Repeatable and reproducible evidential clustering for a healthcare application

Phd project summary

This PhD project is set in the context of longitudinal data analysis for profiling patients with chronic pain, with the aim of forecasting the progression of the disease and thereby personalizing patient care according to their profile and expected trajectory. The input data consist of multivariate ordinal time series of variable lengths. These time series are transformed into sequences of states in order to better identify common patterns across patients.

The objective of the PhD is to develop a clustering method adapted to these data in order to group patients according to similar trajectories. These groups will then be interpreted using explanatory variables provided by the French National Health Data System (SNDS).

In this application context, several scientific challenges have been identified:

- The data are inherently difficult to separate into distinct clusters. Indeed, human-related data involve high individual variability and frequent overlap between clusters [1,2]. To clearly distinguish overlapping regions between clusters, belief function theory (evidence theory) [3,4] may be used.
- As is common in large real-world patient cohorts, the amount of missing data is substantial.
- Clustering methods commonly used in the medical field are generally linear methods [5]. However, the assumption of linear boundaries between clusters is questionable [6]. While non-linear methods appear more appropriate, they introduce issues of repeatability and reproducibility, particularly in the presence of overlapping clusters.

The goal is therefore to develop robust non-linear clustering methods capable of handling ordinal temporal data (or nominal data in the case of state sequences), missing data, and providing a clear representation of overlapping regions between clusters.

The work will begin with a study of the repeatability and reproducibility of existing algorithms. New algorithms will then be developed and evaluated on synthetic datasets in order to thoroughly characterize their properties. Finally, the proposed methods will be applied to the medical dataset.

Application deadline: June 30, 2026

Required profile and skills:

- Applied mathematics statistics, data mining, machine learning
- Programming skills (python and/or matlab)
- Ability to work independently, teamwork skills, interest in applied research

Location :

La PhD will be carried out at LIMOS, located at Clermont-Ferrand, France. The subject is funded by the AI4health project, a chair from the MIAI cluster, and involves a collaboration between the LIMOS, the Analgesia Institute located at the CHU of Clermont-Ferrand, and the LISTIC in Annecy. Regular meetings with partner institutions are planned.

Contacts :

- Desired start date: October 2026

- Salary: 2300 euros gross per month
- Contacts:
  - LIMOS : Violaine Antoine violaine.antoine@uca.fr
  - CHU : Nicolas Kerckhove nkerckhove@chu-clermontferrand.fr
  - LISTIC : Didier Coquin didier.coquin@univ-smb.fr

Bibliography :

[1] Campagner, Andrea, et al. Everything is varied: The surprising impact of instantial variation on ML reliability. Applied Soft Computing 146 : 110644, 2023.
[2] A. Soubeiga, V. Antoine, A. Corteval, N. Kerckhove, S. Moreno, I. Falih, J. Phalip: Clustering and Interpretation of time-series trajectories of chronic pain using evidential c-means. Expert System and Applications 260: 125369, 2025.
[3] Glenn Shafer. A mathematical theory of evidence. Volume 42. Princeton university press, 1976.
[4] Zhe Liu, and Sukumar Letchmunan. Representing uncertainty and imprecision in machine learning: a survey on belief functions. Journal of King Saud University – Computer and Information Sciences, 36(1):101904, 2024.
[5] Tanguay-Sabourin C, Fillingim M, Guglietti GV, Zare A, Parisien M, Norman J, Sweatman H, Da-Ano R, Heikkala E; PREVENT-AD Research Group; Perez J, Karppinen J, Villeneuve S, Thompson SJ, Martel MO, Roy M, Diatchenko L, Vachon-Presseau E. A prognostic risk score for development and spread of chronic pain. Nature Medicine. 2023 Jul;29(7):1821-1831. Doi: 10.1038/s41591-023-02430-4. Epub 2023 July 6. PMID: 37414898
[6] Krakovska O, Christie G, Sixsmith A, Ester M, Moreno S. Performance comparison of linear and non-linear feature selection methods for the analysis of large survey datasets. PLoS One. 14(3):e0213584, 2019.