Approche ensemble pour le co-clustering par blocs sur des données textuelles: Application au biomédical

Séverine Affeldt*, Lazhar Labiod*, Mohamed Nadif*

*Université de Paris, CNRS, Centre Borelli, F-75006 Paris
<prénom.nom>@u-paris.fr

Résumé. Nous proposons un co-clustering par blocs via une approche ensemble qui fusionne plusieurs co-clusterings élémentaires en une matrice d'affinité *consensus* structurée. Les co-clusterings de base sont issus des mêmes données textuelles et générés par la même méthode de co-clustering. Ce processus de fusion renforce la qualité individuelle des co-clusterings par blocs au sein d'une seule matrice consensus. Notre approche permet un co-clustering complètement non supervisé, où le nombre de co-clusters est automatiquement déduit d'un critère de modularité non trivial généralisé. La fonction objective associée permet l'apprentissage conjoint de l'agrégation des co-clusterings élémentaires et du co-clustering consensus. Les résultats expérimentaux sur plusieurs jeux de données réelles démontrent l'intérêt de notre approche comparée à des méthodes compétitives de co-clustering (Affeldt et al., 2020).

1 Introduction

La classification non supervisée ou *clustering* permet le regroupement d'objets similaires en *clusters* homogènes. C'est une approche incontournable dans la science des données et elle est particulièrement utile pour le traitement des données massives. Le choix d'un algorithme de clustering efficace pour l'obtention d'un partitionnement *profitable* n'est pas une tâche simple. En effet, plusieurs algorithmes de clustering ne produisent pas nécessairement la même partition optimale. L'approche *ensemble*, très populaire en apprentissage supervisé, s'avère un très bon moyen pour pallier cet inconvénient. Ainsi, des méthodes proposant une partition *consensus* ont été proposées pour améliorer la pertinence du partitionnement (Strehl et Ghosh, 2003). Toutefois, de telles méthodes ne sont pas conçues pour des données telles que les données textuelles, notamment biomédicales, qui sont généralement clairsemées (*sparse*) et de très grande dimension.

Pour ce type de données, le *co-clustering* ou classification croisée (Govaert et Nadif, 2013) permet, cependant, d'exploiter efficacement la relation naturellement pré-existante entre l'ensemble des *objets* (eg. documents) et leurs *caractéristiques* (eg. termes). Ainsi à partir des matrices document-terme et une classification simultanée des deux ensembles, on obtient généralement une meilleure réorganisation en blocs que par les approches de clustering appliquées séparément sur les deux ensembles (Labiod et Nadif, 2011; Ailem et al., 2016; Salah et Nadif, 2017; Govaert et Nadif, 2018). Pour un corpus de documents, chaque co-cluster fournit un