Ege Aktan
GE - 461
Project V

Assessments On Different Methodical Approaches Among Data Stream Mining

Introduction

The report you are about the read contains extensive analysis about the methods and approaches made among data stream mining. Three random RBF datasets generated with drift speed 0, 10, 70 respectively (RBF, RBF 10, RBF 70). Methods and approaches utilized on these generated datasets for the assessments made on the accuracy retrieved from the applications.

Three Separate Online Single Classifiers

Three different classifiers, which are Hoeffding Tree (HT), Naive Bayes (NB) and Multi-Layer Perceptron (MLP), applied on three different datasets which have been generated beforehand. For the sake of online learning, models were tested first then trained (partially fitted) accordingly. As the drift speed of the dataset increases, accuracy drops down due to the changing characteristics of the data.

HT and NB retrieves similar accuracy scores (around 75%) on the static dataset (RBF) as MLP fits slightly worse than them (65%). However, as the drift speed increases as on datasets RBF 10 and RBF 70, accuracy score drops in every model and ranges in between 50-65%. On the other hand, MLP responds drifts slightly better than the other models (HT and NB) implemented online. On the other hand, MLP is much more costly compared to others. Even though maximum iteration parameter defined as 2 (by default it is 200), process took much more time to accomplish. Temporal accuracy plots of application can be viewed in the appendix.

Online Ensemble Classifiers

Online ensemble classifiers work in a unison and contribute the prediction by voting their estimated target value in the model. In this section, two different ensemble classifiers conducted; Weighted Majority Voting Rule (WMV) and Majority Voting Rule (MV). MV receives the vote of three individual models and decides on the majority decision. In our case, model needs two or more votes on a target estimation to decide on the prediction. WMV works in a similar manner but at the same time it takes a parameter of weights to differentiate the amount of impact caused by the classifiers within. In our case, accuracy scores retrieved from the previous section (online single classifiers) to determine the weights. Thus, models which fitted more accurately for the specific dataset has a more impact on the output of the ensemble classifier. Another parameter defined for ensemble classifiers is voting type. The parameter takes two input into consideration which are *soft voting* and *hard voting*. In the hard voting implementation models vote their predictions regardless of the estimation probability, either one or zero. However, in soft voting implementation the ensemble classifier collects estimation probabilities from the single classifiers and makes a decision according to the probability predictions. Soft voting works in a more precise manner when compared with hard voting. However, if the single classifiers within the ensemble do not achieve

high accuracy scores, it is not advised to benefit from soft voting. Since that is the case on this project as the accuracy scores strictly less than 75%, hard voting utilized only in this section.

Results of the ensemble classifiers conducted on every dataset retrieved similar results to the single classifiers. Advantageous aspect of utilizing MV is that there is a good chance for better fitting compared to single classifiers (HT, NB and MLP) separately. Since MLP responds better with drifting datasets the ensemble model is a better choice for using HT or NB separately. Similarly, NB and HT is a better fit on the static datasets, implementing all three of them is more preferable rather than using only MLP.

Even though weights were given according to the accuracy scores achieved by the single classifiers, WMV retrieved worse results than MV in terms of accuracy especially on the dataset which has no drift (RBF).

Three Separate Batch Single Classifiers

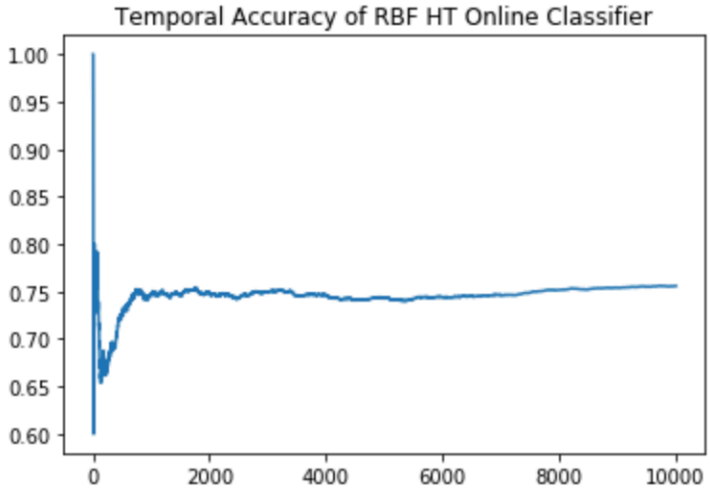
Data conventionally split in two as train set and test set with proportions 70%, 30% respectively. Models fitted on the train set as a whole and measured through the test set. MLP is superior to the other classifiers in terms of accuracy scores as all three of them poorly fits on the drifting datasets.

