Nouvelle approche de bi-partitionnement topologique

Amine Chaibi[*,**], Mustapha Lebbah[*], Hanane Azzag[*]

[*] {prenom.nom}@lipn.univ-paris13.fr
[*]Université Paris 13, Sorbonne Paris Cité - CNRS
LIPN-UMR 7030
99, av. J-B Clément - F-93430 Villetaneuse
[**] Anticipeo
4 bis, impasse Courteline 94800 Villejuif, France

Résumé. Dans ce papier, nous proposons une nouvelle approche topologique de bi-partitionnement (bi-clustering) appelée BiTM en utilisant les cartes auto-organisatrices. L'idée principale de l'approche est d'utiliser une seule carte pour le partitionnement simultané des lignes (observations) et des colonnes (variables). Contrairement aux approches utilisant les cartes topologiques, notre modèle ne nécessite pas de pré-traitement de la base de données. Ainsi, une nouvelle fonction de coût est proposée. De plus, BiTM fournit une visualisation topologique des blocs ou bi-clusters facilement interprétable. Les résultats obtenus sont très encourageants et prometteurs pour continuer dans cette optique.

1 Introduction

Les approches de bi-partitionnement sont devenues un sujet d'intérêt en raison de ses nombreuses applications dans le domaine de l'exploration des données. Une méthode de bi-partitionnement, aussi appelée classification croisée, bi-clustering ou co-clustering, 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. 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 bayésiens (Shan et al. (2010)) et d'autres (Angiulli et al. (2006),Charrad et al. (2008)). L'approche Block Clustering (Hartigan (1972)) permet de diviser la matrice des données en plusieurs sous-matrices correspondant à des blocs. Le principe de base de cette méthode est de faire des permutations des lignes et des colonnes afin de définir la structure de bloc. De plus, l'auteur Hartigan (1972) a proposé deux autres algorithmes de bi-partitionnement : le premier (One-Way Splitting) est principalement basé sur le partitionnement des observations en utilisant des fonctions ayant une variance intra-classe supérieure à un seuil donné afin de diviser la classe associée. Le second algorithme (Two-Way Splitting) procède par des divisions successives des lignes et des colonnes. Le même principe a été repris dans l'approche CTWC proposée par Getz et al.