TER - Master 1 IDSM

Implémentation d'un algorithme de co-clustering à base de tri-factorisation non-négative de matrice

Contact : julien.jacques@univ-lyon2.fr

Mots-clefs :

Data mining, clustering, co-clustering, logiciel R.

Contexte :

Le clustering consiste à regrouper des observations (ou individus) en clusters, de sorte que les données d'un même cluster soient les plus semblables possibles, et les clusters soient les plus différents les uns des autres. Le clustering est une méthode de data mining souvent utilisée pour résumer l'information contenue dans une grande base de données, en résumant ainsi un grand nombre d'observations par un petit nombre de clusters.

A l'heure du big data, ce ne sont plus seulement les observations qui sont nombreuses mais également les variables observées : ainsi, il est fréquent d'observer plusieurs centaines voir milliers de variables par individu. Des techniques de co-clustering ont alors été proposées, pour résumer à la fois les individus mais également les variables, en réalisant simultanément un clustering des individus et un clustering des variables.

Une des premières approches est la méthode double k-means, proposé par Vichi (2001) et décrit dans Van Rosmalen *et al.* (2009). La méthode double k-means étend le célèbre algorithme des K-means au cas du co-clustering. Récemment, il a été montré qu'il était possible de réaliser un co-clustering en réalisant des tri-factorisations non négative de matrice (Allab *et al*. 2016).

Objectif du projet :

L'objectif du projet est d'implémenter sous le logiciel R l'algorithme NMTF pour le co-clustering. Des tests seront ensuite réalisés sur des données simulées puis sur des données réelles.

Compétences mises en œuvre :

- Lire et comprendre un article de recherche
- Savoir implémenter une méthodologie décrite dans un article de recherche
- Savoir mettre en application et tester une méthodologie décrite dans un article de recherche

Références :

Van Rosmalen J. *et al*. (2016). Multi-manifold matrix decomposition for data co-clustering. *Pattern Recognition*, 64:386-398.

Vichi M. (2001). Double k-means clustering for simultaneous classification of objects and variables". In *Advances in Classification and Data Analysis*, 43-52.

K. Allab *et al*. (2016). Multi-manifold matrix decomposition for data co-clustering. *Pattern Recognition*, 64:386-398.

Wang H. et al (2011). Nonnegative Matrix Tri-Factorization Based High-Order Co-Clustering and Its Fast Implementation. 11th IEEE International Conference on Data Mining.

Ding C. et al (2011). Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering.