Machine Learning for Streaming Data
Tutorial IJCAI 2024

Albert Bifet[1,3] , **Bernhard Pfahringer**[1] , **Heitor Murilo Gomes**[2] , **Nuwan Gunasekara**[1]
[1]AI Institute, University of Waikato, +64478384704.
[2]Victoria University of Wellington, +6448873992.
[3]LTCI, Télécom Paris, IP Paris
abifet@waikato.ac.nz, bernhard@waikato.ac.nz, heitor.gomes@vuw.ac.nz, ng98@students.waikato.ac.nz

Abstract

Machine Learning for Data Streams (MLDS) has been an important area of research since the late 1990s, and its usage in industry has grown significantly over the last few years. However, there is still a gap between the cutting-edge research and the tools that are readily available, which makes it challenging for practitioners, including experienced data scientists, to implement and evaluate these methods in this highly complex domain. Our tutorial aims to bridge this gap with a dual focus. We discuss advanced research topics, such as partially delayed labeled streams, while providing practical demonstrations of their implementation and assessment using Python. By catering to both researchers and practitioners, the tutorial aims to empower them to design and conduct experiments, and extend existing methodologies.

1 Tutorial Description for Conference Registration Brochure

Machine learning for data streams (MLDS) has been a significant research area for the last couple of decades, with increasing adoption in the industry over the past few years due to the emergence of Industry 4.0, where more and more processes are monitored digitally. This tutorial delves into critical, fundamental, and challenging aspects of learning from data streams, such as detecting and adapting to concept drifts, data stream classification and regression, and learning from partially/delayed/unlabeled data streams.

2 Tutorial Description for a Web Page Overview

Machine learning for data streams (MLDS) attempts to extract knowledge from a stream of non-IID data. It has been a significant research area since the late 1990s, with increasing adoption in the industry over the past few years due to the emergence of Industry 4.0, where more industry processes are monitored online. Practitioners are presented with challenges such as detecting and adapting to concept drifts, continuously evolving models, and learning from partially labeled and unlabeled data.

Despite commendable efforts in open-source libraries, a gap persists between pioneering research and accessible tools, presenting challenges for practitioners, including experienced data scientists, in implementing and evaluating methods in this complex domain. Our tutorial addresses this gap with a dual focus. We discuss advanced research topics, such as partially delayed labeled streams, while providing practical demonstrations of their implementation and assessment using Python. By catering to both researchers and practitioners, this tutorial aims to empower users in designing, conducting experiments, and extending existing methodologies.

3 Proposed Length of the Tutorial

1/2 day (two 1:45h slots)

4 Goals and Objectives

In this tutorial, our objective is to familiarize attendees with applying diverse machine-learning tasks to streaming data. Beyond an introductory overview, where we delineate the learning cycle of typical supervised learning tasks, we steer our focus towards pertinent challenges seldom addressed in conventional tutorials, such as:

- **Prediction Intervals** for regression tasks
- **Concept drift** detection, visualization and evaluation
- Modelling and addressing **partially and delayed labeled data streams** using semi-supervised and active learning
- The idiosyncrasies of applying and evaluating **clustering** on a data stream

5 Outline

This proposal is for a 3 1/2 hours session. The first part includes an introduction to data stream learning, supervised learning and concept drifts, thus it is the longest with 1 hour and 40 minutes in total. We save 20 minutes for questions and discussion (roughly 5 minutes after each part).

1. **Part 1: Machine Learning for Data Streams**
   - Supervised learning (50 minutes + 10 minutes demo)
     - Learning cycle

  - Evaluation procedures
  - Classification algorithms
    - Incremental decision trees
    - Ensemble methods [Gunasekara *et al.*, 2024]
  - Regression algorithms [Sun *et al.*, 2022]
  - Prediction Intervals [Sun *et al.*, 2024]
  - Practical examples
- QA (5 minutes)
- Concept drifts (30 minutes + 10 minutes demo)
  - Basic concepts (definitions and categorizations)
  - Detecting concept drifts
  - Evaluating and visualizing detections
  - Using drift detectors in beyond data streams
  - Practical examples

