Meta-Features For Data Streams

Internship

Supervision	Maroua Bahri
	Lars Kotthoff
Mail	maroua.bahri@lip6.fr, lars.kotthoff@lip6.fr
Address	LIP6, Sorbonne Université

Keywords	Meta-features, Data streams, Meta-Learning, AutoML

Context

Meta-learning is an important component in automated machine learning (AutoML), enabling systems to characterize datasets using meta-features and to select/adapt algorithms accordingly. While meta-features are well studied for static datasets (e.g., statistical, landmarking, model-based) [1, 2], less work has been done for data streams and many open challenges remain.

Data streams differ from static datasets because they are potentially infinite and cannot be stored entirely, evolve over time due to concept drift, and are processed under time and memory constraints. These requirements make classical meta-features unusable or computationally too expensive [3].

Preliminary efforts introduced the notion of stream meta-features, such as *decision stumps*, which summarize classifier behavior over time through error-rate evolution, confidence estimates, or margin statistics [4]. Additional attempts rely on incremental statistics or drift-sensitive measurements. However, existing methods suffer from: (i) limited coverage, (ii) insufficient computational efficiency, (iii) weak sensitivity to different types of concept drift, and (iv) poor integration within modern meta-learning and AutoML frameworks for data streams. This gap currently limits the development of reliable and adaptive Stream AutoML systems which require fast, robust meta-features of evolving data characteristics.

Objectives

This internship aims to design, implement, and evaluate new efficient stream-compatible meta-features to provide a foundation for meta-learning and AutoML in data streams [4].

The work is organized around three tightly connected research axes. First, you will perform a thorough analysis and extension of existing meta-features for data streams. This involves reviewing the literature for current approaches such as decision stumps, drift-sensitive, and online statistical or structural indicators. The goal of this review is to understand their computational limitations, identify missing categories, and uncover opportunities for more expressive or efficient descriptors of evolving data.

Building on this analysis, you will design of novel stream meta-features or efficient incremental versions of established ones. These may include drift-aware descriptors capturing the magnitude and recurrence of concept drift, online adaptations of classical sta-

tistical measures, dimensionality estimates for evolving distributions, clustering-based indicators, and behavioral descriptors derived from online model dynamics. All proposed meta-features must operate in real time under tight memory constraints and remain compatible with prequential evaluation (Test-Then-Train).

Expected outcomes and contributions:

- A comprehensive taxonomy and analysis of existing stream meta-features, including decision stumps and drift descriptors [5].
- To support transparency and further research, you will provide a fully documented open-source implementation (e.g., with River or MOA). This will include all source code, datasets, and experimental configurations used, allowing researchers and practitioners to reproduce and build upon our work easily.
- A research report (and potentially a publication) documenting theory, algorithms, and experiments.

Internship

The project is intended for a Master's thesis (or equivalent) student dedicated to the objectives of the project. The student will do his/her master's thesis in the RO team at LIP6.

Skills

- Master level research internship M2 or equivalent (stage de fin d'études ingénieur).
- Strong programming skills in Python.
- Familiarity with reading, understanding, and building on academic publications.
- Sound knowledge of machine learning and data streams (appreciated).

Gratification

According to current regulations.

Contact to apply

Send the following documents (exclusively in PDF format) to maroua.bahri@lip6.fr and lars.kotthoff@lip6.fr:

- A cover letter explaining your qualifications, experiences, and motivation for this topic.
- Curriculum vitae.
- Transcripts of grades from the third year of your bachelor's degree, the first year of your master's degree, and any available grades from the second year of your master's degree (or equivalent for engineering schools).
- If possible, recommendation letters.
- If possible, a link to repositories of personal projects (e.g., GitHub).

References

[1] A. Rivolli, L. C. Garcia, J. Vanschoren, and A. C. de Carvalho, "Towards reproducible empirical research in meta-learning," *JMLR*, 2022.

[2] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, *Metalearning: Applications to Data Mining*. Springer, 2009.

[3] M. Bahri, A. Bifet, J. Gama, H. M. Gomes, and S. Maniu, "Data stream analysis: Foundations, major tasks and tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 3, p. e1405, 2021.

[4] V. M. A. Souza *et al.*, "Data stream meta-features and concept drift characterization," *Machine Learning*, 2017.

[5] J. Komorniczak and P. Ksieniewicz, "On metafeatures' ability of implicit concept identification," *Machine Learning*, vol. 113, no. 10, pp. 7931–7966, 2024.