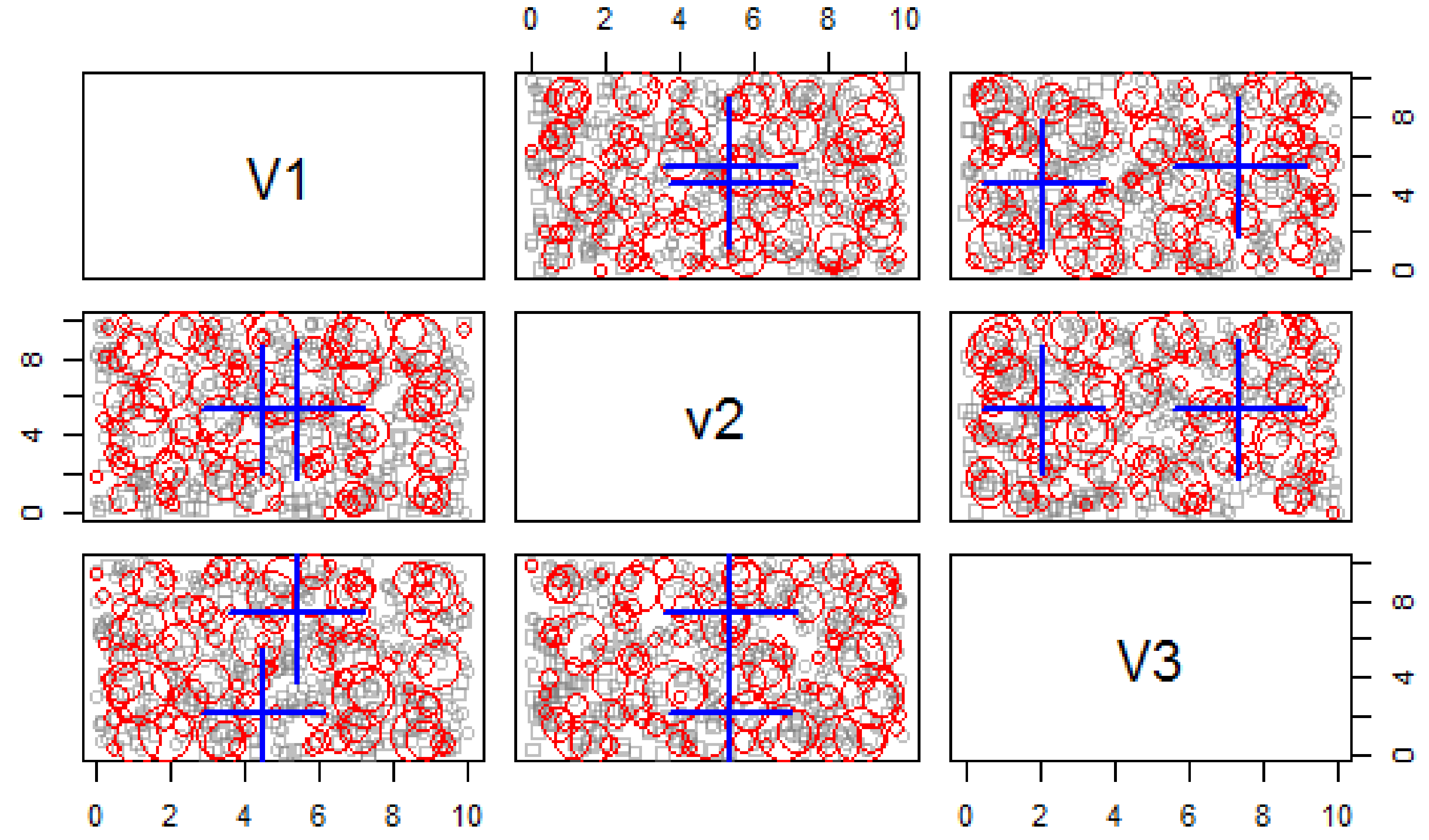
The train stream is then used as input to the clustream + K-means clustering algorithm from the streamMOA package in R. The labels of the input data arrive with some delay and these labels are then compared with the labels assigned by the clustering algorithm to find the actual ones. These assigned classes are then given to the classifier along with the training stream to train the classifier. To handle the issue of imbalance, the algorithmic level approach of ensemble classifier with boosting methodology is used. The classifier predictions are evaluated with the help of true labels on the basis of confusion matrix and F-score.