2. **Part 2: Unsupervised learning (Clustering and Drift Detection)**
   - Unsupervised drift detection (20 minutes)
     - Assumptions and limitations
     - Algorithms
       - STUDD [Cerqueira *et al.*, 2023]
     - Practical examples
   - QA (5 minutes)
   - Clustering (20 minutes + 10 minutes demo)
     - Micro-clusters
     - Online and offline steps
     - Algorithms
       - CluStream
     - Evaluation procedures
     - Practical examples
3. **Part 3: Partially and delayed labeled data (30 minutes + 10 minutes discussion)**
   - Assumptions
   - Semi-supervised learning
   - Active learning
   - Algorithms
     - Cluster-and-label [Le Nguyen *et al.*, 2019]
   - Evaluation procedures
   - Practical examples
4. QA (10 minutes)

5.1 Part 1: Machine Learning for Data Streams

We will introduce the basic supervised setting for data stream classification. This includes evaluation procedures, the learning cycle, and some key algorithms (incremental trees and ensembles). We will also discuss prediction intervals. These intervals enhance the interpretability and confidence of regression algorithms. Furthermore, we discuss concept drift detection, differing from the common practice of treating drift detectors as a mere mechanism of an adaptive learning algorithm. We'll discuss how drift detection extends beyond streaming data, enabling the assessment of traditional ML pipeline deterioration.

Key references

- Issues in evaluation of stream learning algorithms [Gama *et al.*, 2009]
- Efficient online evaluation of big data stream classifiers [Bifet *et al.*, 2015]
- A survey on ensemble learning for data stream classification [Gomes *et al.*, 2017]
- Machine learning for data streams: with practical examples in MOA [Bifet *et al.*, 2017]
- Scikit-multiflow: A multi-output streaming framework [Montiel *et al.*, 2018]
- River: machine learning for streaming data in python [Montiel *et al.*, 2021]
- SOKNL: A novel way of integrating K-nearest neighbours with adaptive random forest regression for data streams [Sun *et al.*, 2022]
- Adaptive Prediction Interval for Data Stream Regression [Sun *et al.*, 2024]
- Gradient Boosted Trees for Evolving Data Streams [Gunasekara *et al.*, 2024]

5.2 Part 2: Unsupervised learning (Clustering and Drift Detection)

In part 2, we will explore unsupervised concept drift detection and segue into the discussion on clustering for data streams. Unsupervised drift detection is a key challenge in the field and a natural extension of the traditional supervised approach. Applying clustering in a streaming setting involves various nuances, including online (micro-clustering) and offline (macro-clustering) steps. We aim to discuss those in detail while showing examples of how algorithms can be used in practice.

Key references

- Machine learning for data streams: with practical examples in MOA [Bifet *et al.*, 2017]
- An evaluation of data stream clustering algorithms [Mansalis *et al.*, 2018]
- A framework for clustering evolving data streams [Aggarwal *et al.*, 2003]
- The clustree: indexing micro-clusters for anytime stream mining [Kranen *et al.*, 2011]
- STUDD: A student–teacher method for unsupervised concept drift detection [Cerqueira *et al.*, 2023]

5.3 Part 3: Partially and delayed labeled data

We will explore existing approaches for handling partially and delayed labeled data. This includes suggestions on evaluating algorithms and identifying the desired characteristics of the methods. Lastly, we will discuss AutoML, a de facto approach for batch machine learning. Its application to data streams is still in its early stages, presenting several challenges.

Key references

- Change with delayed labeling: When is it detectable? [Žliobaite, 2010]
- Active learning with drifting streaming data [Žliobaitė *et al.*, 2013]
- Realistic evaluation of deep semi-supervised learning algorithms [Oliver *et al.*, 2018]
- Semi-supervised learning over streaming data using MOA [Le Nguyen *et al.*, 2019]
- Delayed labelling evaluation for data streams [Grzenda *et al.*, 2020]
- A survey on semi-supervised learning for delayed partially labelled data streams [Gomes *et al.*, 2022]

6 Potential Target Audience and Prerequisites

This tutorial's target audience includes researchers and practitioners, especially those interested in learning from data streams, evolving data, and/or IoT applications. No previous experience in machine learning for data streams is required, but familiarity with traditional machine learning concepts and frameworks (like scikit-learn) is expected.

