Efficient online tree, rule-based, and distance-based algorithms

Saulo Martiello Mastelini[1], **André Carlos Ponce de Leon Ferreira de Carvalho**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (ICMC-USP)
400, Trabalhor São Carlense av., São Carlos - SP, 13566-590

mastelini@alumni.usp.br, andre@icmc.usp.br

Abstract. *The fast development of technology resulted in the constant production of data in different forms and from different sources. Contrary to what was observed in the first machine learning (ML) research works, there might be too much data to handle with traditional algorithms. Changes in the underlying data distributions might also render traditional ML solutions useless in real-world applications. Online ML (OML) aims to create solutions able to process data incrementally, with limited computation resource usage, and to deal with time-changing data distributions. Unfortunately, we have seen a recent growing trend in creating OML algorithms that solely focus on predictive performance and overlook computational costs. In regression tasks, the problem is even more pronounced when considering some of the most popular OML solutions: decision trees, decision rules, and ensembles of these models. In this thesis, we created improved and efficient OML algorithms from the mentioned algorithmic families by focusing on decreasing time and memory costs while keeping competitive predictive performance. Our proposals are either novel standalone OML algorithms or additions that can be paired with any existing tree or decision rule regressors.*

1. Introduction

In the last decades, we have seen a steep increase in the produced and monitored data [Bahri et al. 2021, Palli et al. 2024]. As the years passed, rich and varied data sources concerning different real-world phenomena became widely available. With the prevalence of data and the development of technology concerns about data privacy have arisen[1,2], as most of the gadgets and digital tools we use nowadays are also data sensors. In response to the data availability arising and the infrastructure and hardware aspects related to this fact, computational and pattern recognition tools also had to evolve [Gama 2010, Russell and Norvig 2016].

Traditional Machine Learning (ML) algorithms were originally intended to use all available data for training and, thus, may not be able to deal with the huge amount of incoming training data [Gama 2010, Bifet et al. 2018]. Indeed, the fact data nowadays arrives continuously as a stream led to the creation of incremental or online ML algorithms. Online ML (OML) seeks to create learning models that are trained incrementally, that is, one instance at a time. These algorithms also operate under constrained

[1]https://gdpr-info.eu/

[2]https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

computational resources due to the potentially unbounded nature of their input data. Ideally, an OML algorithm should only process each datum once and then discard that input [Bifet and Gavaldà 2009]. By doing so, the learning model must be able to extract and store efficiently any useful information that may be extracted from the inputs. It must also be ready to provide predictions at any time. As a direct consequence, reducing the time requirements of OML solutions is also of paramount importance. That is, besides using a limited amount of memory OML models should be able to process data as fast as possible so that the incoming data streams can be processed in their entirety.

In the last two decades, plenty of OML solutions were designed to deal with the mentioned aspects of data stream mining. Nonetheless, in recent years we have seen only small changes and improvements in the available solutions with not nearly as many breakthroughs as the first OML algorithms offered in comparison to their original, batch-based, traditional ML counterparts. Most of the recent new OML algorithms focus mostly on predictive performance alone. To exacerbate the situation, in many cases, the slight obtained predictive performance improvements come with a steep increase in memory and time costs [Korycki and Krawczyk 2020, Cano and Krawczyk 2022]. Prior to the author's thesis, this situation was even more pronounced for regression, as there were no known effective solutions to mitigate training costs in these tasks, unlike the solutions already proposed for classification tasks [Pfahringer et al. 2008].

The thesis did not follow the same trend, actually taking the opposite path by exploring ways to reduce the memory and time costs of tree, rule-based, and distance-based OML algorithms. By doing so, we believe we can leverage the best recent developments in the OML research field, while also improving their efficiency and potential for application in real-world learning tasks. To achieve the desired reduction in computational costs we employed multiple strategies to improve existing OML algorithms, and we also devised a new accurate and efficient ensemble regressor [Mastelini et al. 2022]. Most prominently we have focused on the family of Hoeffding bound-based [Hoeffding 1963] regressors, namely, Hoeffding Trees (HT), decision rules, and ensembles thereof. By the end of the thesis development, we also tackled the computational cost reduction targeting distance-based algorithms. More specifically, we focused on incremental k-Nearest Neighbors algorithms that keep a data window with the most recent observed data stream elements, regardless of the final learning task, e.g., classification and regression.

