Postdoctoral offer 2024
Ensemble constrained clustering for time series analysis, with application to industry data

1 CONTEXT

Automated data acquisition systems and increasing storage capacities have made time series data available across a wide range of domains, from earth observation to industry. However, this data is often provided with insufficient or no labels, thus preventing the use of supervised methods. In this context, unsupervised methods can be valuable to help users extract information, such as identifying different behaviors on a production line. Nevertheless, when it comes to time series analysis, these methods face several drawbacks.

First, the diverse nature of sensors and sources used to generate temporal data results in significant heterogeneity in terms of format, volume, quality, and richness of information. For example, a single production line can include a large set of different sensors, each constrained by its manufacturer's API. This diversity has led to a wide range of categorization methods for analyzing time series, e.g., based on elastic metrics [1], frequency decomposition [2], and pattern extraction [3], each with its own advantages and limitations, which can also complement one another.

Secondly, clustering approaches often yield results that do not align with the experts' expectations or intuitions. This is especially true when considering the aforementioned heterogeneity of time series data. Therefore, incorporating some expert knowledge, even if it doesn't encompass the full spectrum of actual classes, can significantly enhance the quality of the clustering results [4]. This knowledge is often expressed in the form of constraints [5]. However, these methods often suffer from the negative impact of constraints, resulting in a decrease in quality when constraints are added [4, 6].

Finally, asking experts to define all classes at the outset of the project is unreasonable. It is indeed often the case that not all classes can be semantically defined before a data analysis has been carried out. It is more practical to engage experts throughout the entire process as they progressively unfold the data processing and analysis within an iterative cycle of interactions between the expert and the learning system [7]. The goal of this interaction is to bridge the gap between the results generated by the algorithms and the expert's thematic insights. This process is designed to make the results more comprehensible to the expert.

2 ORGANISATION

2.1 GOALS

The main task of this post-doc is to develop an ensemble clustering method that relies on a diversity of viewpoints (i.e. representations or metrics). It will use constraints given iteratively by the user to select and combine the proper viewpoints. This should result in a better clustering that is a consensus of the most suitable viewpoints, in adequacy with the expert's knowledge, to leverage potential negative effects of constraints. To achieve this goal, we need to fulfill four objectives:

- Select a subset of sufficiently independent/diverse existing metrics/representations (required to have complementary viewpoints) relevant to clusterize time series;

- Define a generic ensemble method to obtain a consensus clustering result from the previously selected viewpoints that maximize the respect of the expert's knowledge;
- Propose a generic method to iteratively update the clustering by integrating new expert's knowledge in interaction with the expert;
- Validate the method operability by focusing on Industry data, mainly relying on a demonstration production line of one of our industry partners.

2.2 COLLABORATION AND SUPERVISION

The person recruited will be co-directed by Nicolas Lachiche (50%), specialist of complex data mining, and Baptiste Lafabrègue (50%), time series analysis specialist. He or she will actively collaborate with the SDC team at ICube in Strasbourg, and more particularly with Nassime Mountasir, a 3rd-year PhD student working on predictive maintenance issues.

3 TO APPLY

3.1 CANDIDATE PROFILE

- PhD in Computer Science, specializing in machine learning/explainability.
- Solid knowledge of Machine Learning methods. Experience in time series analysis and/or predictive maintenance would be also valuable.
- Good verbal (English or French) and written (English) communication skills.
- Interpersonal skills and the ability to work individually or as part of a project team.

3.2 GENERAL INFORMATIONS

- Location: Illkirch, south of Strasbourg (Pôle API, 300 Bd Sébastien Brant, 67400 Illkirch-Graffenstaden)
- Duration: One year (renewable once)
- Gross salary: 3200€/month
- To apply: Interested candidates should submit (by e-mail) their curriculum vitae, a list of publications, a letter of motivation and contact details for two references. Applications will be accepted until the position is filled. The position will start as soon as possible, hopefully in October 2024.
- Contact :
    - Baptiste Lafabregue, lafabregue@unistra.fr
    - Nicolas Lachiche, nicolas.lachiche@unistra.fr

4 BIBLIOGRAPHIC REFERENCES

[1] Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), 43-49.
[2] Daubechies, I. (1992). Ten lectures on wavelets. Society for industrial and applied mathematics.
[3] Ye, L., & Keogh, E. (2009, June). Time series shapelets: a new primitive for data mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 947-956).
[4] Lampert, T., Dao, T. B. H., Lafabregue, B., Serrette, N., Forestier, G., Crémilleux, B., ... & Gançarski, P. (2018). Constrained distance based clustering for time-series: a comparative and experimental study. Data Mining and Knowledge Discovery, 32, 1663-1707.

[5] Koptelov, M., … & Teisseire, M. (2023, October). Towards a (Semi-)Automatic Urban Planning Rule Identification in the French Language. In Proceedings of the 10th IEEE International Conference on Data Science and Advanced Analytics (DSAA)
[6] Davidson, I., Wagstaff, K. L., & Basu, S. (2006, September). Measuring constraint-set utility for partitional clustering algorithms. In European conference on principles of data mining and knowledge discovery (pp. 115-126).
[7] Lafabregue, B., Gançarski, P., Weber, J., & Forestier, G. (2022, Nov.). Incremental constrained clustering with application to remote sensing images time series. In 2022 IEEE International Conference on Data Mining Workshops (pp. 814-823).