7 Importance to IJCAI audience and Covered IJCAI Tutorial Objectives

We believe this tutorial on Machine Learning for Data Streams goes beyond previously discussed IJCAI topics such as concept drift detection, classification, and regression methods for data streams.

During the tutorial, we intend to educate experts and non-experts on the latest developments in Machine Learning for Streaming Data by providing the relevant conceptual framework with practical examples.

8 Ethical Concerns

No ethical concerns.

9 Organizers

All the organizers will attend the conference.

9.1 Albert Bifet

Professor Albert Bifet is the Director of the Te Ipu o te Mahara AI Institute at the University of Waikato and Co-chair of the Artificial Intelligence Researchers Association (AIRA). His research focuses on Artificial Intelligence, Big Data Science, and Machine Learning for Data Streams. He is leading the TAIAO Environmental Data Science project and co-leading the open source projects MOA (Massive Online Analysis), StreamDM for Spark Streaming and SAMOA (Scalable Advanced Massive Online Analysis). He is the co-author of a book on Machine Learning from Data Streams published at MIT Press. He is one of the winners of the best paper award at the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) 2023, and he will be the general co-chair of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) 2024.

Website: https://albertbifet.com/

List of tutorials: https://albertbifet.com/tutorials/

9.2 Bernhard Pfahringer

Bernhard Pfahringer received his PhD degree from the University of Technology in Vienna, Austria, in 1995. He is a Professor with the Department of Computer Science, and a co-director for the AI Institute, at the University of Waikato in New Zealand. His interests span a range of data mining and machine learning sub-fields, with a focus on streaming, randomization, and complex data. Bernhard is the co-author of the book "Machine Learning from Data Streams" published at MIT Press.

Website: https://www.cs.waikato.ac.nz/~bernhard/

Short list of tutorials:

- Bifet A., Pfahringer B.: Hands-on Tutorial on Massive Online Analytics. KDD 2017. Pfahringer B.:
- Weka: A Tool for Exploratory Data Mining. IEEE Symposium Series on Computational Intelligence 2007.
- Witten I.H., Frank E., Pfahringer B., Hall M.: Inside WEKA – and Beyond the Book, Tutorial at ICML 2002.

9.3 Heitor Murilo Gomes

Heitor is a senior lecturer at the Victoria University of Wellington (VuW) in New Zealand. Before joining VuW, Heitor was a senior research fellow and co-director of the AI Institute at the University of Waikato were he taught from 2020 to 2022 the "data stream mining" (COMPX523) course. Heitor's main research area is the application of machine learning for data streams in a variety of tasks. In this field, he has contributed to ensemble learning for both regression and classification tasks, worked on unsupervised drift detection, and in 2023, he was awarded a grant to conduct research on developing novel theories and algorithms for partially delayed labeled streams. Besides participating as PC member of a multitude of conferences (KDD, IJCAI, ECML, PAKDD, ...) Heitor is also an active contributor to open-source projects like MOA (Massive Online Analysis), StreamDM (a real-time analytics open-source software library built on top of Spark Streaming), and river (where he supervises students and postdocs since the inception of the project).

Website: http://www.heitorgomes.com

9.4 Nuwan Gunasekara

Nuwan earned his PhD in "Advanced Adaptive Classifier Methods for Data Streams" from the University of Waikato, and he currently works at the AI Institute of the same university. His research interests primarily revolve around Stream Learning, Online Continual Learning, and Online Streaming Continual Learning. He has delivered a guest lecture and talks at the University of Waikato's Data Stream Mining (OMPX523 Masters) course and Cardiff University's Machine Learning Seminar. Nuwan has contributed to this field by working on streaming gradient boosted trees for classification and regression, developing neural network-based methods for data streams, and exploring the intersection between

Stream Learning and Online Continual Learning. Nuwan's research has been featured in esteemed publications like IJCAI, Springer Machine Learning, and IJCNN. He is currently serving as a PC member for IJCAI 24 Survey Track. He also actively contributes to and maintains the MOA (Massive Online Analysis) Stream Learning Platform.

Website: https://nuwangunasekara.github.io

References

[Aggarwal *et al.*, 2003] Charu C Aggarwal, S Yu Philip, Jiawei Han, and Jianyong Wang. A framework for clustering evolving data streams. In *Proceedings 2003 VLDB conference*, pages 81–92. Elsevier, 2003.

