Report of Lab Massive Online Analysis

Student : ZHAO Mengzi

1 - Classifiers selected for the experiments

I used different classifiers as below to find the best classifiers for the Covertype dataset : HeterogeneousEnsembleBlastFadingFactors, LimAttClassifier, kNNwithPAWandADWIN, SAMkNN, NoChange, AccuracyWeightedEnsemble, LeveragingBag.

2 - Results of the experiments

Learner : meta.HeterogeneousEnsembleBlastFadingFactors
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 12m24s
Mean of accuracy : 93.02
Mean of Kappa : 85.5
Memory : 4.79

Learner : LimAttClassifier
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 2m23s
Mean of accuracy : 89.14
Mean of kappa : 77.30
Mean of kemory : 0.44

Learner : lazy.kNNwithPAWandADWIN
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 15m55s
Mean of accuracy : 92.83
Mean of kappa : 85.48
Mean of kemory : 4.51

Learner : lazy.SAMkNN
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000

Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 29m25
Mean of accuracy : 93.40
Mean of kappa : 86.21
Mean of kemory : 203.17

Learner : functions.NoChange
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 5.34s
Mean of accuracy: 95.66
Kappa : 91.32
Memory : 0.03

Learner : meta.AccuracyWeightedEnsemble -l (trees.HoeffdingTree -e 1000 -g 100 -c 0.01 -l NB)
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 8m24s
Mean of accuracy : 81.45
Kappa : 62.30
Memory : 0.89

Learner : meta.LeveragingBag
Evaluator : WindowClassificationPerformanceEvaluator
Instance limit : 100,000,000
Sample Frequency : 10,000
Mem check frequency : 100,000
Time : 5m34s
Mean of accuracy: 92.50
Kappa : 85.90
Memory : 2.90

3 - a discussion about the results

We can notice that the classifier has the highest accuracy is nochange and its value of kappa is 91.32, it is also very high comparing to other classifiers, the use of memory is low, it is about 0.53. This classifiers uses little time. I found some information about this classifier to know why its accuracy stays always high. “The nochange classifier needs complete label information, it is a 1-bit finite state machine seeded by the class label. The state machine predicts this class for the next exemplar until there is a change in the exemplar class label. The process then repeats with each change in class

label for the current exemplar being assumed as prediction for the next exemplar." So when there are successive sequence of data in the stream with the same label, the nochange classifier can have a pretty high accuracy.

According to these results, we can also know the classifier AccuracyWeightedEnsemble has the accuracy about 81.5, but its value of kappa is low, it is 62.3. The value of kappa measures the inter-rater agreement for categorical items. It takes more than 8 minutes, comparing to the other classifier, this classifier is not good to use for this dataset, because there are some other classifiers who use less time and who have high accuracy.

We can notice that the classifiers HeterogeneousEnsembleBlastFadingFactors, kNNwithPAWandADWIN, lazy.SAMkNN have high accuray and their values of kappa are high, but they spend too much time. And the classifier LimAttClassifier uses just 2 minutes 23s, it has accuracy about 90, and the value of kappa is not too low, it doesn't use too many memory. The classifier LeveragingBag uses more than 5 minutes, it has very high accuracy, and its value of kappa is high too, the use of memory is OK, it is not too much.

4 - Classifier that I recommend to use with the Covertype dataset

After discussing different results of these classifiers, I recommend the LeveragingBag classifier, as I said in Part 3, the accuracy and the value of kappa of this classifier are really good, this classifier uses about 5m30s on this dataset, comparing to the other classifier, the use of time is not really good but not too bad, also for the use of the memory, it use about 3, it's acceptable too.