Adversarial effects of intermediate latency in Active Learning on Data Streams

Pedro Henrique Parreira and Ronaldo Cristiano Prati
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 large volumes of data.

Besides, the distribution of data can change over the data stream, that is, \( p_0(x, y) \neq p_1(x, y) \), 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 obtaining 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 (Zliobaitė 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 (Zliobaitė, 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 stream that can incorporate delayed labels. Furthermore, the paper also provides a taxonomy for the different types of intermediate latency.

Zliobaitė 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 (Zliobaitė 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.

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.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>300</td>
<td>150</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>600</td>
<td>300</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>1.200</td>
<td>600</td>
</tr>
<tr>
<td>Scenario 7</td>
<td>1.800</td>
<td>900</td>
</tr>
<tr>
<td>Scenario 8</td>
<td>2.400</td>
<td>1.200</td>
</tr>
<tr>
<td>Scenario 9</td>
<td>3.000</td>
<td>1.500</td>
</tr>
</tbody>
</table>

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


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.


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.


