Biclustering itératif par une approche évolutionnaire

Alexandre Blansché et Lydia Boudjeloud-Assala

Université de Lorraine
CNRS, LORIA, F-57000 Metz, France
prénom.nom@univ-lorraine.fr

1 Problématique

Le *biclustering* traite le problème de la haute dimensionnalité en détectant des *clusters* définis par un ensemble d'objets (lignes) R décrits par un ensemble de variables (colonnes) C. Lors de la recherche de *biclusters* dans un jeu de données, différents modèles peuvent être identifié : *biclusters* constants, lignes ou colonnes constantes, valeurs cohérentes. En outre, les *biclusters* peuvent être exhaustif ou non (chaque ligne et colonne appartient à au moins un *bicluster* ou non), exclusif ou non (chaque ligne et colonne ne peut appartenir qu'à un *bicluster* au plus ou à plusieurs). Dans cet article, nous nous intéressons au cas des *biclusters* cohérents, non exhaustifs, non exclusifs. La méthode de Cheng et Church (Cheng et Church, 2000), notée C&C dans la suite de l'article, extrait les *biclusters* les uns après les autres par un processus itératif. Chaque *bicluster* est un ensemble maximum de colonnes et de lignes avec une mesure d'homogénéité H se situant sous un seuil δ défini. Nous pouvons pointer deux défauts à cette méthode. D'une part, la valeur δ doit être fixée, mais elle est difficile à déterminer. D'autre part, pour éviter les extractions multiples, les *biclusters* extraits sont masqués avec du bruit. Quand des *biclusters* se chevauchent, il y a une perte d'efficacité.

2 Approche proposée

Nous proposons une approche itérative qui extrait les *biclusters* les uns après les autres au lieu d'extraire les *biclusters* en même temps. Le principe général de la méthode proposée reprend l'approche introduite dans Boudjeloud-Assala et Blansché (2012). L'extraction d'un *bicluster* est guidée par trois fonctions objectif : la mesure d'homogénéité H, définie dans Cheng et Church (2000), une mesure le chevauchement des *biclusters*selon l'indice de Jaccard corrigé, proposé dans Hanczar et Nadif (2013) et la racine carrée de la taille du *bicluster*. Le processus d'extraction sera alors une méthode d'optimisation à objectifs multiples. Nous avons choisi d'utiliser un algorithme évolutionnaire multi-objectif, une solution étant décrite par un sous-ensemble d'observations (lignes) et un sous-ensemble de variables (colonnes). Pour accélérer le temps de convergence, nous avons combiné l'approche évolutionnaire avec l'algorithme déterministe de Cheng & Church, pour former un algorithme évolutionnaire lamarckien. Ainsi, pour un individu I, on applique C&C sur sous-ensemble de données ainsi formé, avec une valeur de δ choisie aléatoirement, pour obtenir un individu I' qui remplacera