Application des Pattern Structures à la découverte de biclusters à changements de signes cohérents

Nyoman Juniarta*, Miguel Couceiro*, Amedeo Napoli*

*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
firstname.lastname@loria.fr

Résumé. Le “biclustering” joue un rôle majeur dans beaucoup d’applications du monde réel. Il est lié au “clustering” qui regroupe des lignes similaires dans une matrice de données numériques, tandis que le biclustering cherche à regrouper simultanément des lignes et colonnes similaires, c’est-à-dire trouver des sous-matrices où émerge une corrélation entre les entrées. Le biclustering s’appuie sur un critère de similarité, et dans cet article, nous nous intéressons au biclustering “à colonnes constantes” (CC), où les valeurs numériques dans les colonnes des sous-matrices sont constantes pour chaque ligne. L’étude est ensuite étendue au biclustering à “changements de signes cohérents” (CSC), où la différence entre les valeurs de deux colonnes consécutives de la sous-matrice est du même signe pour chaque ligne.

1 Introduction

Le “biclustering” est une technique de fouille de données qui permet de retrouver des motifs sous la forme d’une sous-matrice dans une matrice ou un tableau de données à deux entrées [Madeira et Oliveira (2004)]. Il est lié au “clustering” [Arabie et al. (1996)] dont l’objectif est de regrouper des lignes de la matrice en fonction des similarités qu’elles présentent. Pour une sous-matrice considérée, la similarité entre les lignes doit alors être vérifiée pour toutes les colonnes. D’une façon alternative, le biclustering consiste à regrouper les lignes et les colonnes d’une sous-matrice simultanément, en fonction d’un certain critère de similarité. Par exemple, dans le biclustering à colonnes constantes (CC), pour chaque ligne d’une sous-matrice les valeurs dans les colonnes sont constantes et (possiblement) différentes d’une colonne à l’autre.

Pour mettre en oeuvre le biclustering, nous utilisons dans cet article l’analyse formelle de concepts (FCA pour “Formal Concept Analysis” [Ganter et Wille (1999)]) et les “Pattern Structures”, une extension de la FCA pour traiter les tableaux numériques [Ganter et Kuznetsov (2001); Kaytoue et al. (2011)]. La FCA est très liée au biclustering car elle cherche aussi à regrouper les lignes et colonnes d’un tableau de données binaires en fonction d’une similarité qui consiste à partager les mêmes croix (ou 1) pour chaque ligne d’une sous-matrice.

Nous présentons ici deux méthodes pour découvrir des biclusters dits “à changements de signes cohérents” (CSC), comme défini dans [Madeira et Oliveira (2004)], à partir d’une matrice de signes. Cette matrice est obtenue en deux étapes : (i) en “échelonnant” (“scaling”) la matrice numérique d’origine où ont été recherchés les biclusters à colonnes constantes (CC),