The next sections are organized as follows. Section 2 presents the problem tackled in the thesis, as well as the objective, research questions, and hypothesis that guided all the research development. Section 3 presents the related work upon which the research was built. Last, we summarize our main contributions to the research and application fields in Section 4.

2. Problem, objective, research questions, and hypotheses

Although many new streaming-based OML solutions have been created, they mostly focus on predictive performance and often overlook the needed computational resources. This becomes more apparent when considering resource-constrained applications and reducing the environmental impacts related to energy usage [García-Martín et al. 2019, García Martín 2020]. By surveying the available literature, we identified the following specific problems related to the scope of the thesis.

Problem: most existing HT regressors (HTR) and ensembles thereof were too computationally costly for real-world applications. Besides, in the specific case of sliding window-based k-NN-based algorithms, we realized that the existing solutions perform a complete scan of the data buffer elements, which might become impractical as the amount of stored examples increases.

Hence, we envisioned a central principle that guided the whole development of the thesis.

Objective: the main objective of the thesis was **to develop efficient tree and rule-based incremental regressors, and distance-based algorithms.**

We proposed the following research questions to guide our research during the development of the thesis:

Q1: Can the computational costs of HTRs be reduced without significant impacts on predictive performance?

Q2: Can efficient online tree-based ensembles be created while keeping competitive predictive performance to state-of-the-art solutions?

Q3: Is there an efficient and alternative strategy to perform nearest neighbor search queries on a data buffer that is constantly updated using a first-in, first-out data ingestion policy?

Given the research questions and the literature review, we formulated the following hypotheses to pursue further:

(Hyp. 1) *The use of summarization and data sampling techniques can significantly reduce computation costs for building incremental trees and decision rule-based regressors, without significantly impacting prediction accuracy.*

(Hyp. 2) *Efficient and incremental graph-based search structures can be created to perform nearest neighbor search queries using arbitrary distance measures.*

In the thesis, we used **Hyp. 1** to answer **Q1** and **Q2**, and **Hyp. 2** as the starting point to address **Q3**.

At the beginning of the thesis developments, we performed an in-depth review of existing literature to obtain a broad view of the research field. This was a continuous and ongoing action as time progressed and we developed our research contributions. During the thesis development, we performed several exploratory analyses of sampling, data compression, and other statistical and computation tools to fulfill our objective.

3. Related work

HTs, the main base algorithm for thesis development, work by creating a single root node and posteriorly creating additional decision nodes and leaves [Ikonomovska et al. 2011b]. To do so, HTs keep structures called attribute observers (AO) [Mastelini and de Carvalho 2021] to monitor input statistics. Namely, each leaf node in an HT carries one AO per input feature. Categorical features are easier to monitor due to their inherent discrete nature. Continuous features, i.e., real numbers, are harder to keep track of as there are no predefined partitions. HTRs and decision rule regressors work by monitoring how each input contributes to reducing the overall variance in the target variable [Ikonomovska et al. 2011b, Ikonomovska et al. 2011a, Duarte et al. 2016].

The input feature and cut point combination that provides the maximum Variance Reduction (VR) are chosen to create splits and expand the tree or decision rule structure. Hence, regression AOs use incremental variance estimators to evaluate the VR for each monitored split candidate.

There were more efforts to improve the original HTs [Domingos and Hulten 2000] in classification tasks [Pfahringer et al. 2008] compared to regression. Indeed, the usual configuration of HT classifiers nowadays has a $O(1)$ cost to both monitor a new instance and to query split candidate points. Most of the research effort has been put toward reducing the predictive error, often overlooking training costs [Ikonomovska et al. 2015, Osojnik et al. 2018, Gomes et al. 2018].

The original version of HTR [Ikonomovska et al. 2011b] introduced a binary search tree (BST) structure to monitor tree-splitting statistics. This algorithm, dubbed Extended BST (E-BST) also had a mechanism to deactivate nodes that did not hold promising split points. Still, a BST has a $O(n)$ memory and split point query costs and an $O(\log n)$ cost per point insertion. The insertion cost, in the worst-case scenario, can become $O(n)$ when the input data is already ordered. There were also attempts to limit the number of stored BST nodes to reduce computation costs [Duarte and Gama 2015, Duarte et al. 2016]. An improved version of E-BST was also proposed to reduce the number of stored nodes [Osojnik 2017].

