Clusterings évidentiels répétables et reproductibles pour une application en santé

Résumé du projet de thèse

La thèse s'inscrit dans un contexte d'analyse de données longitudinales pour le profilage de patients douloureux chroniques, dans le but d'établir des projections sur l'évolution de la pathologie et ainsi de personnaliser la prise en charge des patients selon leur profil et leur évolution théorique en découlant. Les données en entrée correspondent à des séries temporelles ordinales et multivariées de longueur variables. Ces séries temporelles sont transformées en une suite d'état de manière à mieux identifier des patterns identiques entre patients.

L'objectif de la thèse est de proposer une méthode de clustering adaptée aux données afin de grouper les patients selon des évolutions similaires. Ces groupes doivent ensuite être interprétés via l'utilisation de variables explicatives fourni par le SNDS.

Dans ce contexte applicatif, plusieurs défis scientifiques sont identifiés :

- les données sont intrinsèquement peu séparables en cluster. En effet, les données liées à l'humain impliquent une forte variabilité individuelle et de fréquentes zones de chevauchement entre clusters [1,2]. Afin de distinguer de manière claire les zones de chevauchement de clusters, on pourra utiliser la théorie des fonctions de croyance [3,4].
- inhérent aux grandes cohortes de patients en vie réelle, le nombre de données manquantes est important.
- les méthodes de clustering généralement utilisés dans le domaine médical sont des méthodes linéaires [5]. Or, l'hypothèse de frontière linéaire entre clusters est questionnable [6]. L'utilisation de méthodes non linéaires semblent plus appropriée, mais introduit des problèmes de répétabilité et de reproductibilité liés à la présence de recouvrement en les clusters.

L'idée est donc de proposer des méthodes de clustering non linéaires, robustes, prenant en compte des données temporelles ordinales (ou nominales dans le cas d'une suite d'état), les données manquantes et permettant de représenter clairement les zones de chevauchement entre clusters.

Dans un premier temps, une étude de la répétabilité et de la reproductibilité des algorithmes existants sera réalisée. De nouveaux algorithmes seront ensuite proposés et testées sur des jeux de données synthétiques afin de bien identifier leurs caractéristiques. Enfin, les méthodes seront appliquées au jeu de données médicales.

Date limite pour postuler : 30/06/2026

Profil et compétence requis :

- Mathématiques appliquées en informatique, statistiques, data mining, machine learning
- programmation (python et/ou matlab)
- bon niveau d'anglais
- autonomie, capacité à travailler en équipe, intérêt pour le domaine appliqué

Localisation :
La thèse se déroule au LIMOS, à Clermont-Ferrand. Le sujet est financé par le projet AI4health, une chaire du cluster MIAI, en collaboration entre le LIMOS et l'institut Analgésia, situé au CHU de Clermont-Ferrand, et le LISTIC à Annecy. Des réunions régulières sont prévues avec les deux entités.

Contacts :

- Date de début souhaitée : octobre 2026
- Salaire :2300 euros brut par mois
- Contacts :
  - LIMOS : Violaine Antoine violaine.antoine@uca.fr
  - CHU : Nicolas Kerckhove nkerckhove@chu-clermontferrand.fr
  - LISTIC : Didier Coquin didier.coquin@univ-smb.fr

Bibliographie :

[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.