

Proposition de stage

Création d'un package R "Online Relational SOM"

Contexte :

Avec l'affluence croissante des données complexes (corpus, réseaux sociaux, parcours de vie ou professionnels, ...), les méthodes de data-mining doivent s'adapter afin de gérer à la fois la complexité et la taille des données. Le sujet du stage porte sur les cartes auto-organisées [2], des algorithmes qui permettent de visualiser les données via une projection non-linéaire ainsi que de construire une classification non-supervisée. Initialement, les cartes auto-organisées ont été développées pour des données euclidiennes, mais, depuis quelques années, des généralisations pour des données caractérisées par des matrices de distances ou de dissimilarités [1] ont été proposées. Dans un premier temps, ces développements ont été décrits uniquement en version BATCH. Récemment, une version Online (Online Relational SOM) a été introduite [3]. Avec cette dernière méthode, les temps de calcul sont significativement réduits et l'organisation de la carte est meilleure.

- [1] B. Hammer and A. Hasenfuss. Relational topographic maps. Technical Report IfI-07-01, Clausthal University of Technology, 2007.
- [2] T. Kohonen. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.
- [3] M. Olteanu, N. Villa-Vialaneix, and M. Cottrell. On-line relational som for dissimilarity data. In *Proceedings of the WSOM, 2012*. To appear.

Mission :

Le stagiaire sera amené à travailler sur l'algorithme Online Relational SOM. Il participera notamment au développement d'un package R qui sera soumis sur CRAN (<http://cran.r-project.org/web/packages/>) : des versions préliminaires des fonctions à implémenter existent déjà et pourront être utilisées comme base à la programmation. Dans un second temps, une version sparse de l'algorithme sera développée et implémentée pour des données de grande dimension.

Niveau d'études :

BAC+5 avec des compétences en statistique, machine learning, programmation en R. Des compétences en programmation C/C++ seraient appréciées.

Durée et localisation :

Trois à six mois. SAMM, Université Paris 1.

Contacts :

Madalina Olteanu (madalina.olteanu@univ-paris1.fr)
Nathalie Villa-Vialaneix (nathalie.villa@univ-paris1.fr)