

Partitionnement sous contrainte de similarité

Application à la classification des nœuds de consommation et de production
d'un réseau de distribution d'électricité

Financement : LabCom @LIENOR et Région Nouvelle-Aquitaine

Laboratoire : Laboratoire d'Informatique et d'Automatique pour les Systèmes (LIAS)

Localisation : Chasseneuil-du-Poitou

Date de début : septembre 2022 (flexible)

Date limite d'envoi des candidatures : 24 juin 2022

Salaire net mensuel : 1794 euros

Équipe encadrante : Mickael BARON et Brice CHARDIN

(L'encadrant dont le nom est souligné est l'encadrant référent, contact : brice.chardin@ensma.fr)

Contexte industriel

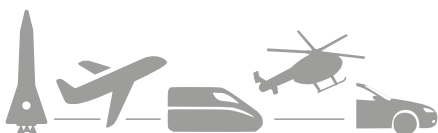
SRD est un gestionnaire de réseau de distribution d'électricité chargé de gérer, exploiter, entretenir et développer un réseau électrique couvrant 90% de la Vienne. Pour l'optimisation de son réseau et la planification d'investissements, SRD cherche à modéliser le comportement des consommateurs et producteurs qu'il dessert. Ce problème de modélisation se rencontre à différents niveaux de granularité : du compteur électrique particulier ou de l'installation photovoltaïque jusqu'au poste de transformation.

Bien que cette modélisation soit principalement basée sur les valeurs historiques de puissance transitant sur le réseau, SRD s'intéresse plus particulièrement à son pouvoir prédictif, c'est-à-dire sa capacité à capturer le comportement futur des éléments considérés. Le caractère variable, aléatoire ou incertain des prévisions dans le domaine de l'énergie devra donc être pris en compte pour aboutir à une catégorisation exploitable par SRD.

Partitionnement (*equiwide clustering*)

Objectif scientifique L'objectif principal de cette thèse est d'élaborer des techniques de classification permettant d'identifier des groupes d'éléments avec une garantie de dissimilarité maximale entre deux éléments d'un même groupe, et de positionner ce type d'approche par rapport aux algorithmes de partitionnement existants – notamment les approches par densité. Les techniques considérées ici sont basées sur un partitionnement sous contrainte [1], et plus spécifiquement sous contrainte de dissimilarité intra-cluster maximale [2]. Ce type de partitionnement garantit une certaine proximité entre les membres d'un groupe et le représentant désigné *in fine* pour les remplacer. Pour simplifier le processus de modélisation résultant de cette classification, l'algorithme cherche à minimiser le nombre de catégories fournies en sortie.

La dissimilarité intra-cluster capture une notion de distance désignée en théorie des graphes par *diamètre* ou *rayon*. Pour une contrainte de diamètre, le problème est équivalent au problème de couverture par cliques minimale ou de coloration de graphe [3]. Pour une contrainte de rayon, le problème est équivalent à la détermination d'un ensemble dominant de cardinalité minimale.



Ces équivalences ouvrent des perspectives sur des implémentations efficaces basées sur des approximations, qui auront un intérêt pratique étant donné la difficulté théorique de ces deux problèmes (NP-difficile).

La dissimilarité et les seuils choisis doivent avoir un sens pour SRD afin que les opérateurs de conduite puissent raisonner sur les résultats obtenus. Une telle dissimilarité pourrait être basée sur l'impact pratique d'une interversion de deux éléments, que ce soit sur la topologie du réseau – par exemple la différence entre deux topologies optimales – ou les flux de puissance – par exemple les pertes par effet Joules à topologie identique. De manière alternative ou complémentaire, des approches semi-supervisées [4] peuvent être envisagées pour estimer une dissimilarité pertinente à partir d'exemples annotés par les opérateurs de conduite.

Contribution technique Un objectif technique du projet consiste à mettre à disposition les techniques de résolution élaborées pendant cette thèse en enrichissant une bibliothèque open source existante : scikit-learn (via scikit-learn-contrib). Pour qu'une telle intégration puisse être réalisée, le projet devra en satisfaire les exigences, notamment en ce qui concerne la qualité de l'implémentation ou de la documentation, les tests unitaires et l'intégration continue.

Profil du candidat

Le candidat devra être titulaire d'un master en informatique ou d'un diplôme d'ingénieur et posséder des connaissances en développement logiciel, systèmes d'information, statistiques et analyse de données. Un bon niveau en français et en anglais est nécessaire.

Documents à fournir :

- Curriculum Vitae;
- lettre de motivation;
- notes de master ou équivalent;
- tout autre document jugé utile pour enrichir le dossier de candidature (e.g. lettres de recommandation, articles scientifiques, rapport de master).

Références

- [1] T.-B.-H. Dao, K.-C. Duong, and C. Vrain, "Constrained clustering by constraint programming", *Artificial Intelligence*, vol. 244, pp. 70–94, 2017.
- [2] J. Andersen, B. Chardin, and M. Tribak, "Clustering to the fewest clusters under intra-cluster dissimilarity constraints", in *Proceedings of the Thirty-Third IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021.
- [3] P. Hansen and M. Delattre, "Complete-link cluster analysis by graph coloring", *Journal of the American Statistical Association*, vol. 73, no. 362, pp. 397–403, 1978.
- [4] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering", in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, 2004, pp. 839–846.