Nonetheless, none of the proposed solutions to alleviate the E-BST costs change its asymptotic costs. In the thesis, we propose several alternatives to the original and modified versions of the E-BST algorithm. Our proposed AOs effectively reduce the asymptotic costs involved in building HTRs, regarding the memory requirements and the running time. Naturally, our proposed AOs can be directly applied to Hoeffding bound-based decision rule algorithms and tree ensembles.

Lastly, in the realm of distance-based algorithms, we leverage existing state-of-the-art graph search indices [Dong et al. 2011, Shimomura et al. 2021] to create a novel, efficient, and general-purpose incremental structure to perform nearest neighbor search queries [Mastelini et al. 2024]. To the best of our knowledge, the proposed solution is the first of its kind available for incremental nearest-neighbor search problems that can work with arbitrary distance metrics.

4. Main contributions

Our main contributions, detailed in the thesis' chapters, were mainly reflected by papers published in renowned scientific diffusion venues, as we list next:

1. We developed strategies to speed up HTR construction and save memory costs while also keeping competitive predictive performance compared to the original HTR [Mastelini and Ponce de Leon Ferreira de Carvalho 2020, Mastelini and de Carvalho 2021, Mastelini et al. 2021];
2. We introduced a robust incremental variance estimator to use in the AOs which is much more accurate than the one used in the original HTRs [Mastelini and de Carvalho 2021], and is not prone to numerical cancellation issues as the preceding estimator. Our variance estimator and the related combination formulae are based upon the renowned Welford algorithm [Knuth 2014];

3. We proposed a novel tree-based ensemble regression algorithm that leverages, among other aspects, sub-bagging to significantly speed up the training step and reduce the memory costs while also boosting predictive performance [Mastelini et al. 2022];
4. We proposed an algorithm to efficiently perform nearest neighbor search queries in a sliding window while also being able to handle frequent element addition and removal [Mastelini et al. 2024];
5. In collaboration with researchers from multiple countries, the thesis author created the River [Montiel et al. 2021] Python library for OML, from which he is an active maintainer. River is one of the most popular tools for researchers and practitioners who want to develop OML solutions.

Additionally, the research development allowed us to contribute with researchers from multiple countries, among which we highlight the following:

- Prof. Dr. João Gama (University of Porto - Portugal);
- Prof. Dr. Albert Bifet (University of Waikato - New Zealand);
- Prof. Dr. Heitor Murilo Gomes (Victoria University of Wellington - New Zealand);
- Dr. Max Halford (Carbon Fact - France);
- Prof. Dr. Jesse Read (École Polytechnique, Institut Polytechnique de Paris - France);
- Prof. Dr. Celine Vens (Faculty of Medicine at KULAK, KU Leuven - Belgium);
- Prof. Dr. Sylvio Barbon Jr. (Department of Engineering and Architecture, University of Trieste - Italy);
- Prof. Dr. Bruno Veloso (University of Porto - Portugal);
- Prof. Dr. Rita P. Ribeiro (University of Porto - Portugal);
- Prof. Dr. Edgar Dutra Zanotto (Federal University of São Carlos - Brazil).

Among many other collaborations with researchers, fellow Ph.D. and Master's students from countries such as Brazil, Portugal, Vietnam, Cameroon, Iran, Egypt, Belgium, Italy, and India, to name a few.

In the final chapter of the thesis, we also outline multiple possible paths for future research and expansion of our work. The active participation in the River project is a constant inspiration to pursue further the original objectives envisioned during the thesis development. The community of users and researchers that gathered around the River project is also an effective source of information to better understand the needs of data stream mining practitioners and the research gaps that still need to be fulfilled.

Acknowledgments

The authors would like to acknowledge the financial support provided by FAPESP via the grants #2018/07319-6 and #2021/10488-7. The authors would also like to thank Professor João Gama for providing valuable guidance during the internship performed by the thesis author at the University of Porto, Portugal.

References

Bahri, M., Bifet, A., Gama, J., Gomes, H. M., and Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3):e1405.

Bifet, A. and Gavaldà, R. (2009). Adaptive learning from evolving data streams. In *International Symposium on Intelligent Data Analysis*, pages 249–260. Springer.

Bifet, A., Gavaldà, R., Holmes, G., and Pfahringer, B. (2018). *Machine Learning for Data Streams with Practical Examples in MOA*. MIT Press. `https://moa.cms.waikato.ac.nz/book/`.

