IOT STREAMING

LAB
MASSIVE ONLINE ANALYSIS

2017/2018

Master 2 Data&Knowledge

PARIS-SACLAY UNIVERSITY

ZHOU Juncheng

Streaming setting:

Instance Limit: 200,000,000

Sample Frequency: 10,000

Mem Check frequency: 100,000

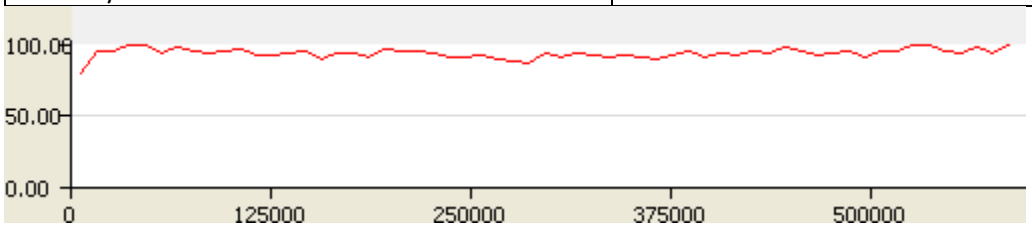
Which classifier

1. HeterogeneousEnsembleBlast (bayes.NaiveBayes,functions.Perceptron,functions.SGD,lazy.kNN,trees.HoeffdingTree)
2. AccuracyUpdatedEnsemble
3. kNNwithPAWandADWIN
4. DynamicWeightedMajority (BaiveBayes)
5. LeveragingBag(HoeffdingTree)

Performance

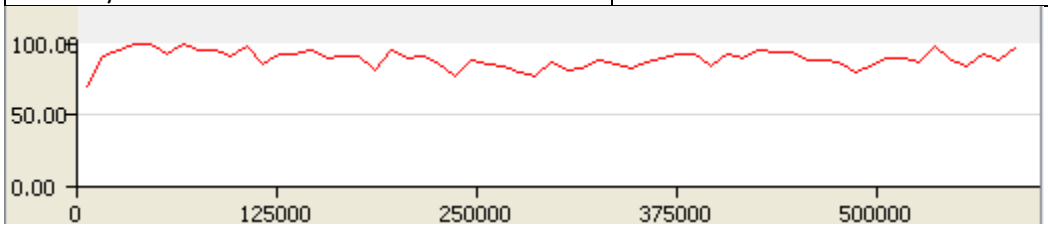
HeterogeneousEnsembleBlast

Measure	Value (Mean)
Accuracy	93
Kappa	85.8
Time	307.93
Memory	4.79



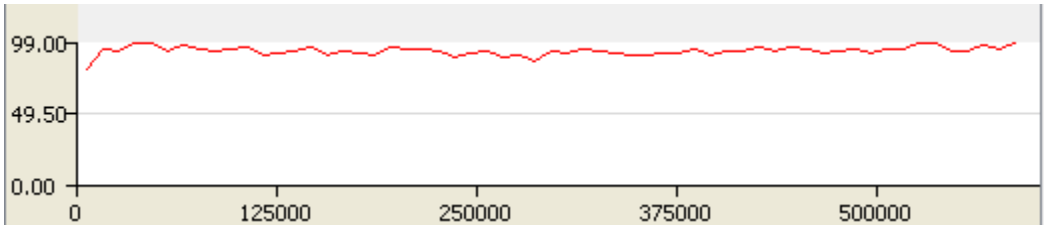
AccuracyUpdatedEnsemble

Measure	Value(Mean)
Accuracy	88.64
Kappa	77.78
Time	71.10
Memory	1.52



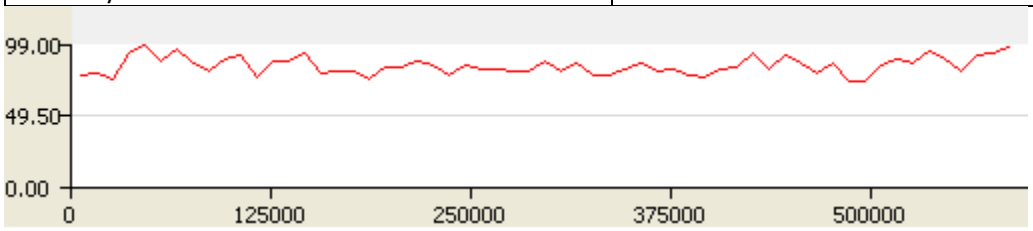
kNNwithPAWandADWIN

Measure	Value(mean)
Accuracy	92.83
Kappa	85.48
Time	370.68
Memory	4.50



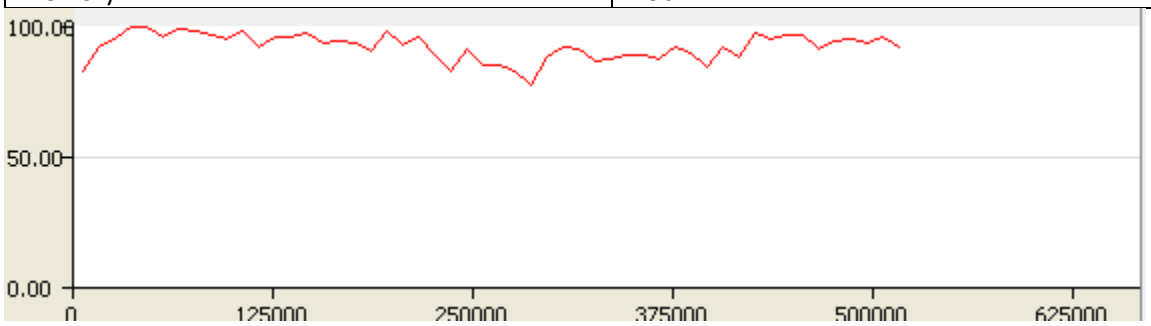
DynamicWeightedMajority

Measure	Value(mean)
Accuracy	83.80
Kappa	68.25
Time	54.75
Memory	0.25



LeveragingBag

Measure	Value(mean)
Accuracy	92.50
Kappa	85.90
Time	84.83
Memory	2.90



Discussion

According to the result of performance. We can find that if we just consider the “Accuracy”, {*LeveragingBag*, *kNNwithPAWandADWIN*, *HeterogeneousEnsembleBlast*} have a good behavior. And at the same time, their Kappa are almost the same, so, their performance is very closed.

But, the { *kNNwithPAWandADWIN* , *HeterogeneousEnsembleBlast* } spent much time to train and predict. LeveragingBag is better than others, and his cost of memory is lower than {kNNwithPAWandADWIN , HeterogeneousEnsembleBlast }.

The only problem of LeveragingBag is the trend of accuracy who is not stable. So maybe some time, the predict is not good. And we can find that the { *kNNwithPAWandADWIN* , *HeterogeneousEnsembleBlast* }’s trend is very stable, so we can always get the best predict at any time.

Conclusion

Based on the analysis, I prefer use { *kNNwithPAWandADWIN* , *HeterogeneousEnsembleBlast* }. Even if they spend much time and take more memory, but they can promise the mode can have a good behavior at any time. Anyway, stable is very import.

So, I will choose *HeterogeneousEnsembleBlast* to train and predict Covertype dataset.