Pondération de blocs de variables en bi-partitionnement topologique

Amine Chaibi, Hanane Azzag, Mustapha Lebbah

{prenom.nom}@lipn.univ-paris13.fr
Université Paris 13, Sorbonne Paris Cité - CNRS
LIPN-UMR 7030
99, av. J-B Clément - F-93430 Villetaneuse

Résumé. Dans cet article, nous proposons une nouvelle approche permettant à la fois le bi-partitionnement topologique (bi-clustering) et la pondération de blocs variables. Le modèle que nous proposons FBR-BiTM (Feature Block Relevance using BiTM) permet de découvrir un espace topologique d'un ensemble d'observations et de variables en associant un nouveau score de pondération à chaque sous ensemble de variables. L'estimation des coefficients de pondération est réalisée dans le même processus d'apprentissage que le bi-partitionnement. Ces pondérations sont locales et associées à chaque prototype. Elles reflètent l'importance locale de chaque bloc de variables pour le bi-partitionnement. L'évaluation montre que l'approche proposée, comparée à d'autres méthodes de bi-partitionnement, obtient des résultats performants.
Mots clés : bi-partitionnement, pondérations de blocs variables, cartes topologiques.

1 Introduction

Les approches de bi-partitionnement sont devenues un sujet d'intérêt majeur en raison de leurs nombreuses applications dans le domaine de la fouille des données. Une méthode de bi-partitionnement, aussi appelée bi-clustering, co-clustering ou classification croisée, est une méthode d'analyse qui vise à regrouper des données en fonction de leur similarité. La stratégie classique des méthodes de bi-partitionnement cherche à trouver des sous-matrices ou des blocs, qui représentent des sous-groupes de lignes et des sous-groupes de colonnes d'une matrice de données.

Un des objectifs d'une méthode de bi-partitionnement est la recherche d'un couple de partitions, l'une sur les observations (les lignes d'une matrice de données), l'autre sur les variables (colonnes d'une matrice de données), tel que la "perte d'information" due au regroupement soit minimale (Charrad et al., 2008) ; c'est-à-dire de sorte que la différence entre l'information apportée par la matrice de données initiale et celle apportée par le regroupement obtenu soit minimale. Depuis le premier algorithme de bi-partitionnement, appelé Block Clustering proposé par Hartigan (1972), de nombreuses techniques ont été proposées telles que l'énumération exhaustive (Tanay et al., 2002), l'analyse spectrale (Greene et Cunningham, 2010), les réseaux