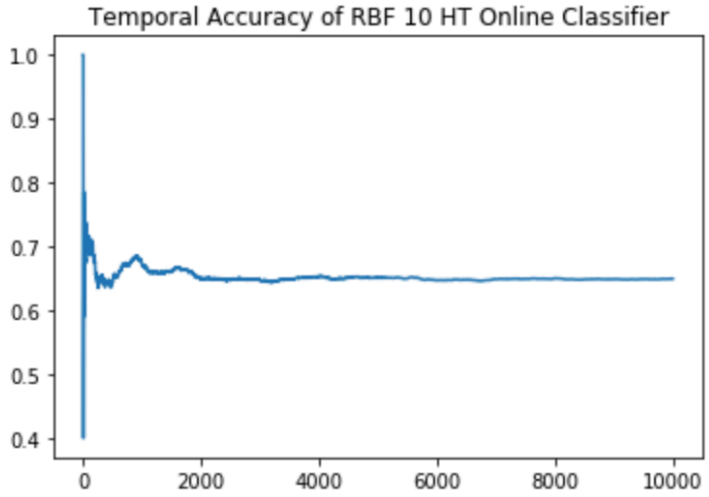
Two Batch Ensemble Classifiers

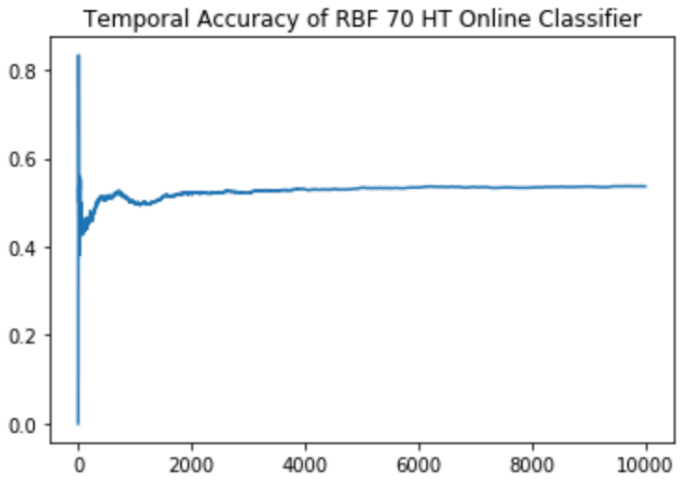
Ensemble classifiers have no significant advantage among individual models. Soft voting provides worse results than hard voting as the drift speed increases on the datasets.

