Heuristique pour l'extraction de motifs ensemblistes bruités

Karima Mouhoubi, Lucas Létocart et Céline Rouveirol

LIPN, UMR CNRS 7030, Université Paris 13, 99 av. J.B. Clément, 93430 Villetaneuse, France
nom.prénom@lipn.univ-paris13.fr

Résumé. La recherche de motifs ensemblistes dans des matrices de données booléennes est une problématique importante dans un processus d'extraction de connaissances. Elle consiste à rechercher tous les rectangles de 1 dans une matrice de données à valeurs dans {0,1} dans lesquelles l'ordre des lignes et colonnes n'est pas important. Plusieurs algorithmes ont été développés pour répondre à ce problème, mais s'adaptent difficilement à des données réelles susceptibles de contenir du bruit. Un des effets du bruit est de pulvériser un motif pertinent en un ensemble de sous-motifs recouvrants et peu pertinents, entraînant une explosion du nombre de motifs résultats. Dans le cadre de ce travail, nous proposons une nouvelle approche heuristique basée sur les algorithmes de graphes pour la recherche de motifs ensemblistes dans des contextes binaires bruités. Pour évaluer notre approche, différents tests ont été réalisés sur des données synthétiques et des données réelles issues d'applications bioinformatiques.

1 Introduction

La recherche de motifs ensemblistes dans des données booléennes consiste à rechercher tous les rectangles de 1 dans une matrice à valeurs dans $\{0, 1\}$ dans laquelle l'ordre des lignes et colonnes n'est pas important. Lorsque les données booléennes sont le résultat de traitements sur des données numériques issues de processus expérimentaux complexes, celles-ci peuvent alors contenir du bruit. L'effet du bruit va être de fractionner des motifs importants vérifiant certaines contraintes, telle que le support minimal, en un nombre exponentiel de petits fragments non pertinents. La figure 1 illustre un exemple d'un contexte booléen non bruité (matrice A) où, pour un support minimal de 0.3, deux motifs fréquents maximaux peuvent être extraits ainsi que la même matrice mais en introduisant du bruit (matrice B).

La prise en compte du bruit pour la découverte de motifs a fait l'objet d'un nombre important de travaux de recherche tels que Mannila et Seppanen (2004), Besson et al. (2006) et Liu et al. (2006). Pour résoudre ce problème, la plupart des travaux ont repris le principe de recherche par niveau de l'algorithme Apriori d'Agrawal et al. (1993) et sont donc limités à l'utilisation de contraintes anti-monotones pour élaguer l'espace de recherche.

Dans le travail de Mannila et Seppanen (2004), les auteurs recherchent toutes les régions de support minimal σ et qui dépassent un seuil de densité $\delta \in [0, 1]$. Cette approche permet d'extraire toutes les régions verifiant les contraintes du support et de densité. cependant le choix de ces paramètres reste une tâche difficile et nécessite une connaissance préalable sur les données. De plus, la méthode reste très coûteuse puisqu'elle utilise une recherche par niveau.