Adversarial effects of intermediate latency in Active Learning on Data Streams

Pedro H. Parreira[1], Ronaldo C. Prati[1]

[1]Federal University of ABC (UFABC), Center of Mathematics, Computing, and Cognition,
Avenida dos Estados 5001, 09210-580, Santo André, SP, Brazil

pedro.parreira@ufabc.edu.br, ronaldo.prati@ufabc.edu.br

Introduction

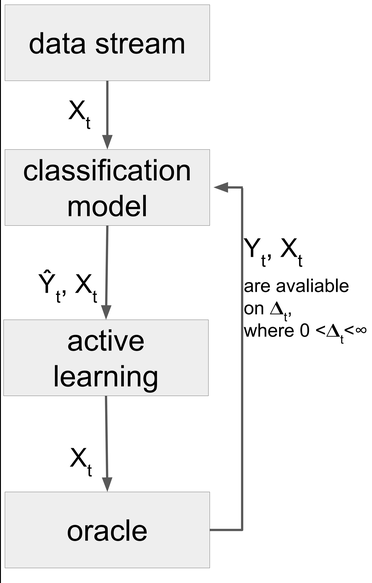
Data Streams (DS) are characterized by the continuous and ininterrupt arrival of instances. As a result a DS produce large volume of data. Due at the high associated cost, it is unlikely that all instances will have their correct labels available for verification in a DS context. In this context, an active learning (AL) approach in DS becomes interesting. The vast majority of AL approaches in DS assumes that a label for a given instance is made available immediately after request. We evaluatesome existing strategies of active learning in datastreams scenarios, with delayed label availability

Data Stream

Data Streams (DS) are characterized by arrival continuous and ininterrupta of instances. Furthermore, in a Data Streams (DS):

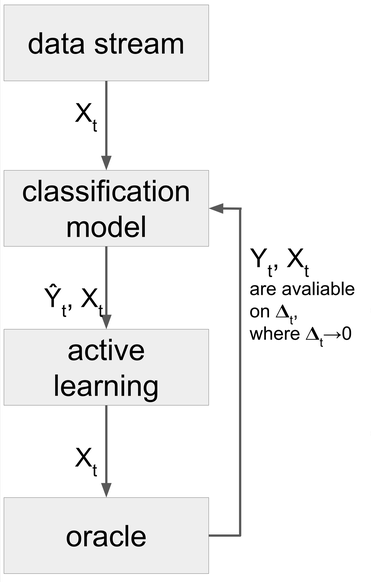
- The distribution of data can change over time, that is, $p_{t0}(x, y) \neq p_{t1}(x, y)$, called *concept drift*.
- The distribution of classes can change over time, which is called *concept evolution*.
- May exist a time interval between the arrival of an instance and its respective label availability, which is called *verification latency*, and the time interval is called *latency*

Active Learning on Data Streams
Scenario of intermediate latency



- Scenario rarely assumed in the AL approaches in DS.
- The label are made available after a interval time after the request.
- The classification model is adapted later with the instance and its label.
- Hypothesis: due to the concept drift, there is an uncertainty in the informativeness of the instances and their label.

Active Learning on Data Streams
Scenario of null latency



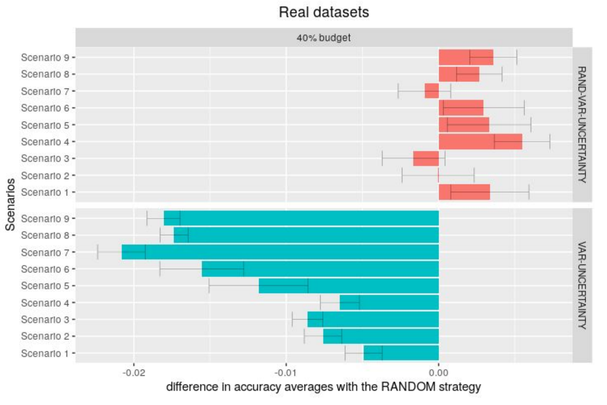
- Scenario assumed in the vast majority of AL approaches in DS.
- The label are made available There no uncertainty in the informativeness of the instances and its label.
- immediately after request.
- The classification model is adapted immediately with the instance and its true label.
- There no uncertainty in the informativeness of the instances and its label.

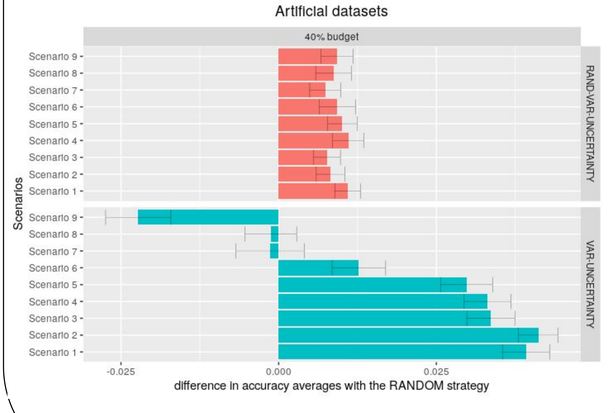
Experiments

- The experiment evaluated three existing strategies of AL, that consider null latency in DS, in the scenario with delayed label availability.
- 3 artificial datasets (SINE, MIXED e STAGGER), in that each dataset have two differents versions according to the type of concept drift (gradual and abrupt), and 2 real datasets (Electricity and Airline).
- 9 latency scenarios (in terms of latency, “Scenario 1” < “Scenario 2” < … < “Scenario 8” < “Scenario 9”)
- The RANDOM, VAR-UNCERTAINTY and RAND-VAR-UNCERTAINTY AL strategies were considered in the experiments. Each strategy can request the label of at most 40% of the instances.

Results

- The results of the experiments were separated by the type of dataset (real and artificial).
- The performance of the VAR-UNCERTAINTY and RAND-VAR-UNCERTAINTY strategies were compared with the RANDOM strategy.
- Each scenario, the mean accuracy obtained of the each strategy was subtracted from the mean accuracy of the RANDOM strategy.
  - In case of the difference is less than zero, the strategy performanced worse than of the RANDOM strategy.
  - Otherwise, the strategy performanced better than of the RANDOM strategy.





Conclusion

- In scenarios with longer latency intervals, the RAND-VAR-UNCERTAINTY strategy, that use the informativeness of the instances combined with a random approach, achieved better performance.
- In scenarios with lower latency interval, the VAR-UNCERTAINTY strategy, that considers only the informativeness of the instances, achieved better performance.
- The results obtained suggest that the informativeness of the instances becomes more uncertain with the increase of the latency interval.