Ege Aktan
GE - 461
Project V

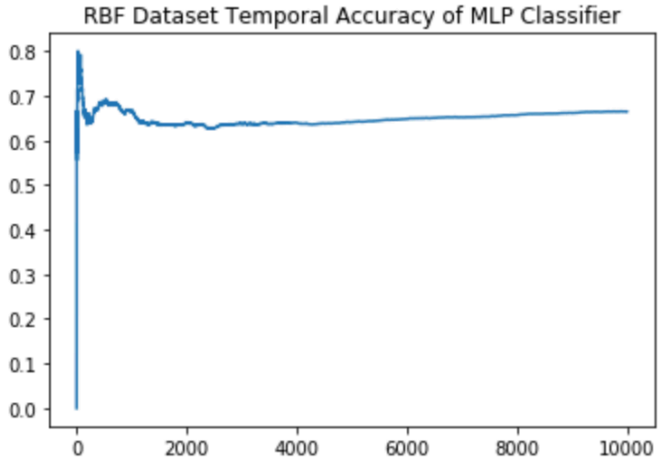
Appendix

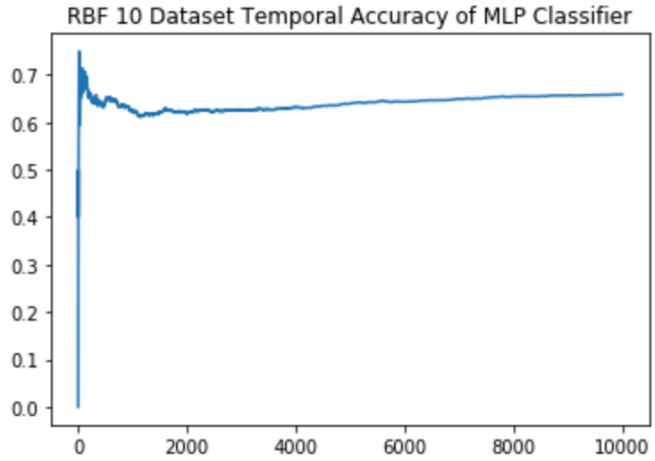


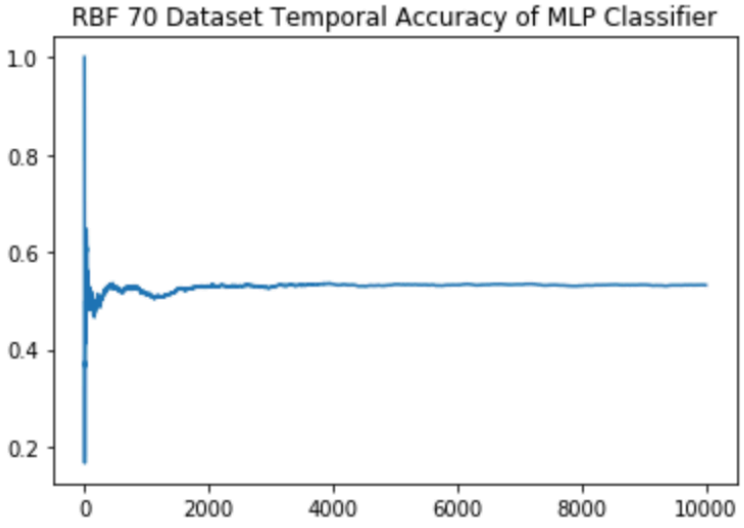


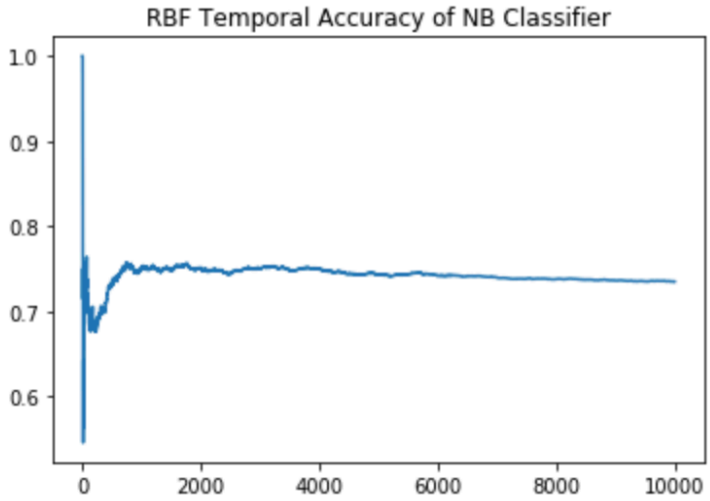


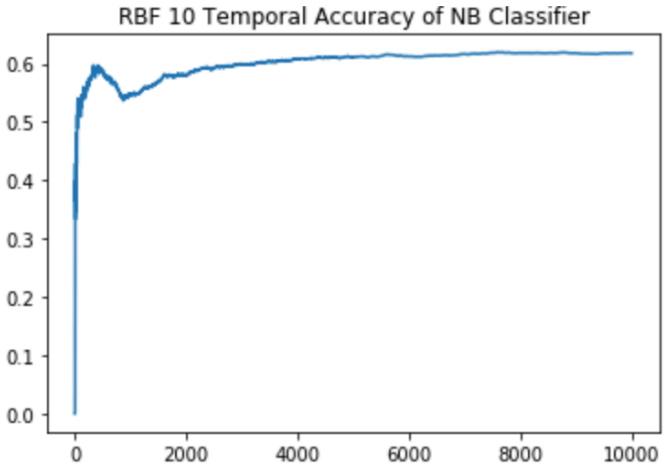
Ege Aktan
GE - 461
Project V

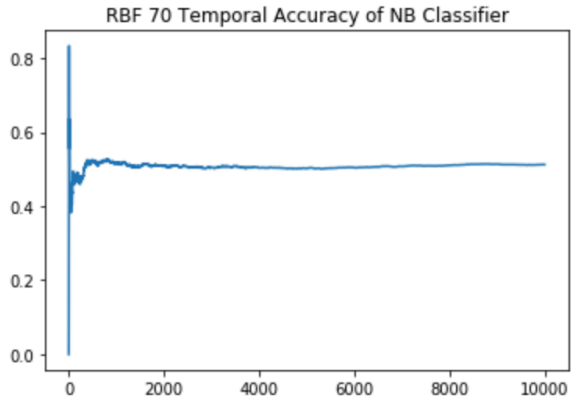







RBF Temporal Accuracy of NB Classifier
1.0
0.9
0.8
0.7
0.6
0
2000
4000
6000
8000
10000


RBF 10 Temporal Accuracy of NB Classifier
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2000
4000
6000
8000
10000


RBF 70 Temporal Accuracy of NB Classifier
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000