[Bifet *et al.*, 2015] Albert Bifet, Gianmarco de Francisci Morales, Jesse Read, Geoff Holmes, and Bernhard Pfahringer. Efficient online evaluation of big data stream classifiers. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 59–68, 2015.

[Bifet *et al.*, 2017] Albert Bifet, Ricard Gavalda, Geoffrey Holmes, and Bernhard Pfahringer. *Machine learning for data streams: with practical examples in MOA*. MIT press, 2017.

[Cerqueira *et al.*, 2023] Vitor Cerqueira, Heitor Murilo Gomes, Albert Bifet, and Luis Torgo. Studd: A student–teacher method for unsupervised concept drift detection. *Machine Learning*, 112(11):4351–4378, 2023.

[Gama *et al.*, 2009] Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues. Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–338, 2009.

[Gomes *et al.*, 2017] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, 50(2):1–36, 2017.

[Gomes *et al.*, 2022] Heitor Murilo Gomes, Maciej Grzenda, Rodrigo Mello, Jesse Read, Minh Huong Le Nguyen, and Albert Bifet. A survey on semi-supervised learning for delayed partially labelled data streams. *ACM Computing Surveys*, 55(4):1–42, 2022.

[Grzenda *et al.*, 2020] Maciej Grzenda, Heitor Murilo Gomes, and Albert Bifet. Delayed labelling evaluation for data streams. *Data Mining and Knowledge Discovery*, 34(5):1237–1266, 2020.

[Gunasekara *et al.*, 2024] Nuwan Gunasekara, Bernhard Pfahringer, Heitor Murilo Gomes, and Albert Bifet. Gradient boosted trees for evolving data streams. In *Machine Learning*. Springer, 2024.

[Kranen *et al.*, 2011] Philipp Kranen, Ira Assent, Corinna Baldauf, and Thomas Seidl. The clustree: indexing micro-clusters for anytime stream mining. *Knowledge and information systems*, 29:249–272, 2011.

[Le Nguyen *et al.*, 2019] Minh Huong Le Nguyen, Heitor Murilo Gomes, and Albert Bifet. Semi-supervised learning over streaming data using moa. In *2019 IEEE international conference on big data (Big Data)*, pages 553–562. IEEE, 2019.

[Mansalis *et al.*, 2018] Stratos Mansalis, Eirini Ntoutsi, Nikos Pelekis, and Yannis Theodoridis. An evaluation of data stream clustering algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(4):167–187, 2018.

[Montiel *et al.*, 2018] Jacob Montiel, Jesse Read, Albert Bifet, and Talel Abdessalem. Scikit-multiflow: A multi-output streaming framework. *The Journal of Machine Learning Research*, 19(1):2915–2914, 2018.

[Montiel *et al.*, 2021] Jacob Montiel, Max Halford, Saulo Martiello Mastelini, Geoffrey Bolmier, Raphael Sourty, Robin Vaysse, Adil Zouitine, Heitor Murilo Gomes, Jesse Read, Talel Abdessalem, et al. River: machine learning for streaming data in python. *The Journal of Machine Learning Research*, 22(1):4945–4952, 2021.

[Oliver *et al.*, 2018] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

[Sun *et al.*, 2022] Yibin Sun, Bernhard Pfahringer, Heitor Murilo Gomes, and Albert Bifet. Soknl: A novel way of integrating k-nearest neighbours with adaptive random forest regression for data streams. *Data Mining and Knowledge Discovery*, 36(5):2006–2032, 2022.

[Sun *et al.*, 2024] Yibin Sun, Bernhard Pfahringer, Heitor Murilo Gomes, and Albert Bifet. Adaptive prediction interval for data stream regression. In *PAKDD*, 2024.

[Žliobaitė *et al.*, 2013] Indrė Žliobaitė, Albert Bifet, Bernhard Pfahringer, and Geoffrey Holmes. Active learning with drifting streaming data. *IEEE transactions on neural networks and learning systems*, 25(1):27–39, 2013.

[Žliobaite, 2010] Indre Žliobaite. Change with delayed labeling: When is it detectable? In *2010 IEEE International Conference on Data Mining Workshops*, pages 843–850. IEEE, 2010.