Cano, A. and Krawczyk, B. (2022). Rose: robust online self-adjusting ensemble for continual learning on imbalanced drifting data streams. *Machine Learning*, pages 1–39.

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 71–80, Boston, MA, USA. ACM.

Dong, W., Moses, C., and Li, K. (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586.

Duarte, J. and Gama, J. (2015). Multi-target regression from high-speed data streams with adaptive model rules. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10, Campus des Cordeliers, Paris, France. IEEE.

Duarte, J., Gama, J., and Bifet, A. (2016). Adaptive model rules from high-speed data streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):1–22.

Gama, J. (2010). *Knowledge discovery from data streams*. Chapman and Hall/CRC.

García Martín, E. (2020). *Energy Efficiency in Machine Learning: Approaches to Sustainable Data Stream Mining*. PhD thesis, Blekinge Tekniska Högskola.

García-Martín, E., Rodrigues, C. F., Riley, G., and Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75–88.

Gomes, H. M., Barddal, J. P., Ferreira, L. E. B., and Bifet, A. (2018). Adaptive random forests for data stream regression. In *26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*.

Ikonomovska, E., Gama, J., and Džeroski, S. (2011a). Incremental multi-target model trees for data streams. In *Proceedings of the 2011 ACM symposium on applied computing*, pages 988–993. ACM.

Ikonomovska, E., Gama, J., and Džeroski, S. (2011b). Learning model trees from evolving data streams. *Data mining and knowledge discovery*, 23(1):128–168.

Ikonomovska, E., Gama, J., and Džeroski, S. (2015). Online tree-based ensembles and option trees for regression on evolving data streams. *Neurocomputing*, 150:458–470.

Knuth, D. E. (2014). *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional.

Korycki, Ł. and Krawczyk, B. (2020). Adaptive deep forest for online learning from drifting data streams. *arXiv preprint arXiv:2010.07340*.

Mastelini, S. M. and de Carvalho, A. C. P. d. L. F. (2021). Using dynamical quantization to perform split attempts in online tree regressors. *Pattern Recognition Letters*, 145:37–42.

Mastelini, S. M., Montiel, J., Gomes, H. M., Bifet, A., Pfahringer, B., and de Carvalho, A. C. (2021). Fast and lightweight binary and multi-branch Hoeffding Tree Regressors. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 380–388. IEEE.

Mastelini, S. M., Nakano, F. K., Vens, C., de Leon Ferreira, A. C. P., et al. (2022). Online extra trees regressor. *IEEE Transactions on Neural Networks and Learning Systems*.

Mastelini, S. M. and Ponce de Leon Ferreira de Carvalho, A. C. (2020). 2cs: correlation-guided split candidate selection in hoeffding tree regressors. In *Brazilian Conference on Intelligent Systems*, pages 337–351. Springer.

Mastelini, S. M., Veloso, B., Halford, M., de Leon Ferreira, A. C. P., Gama, J., et al. (2024). Swinn: Efficient nearest neighbor search in sliding windows using graphs. *Information Fusion*, 101:101979.

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H. M., Read, J., Abdessalem, T., and Bifet, A. (2021). River: machine learning for streaming data in python. *Journal of Machine Learning Research*, 22(110):1–8.

Osojnik, A. (2017). *Structured output prediction on data streams*. PhD thesis, Ph. D. thesis, Joºef Stefan International Postgraduate School.

Osojnik, A., Panov, P., and Džeroski, S. (2018). Tree-based methods for online multi-target regression. *Journal of Intelligent Information Systems*, 50(2):315–339.

Palli, A. S., Jaafar, J., Gilal, A. R., Alsughayyir, A., Gomes, H. M., Alshanqiti, A., and Omar, M. (2024). Online machine learning from non-stationary data streams in the presence of concept drift and class imbalance: A systematic review. *Journal of Information and Communication Technology*, 23(1):105–139.

Pfahringer, B., Holmes, G., and Kirkby, R. (2008). Handling numeric attributes in hoeffding trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 296–307. Springer.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Shimomura, L. C., Oyamada, R. S., Vieira, M. R., and Kaster, D. S. (2021). A survey on graph-based methods for similarity searches in metric spaces. *Information Systems*, 95:101507.