Algorithm

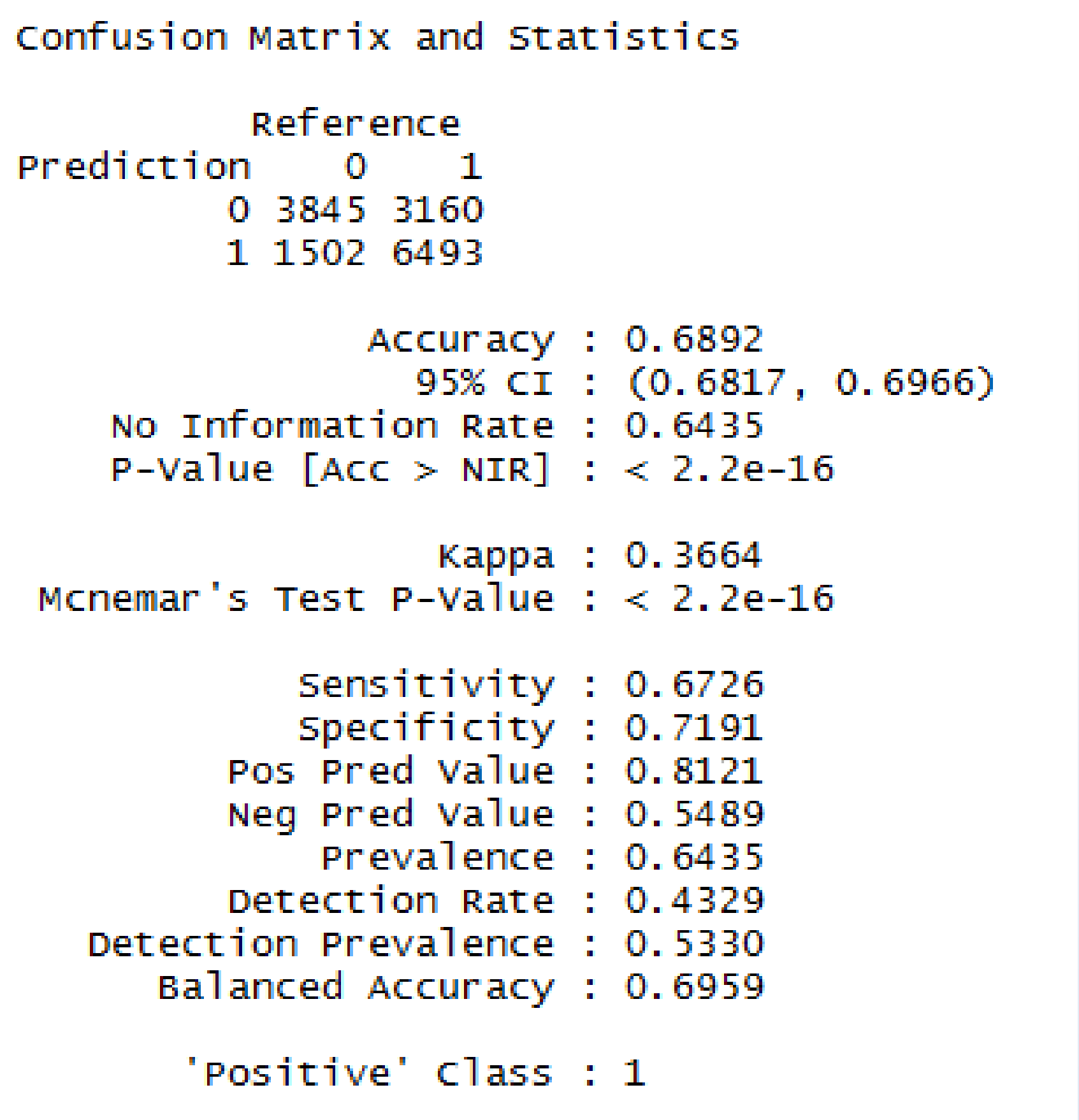
1. Divide the data set into 2 parts (Train and Test) and store the class labels separately as they would be available to the clustering algorithm after a delay.
2. Cluster the train data stream using Clustream + K-means algorithm.
3. Compare the assigned cluster labels with the class labels that arrive after a delay.
4. Train the Ensemble Classifier on train data stream along with cluster labels.
5. Predict the class for the test data stream.
6. Evaluate class predictions on the basis of confusion matrix and F-score using the true labels.

Results

Clustering using Clustream + K-means on sea concepts Imbalanced dataset with 6000 examples over 3 attributes. The red circles indicate micro-clusters obtained through clustream and they are reclustered using K-means to form two macro clusters indicated by "+".

