Adversarial effects of intermediate latency in Active Learning on Data Streams

Pedro Henrique Parreira and Ronaldo Cristiano Prati [1]

Federal University of ABC (UFABC)
pedro.parreira@ufabc.edu.br, ronaldo.prati@ufabc.edu.br

Abstract

In several data mining applications, obtaining the actual labels of examples is a costly task. In a data stream scenario, this task becomes even more challenging due to the vast amount of generated data. Therefore, the active learning approach in this scenario becomes a necessity for acquiring labels for continuous model assessment and updating. However, unlike what is assumed in existing approaches to active learning in a data stream, in many real applications, the label of instances can be made available with delay. We evaluate some existing strategies of active learning in data streams scenarios, with delayed label availability.

1. Introduction

Data streams are stochastic processes in which instances arrive continuously, uninterrupted, and independent of each other (Gama, 2010). Due to such characteristics, data streams have larges volumes of data.

Besides, the distribution of data can change over the data stream, that is, $p_{t0}(x, y) \neq p_{t1}(x, y)$, called *concept drift* (Gama et al., 2014). Another characteristic present in data streams is the change in the distribution of classes over time, which is called *concept evolution* (Gama et al., 2014). Frequently acquiring correct labels of some instances is essential to detect these changes and update the model. However, in a data stream, there 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* (Marrs et al., 2010).

Depending on the *latency*, we can obtain three different scenarios: (1) *null latency*, where the label of x_t is available on $t + \Delta_t$, where $\Delta_t \to 0$; (2) *extreme latency*, where the label of x_t is available on $t + \Delta_t$, where $\Delta_t \to \infty$; (3) *intermediate latency*, label of x_t is available on $t + \Delta_t$, where $0 < \Delta_t < \infty$.

However, due to the high cost associated with the obtaining of labels in several machine learning applications (Settles, 2009) and the massive production of data in the data stream, it is unlikely that all instances will have their correct labels available for verification. In this context, an active learning approach in data streams becomes interesting. The active learning approach aims to select a small portion of unlabeled instances available to be labeled by an Oracle (for example, a specialist) and subsequently used to adapt the classification model (Settles, 2009).

The vast majority of active learning approaches in data streams assumes that, when the systems request a label for a given instance, the Oracle returns the correct label immediately, that is, without any delay (Žliobaitė et al., 2014; Mohamad, 2017; Attenberg & Provost, 2011; Zhao & Hoi, 2013; Hao et al., 2018). In (Parreira & Prati, 2019), we address the use of active learning in the data stream with intermediate latency. However, we consider an Oracle with the capacity to label only one instance simultaneously.

In (Žliobaité, 2010), the author questions whether it is possible and when to detect a *concept drift* from delayed labeled data, besides discusses the relationship between delayed labeling and active learning. In (Gomes et al., 2019), the authors list many data stream research opportunities that take into account *verification latency* but do not mention the use of active learning. In (Plasse & Adams, 2016), the authors provide a *framework* that uses a version of the *Linear Discriminant Analysis*(LDA) algorithm in a data streams that can incorporate delayed labels. Furthermore, the paper also provides a taxonomy for the different types of *intermediate latency*.

Žliobaitė et al. (2014) propose different active learning strategies for acquiring labels in data streams. However, the authors consider *null latency*. To evaluate the effect of *intermediate latency*, in this paper, we evaluate these strategies in scenarios with *intermediate latency*.

2. Active Learning Strategies

In (Žliobaitė et al., 2014), a theoretical support framework for active learning in a data stream is described, as well as some active learning strategies that are capable of handling the *concept drift*.

The authors evaluate three strategies for actively selecting instances in data streams: the *RANDOM* approach selects an instance at random. The *VAR-UNCERTAINTY* strategy uses the informativeness of the instances. The *RAND-VAR-UNCERTAINTY* is a hybrid strategy that use the informativeness of the instances combined with a random approach.

3. Results and future work

To get some insight into the impact of intermediate latency in active learning with data streams, in the experiments were used the real-world datasets *Electricity* and *Airline*, in addition to the artificial datasets *SINE*, *MIXED* e *STAGGER*. Each dataset has two possible classes. For each synthetic database, two versions were generated according to the type of *concept drift*: gradual and abrupt.

