CEA LIST

Thèse

Co-clustering de séries temporelles

Contexte

Le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) est un organisme public de recherche qui est un acteur majeur de l'espace européen de la recherche et exerce une présence croissante à l'international.

Au sein du CEA Tech, l'institut CEA LIST focalise ses recherches sur les systèmes numériques intelligents. Porteurs d'enjeux économiques et sociétaux majeurs, ses programmes de R&D sont centrés sur le *manufacturing* avancé (robotique, réalité virtuelle & augmentée, contrôle non destructif, vision), les systèmes embarqués (sûreté & sécurité, ingénierie logicielle et systèmes, architectures de calcul), l'intelligence ambiante (capteurs, instrumentation & métrologie, communication & interfaces sensorielles, traitement de données & multimédia). En développant des technologies de pointe dont les applications couvrent les secteurs des transports, de la sécurité/défense, du *manufacturing*, de l'énergie et de la santé, le CEA LIST contribue à la compétitivité industrielle de ses partenaires par l'innovation et le transfert technologique (www-list.cea.fr).

Au sein de l'institut CEA LIST, le stagiaire évoluera dans le Laboratoire de Sciences des Données et de la Décision (LS2D) qui comprend une trentaine de personnes.

Sujet de la thèse

Le *co-clustering* (classification croisée) [1,2] est une méthode d'apprentissage automatique non supervisé qui a pour objectif d'identifier la structure en blocs homogènes d'un tableau de données à partir d'une classification jointe des lignes et des colonnes. Depuis 1965, ce problème a été développé sous plusieurs variantes mais son intérêt s'est considérablement accru ces dernières années avec l'arrivée de nombreuses applications comme l'analyse de données textuelles, l'analyse marketing, la génomie, les systèmes de recommandations ou bien encore l'étude de données énergétiques. Ce type d'approches organise simultanément les lignes et les colonnes d'un tableau pour découvrir des blocs homogènes alignés pour proposer une lecture simplifiée des données ou/et en extraire des caractéristiques (*feature engineering*) utilisées par la suite dans des modèles de *machine learning*. Parmi les méthodes développées, on distingue deux types d'approches : les factorisations matricielles et les modèles probabilistes.

Dans le cas particulier des données temporelles, de récentes approches proposent de transformer dans un premier temps ces données en fonctions pour prendre en compte la notion d'ordre due à la temporalité. A partir de ces transformées, il est alors possible d'utiliser une méthode de *co-clustering* en l'appliquant soit sur les coefficients linéaires [3] des différentes fonctions soit sur les projetés de ces fonctions [4]. Dans le cadre de ces deux approches, la méthode de *co-clustering* retenue par les auteurs est le modèle probabiliste de blocs latents.

Pendant la thèse, le doctorant sera amené à développer une méthode alternative (voire plusieurs) pour le *co-clustering* de signaux temporels notamment basée sur la factorisation matricielle [5] et le *signal processing* [6]. Dans un premier temps, il sera intéressant de travailler sur une approche similaire à celles présentées précédemment en proposant une transformation des signaux temporels (par exemple, transformation de Besov) puis en appliquant une méthode de *co-clustering* (soit par une approche probabiliste soit par une approche matricielle). Pour obtenir une

classification croisée pertinente, on pourra considérer une transformation des signaux par morceaux. Se posera alors la question de la définition des morceaux. Une méthode de *co-clustering* spécifique sera aussi envisagée si nécessaire. Dans un second temps, il pourra être à l'étude de travailler sur des matrices de distance entre signaux pour réaliser une classification croisée pertinente. Dans ces deux cas, l'ordre des colonnes et de leurs classes peut être considéré fixe ou variable. Dans le premier cas, les classes des instants de mesure sont ordonnées chronologiquement. Dans le second, la classification croisée autorise dans les mêmes blocs des plages d'instants différents dans le but d'extraction de caractéristiques.

Références :

[1] V. Brault, A. Lomet, Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau, Journal de la Société Française de Statistique, vol.156, issue.3, pp.27-51, 2015

[2] Aurore Lomet, Gérard Govaert, Yves Grandvalet, Model Selection for Gaussian Latent Block Clustering with the Integrated Classification Likelihood, Advances in Data Analysis and Classification, Springer Verlag, 2018, 12 (3), pp.489-508.

[3] C. Bouveyron, L. Bozzi, J. Jacques, F-X. Jollois, The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves, HAL Id: hal-01533438, 2017

[4] Ben Slimen, Y., S. Allio, and J. Jacques. Model-based co-clustering for functional data. In Proceedings of the 48th conference of the French Statistical Society, Montpellier, France, 2018

[5] Wendyam Serge Boris Ouedraogo, Antoine Souloumiac, Meriem Jaidane, Christian Jutten, Non-negative Blind Source Separation Algorithm Based on Minimum Aperture Simplicial Cone, IEEE Transactions on Signal Processing, vol. 62, n°2, pp. 376-389, 2014.

[6] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, Guillaume Gibert, xDawn Algorithm to Enhance Evoked Potentials : Application to Brain-Computer Interface, IEEE Transactions on Biomedical Engineering, vol. 56, n°8, pp. 2035-2043, 2009.

Mots clés : *co-clustering*, apprentissage automatique non supervisé, séries temporelles, *signal processing*

Environnement et Prérequis

- **Lieu de la thèse :** La thèse se déroulera au CEA Saclay, dans le bâtiment DIGITEO
- **Durée :** 3 ans. Les formalités nécessaires au recrutement du candidat étant assez longues, il est recommandé de commencer les démarches au moins 3 mois avant le début de la thèse.
- **Rémunération :** selon profil
- **Prérequis :** Le candidat sera en M2 spécialisé en *machine learning*, *signal processing,* apprentissage automatique non supervisé.
- **Responsables et contact :**
  - Contact : Aurore LOMET, aurore.lomet@cea.fr
  - Antoine Souloumiac, antoine.souloumiac@cea.fr