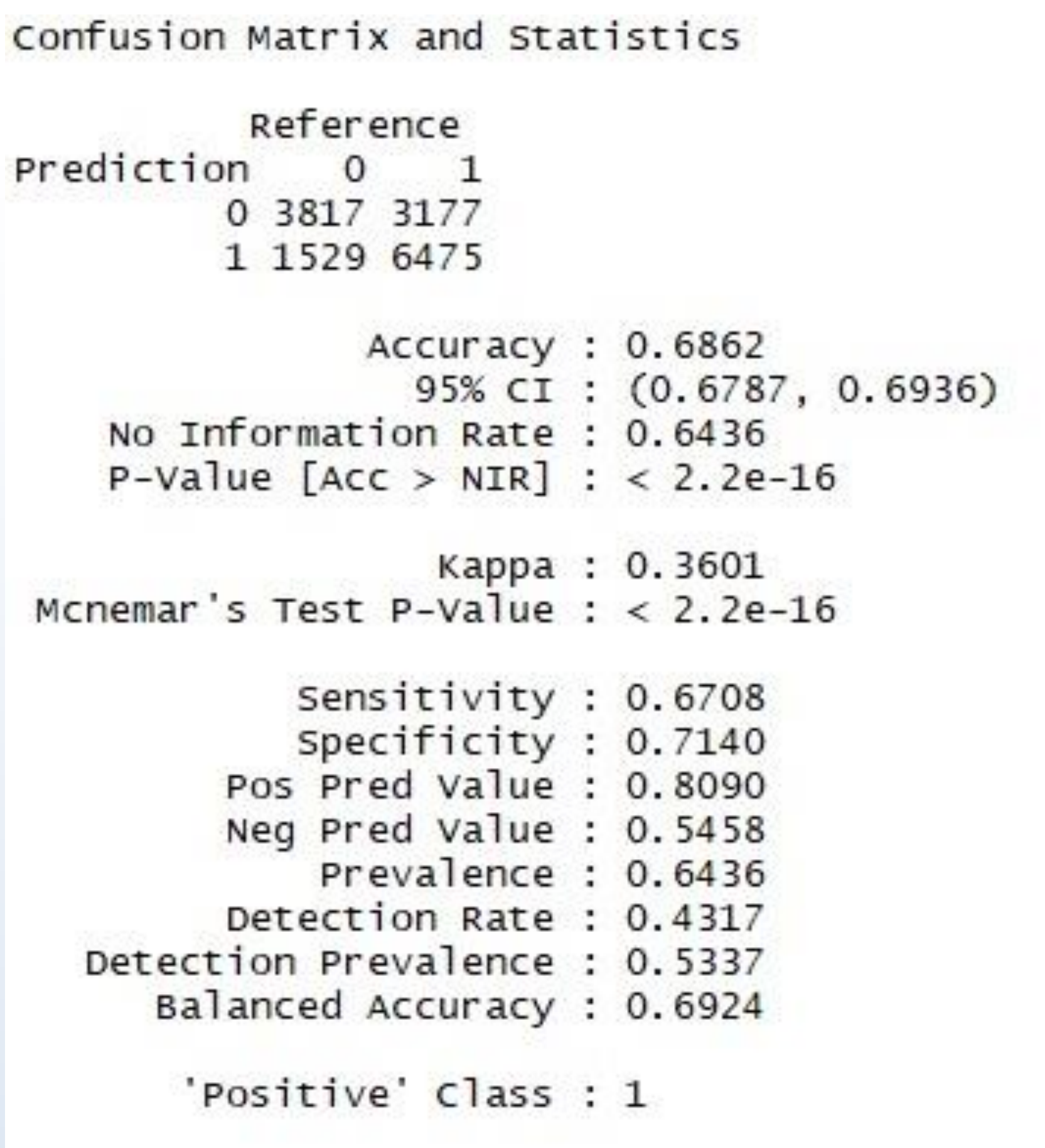


Evaluation Statistics for Naive Bayes Classifier:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3845 3160
         1 1502 6493

               Accuracy : 0.6892
                 95% CI : (0.6817, 0.6966)
    No Information Rate : 0.6435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3664
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6726
            Specificity : 0.7191
         Pos Pred Value : 0.8121
         Neg Pred Value : 0.5489
             Prevalence : 0.6435
         Detection Rate : 0.4329
   Detection Prevalence : 0.5330
      Balanced Accuracy : 0.6959

       'Positive' Class : 1
```

Evaluation Statistics for OzaBoost Ensemble Classifier:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3817 3177
         1 1529 6475

               Accuracy : 0.6862
                 95% CI : (0.6787, 0.6936)
    No Information Rate : 0.6436
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3601
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6708
            Specificity : 0.7140
         Pos Pred Value : 0.8090
         Neg Pred Value : 0.5458
             Prevalence : 0.6436
         Detection Rate : 0.4317
   Detection Prevalence : 0.5337
      Balanced Accuracy : 0.6924

       'Positive' Class : 1
```

Conclusion

The three insights of the problem considered are: class imbalance, data streams and absence of labels. The classifier deals with imbalanced data streams and classifies the data accurately. The ensemble classifier algorithm is used to handle the imbalance data issue. Its performance is judged on the performance metrics like F-score and Confusion matrix.

References

1. Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, Gong Bing, "Learning from class-imbalanced data: Review of methods and applications", International Journal of Expert systems with applications, Elsevier, Dec 2016.
2. Shaza M. Abd Elrahman and Ajith Abraham, "Ä review on class imbalance problem", Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 332-340