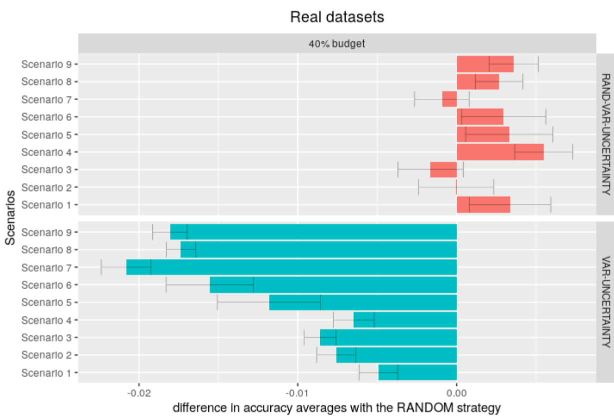

Figure 1. Results obtained for the real datasets

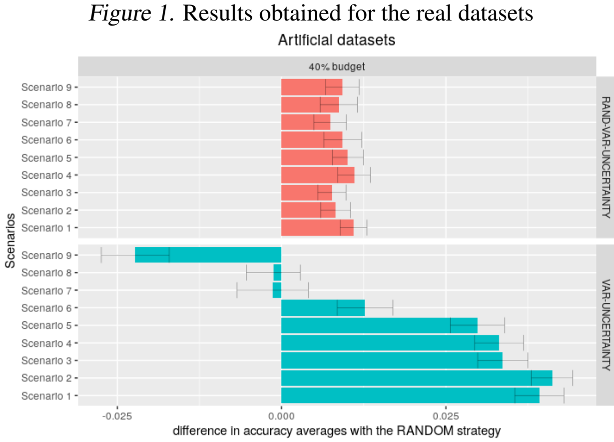

Figure 2. Results obtained for the artificial datasets

In Table 1 are presented the latency scenarios for the datasets. Thus, for example, when considering the dataset *Airline* in *Scenario 5*, the label of the instance belonging to the Class 1 will be made available after the classification model has predicted such a label and, in sequence, the arrival of 600 new instances from the data stream. Furthermore, a 40% budget was established. Therefore, each strategy can request the label of at most 40% of the instances.

	Class 1	Class 2
Scenario 1	50	25
Scenario 2	100	50
Scenario 3	200	100
Scenario 4	300	150
Scenario 5	600	300
Scenario 6	1.200	600
Scenario 7	1.800	900
Scenario 8	2.400	1.200
Scenario 9	3.000	1.500

Table 1. Latency scenarios for the datasets

The performance of the *VAR-UNCERTAINTY* and *RAND-VAR-UNCERTAINTY* strategies was analyzed in comparison with the *RANDOM* strategy. To this end, the average accuracy obtained in each scenario using *VAR-UNCERTAINTY* and *RAND-VAR-UNCERTAINTY* strategies are subtracted from the average accuracy obtained using the *RANDOM* strategy. So, if such a difference is less than zero, the strategy in question achieved a lower performance than the *RANDOM* strategy. Otherwise, the strategy is superior in performance than the *RANDOM* strategy.

Figure 1 shows the results obtained for the real-world datasets. The results show that the impact of increasing the interval of latency is more severe in *VAR-UNCERTAINTY* strategy than *RAND-VAR-UNCERTAINTY* strategy. Figure 2, that depicts the results obtained for the artificial datasets, shows the same pattern.

The results obtained suggest that the informativeness of the instances becomes more uncertain with the increase of the latency interval. Therefore, the *RAND-VAR-UNCERTAINTY* strategy achieves better results in scenarios with longer latency intervals. The *RAND-VAR-UNCERTAINTY* strategy has a random component in its decision criteria, giving less weight to the information of the instances. In contrast, the *VAR-UNCERTAINTY* strategy considers only the informativeness of the instances.

In various real applications of the data stream, the *intermediate latency* is a present problem. Furthermore, due to the massive production of data in the data stream, it is unlikely that all instances will be available with their correct labels. However, few papers address the problem of *intermediate latency* in data streams.

In future work, as the informativeness of the instances becomes uncertain with the increase of the latency interval, we plan to develop new active learning strategies in data streams that consider the cost of obtaining labels and the data that are in process of labeling.

References

Attenberg, J. and Provost, F. Online active inference and learning. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 186–194, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020443. URL http://doi.acm.org/10.1145/2020408.2020443.

Gama, J. *Knowledge discovery from data streams*. CRC Press, 2010.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.

Gomes, H. M., Read, J., Bifet, A., Barddal, J. P., and Gama, J. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter*, 21:6–22, 11 2019. doi: 10.1145/3373464.3373470.

Hao, S., Lu, J., Zhao, P., Zhang, C., Hoi, S. C. H., and Miao, C. Second-order online active learning and its applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(7):1338–1351, July 2018. ISSN 1041-4347. doi: 10.1109/TKDE.2017.2778097.

Marrs, G. R., Hickey, R. J., and Black, M. M. The impact of latency on online classification learning with concept drift. In Bi, Y. and Williams, M.-A. (eds.), *Knowledge Science, Engineering and Management*, pp. 459–469, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15280-1.

Mohamad, S. *Active Learning for Data Streams*. PhD thesis, Bournemouth University, 2017.

Parreira, P. and Prati, R. Active learning in data stream with intermediate latency. In *XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Salvador, 10 2019.

Plasse, J. and Adams, N. Handling delayed labels in temporally evolving data streams. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2416–2424, Dec 2016. doi: 10.1109/BigData.2016.7840877.

Settles, B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Zhao, P. and Hoi, S. C. Cost-sensitive online active learning with application to malicious url detection. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 919–927, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487647. URL http://doi.acm.org/10.1145/2487575.2487647.

Žliobaitė, I. Change with delayed labeling: When is it detectable? In *2010 IEEE International Conference on Data Mining Workshops*, pp. 843–850, 2010.

Žliobaitė, I., Bifet, A., Pfahringer, B., and Holmes, G. Active learning with drifting streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1):27–39, Jan 2014. ISSN 2162-237X. doi: 10.1109/TNNLS.2012.2236570.