Motifs récursifs : extraction ascendante hiérarchique d'ensembles d'items ou d'évènements pour le résumé de données transactionnelles ou séquentielles

Julien Blanchard

Université de Nantes & LINA (CNRS UMR6241) équipe COD
Rue Christian Pauc - 44306 Nantes
julien.blanchard@univ-nantes.fr

Résumé. Nous proposons une méthode originale pour extraire un résumé compact, représentatif et intelligible des motifs fréquents dans des données transactionnelles ou séquentielles. Notre approche consiste à extraire un nouveau type de motifs que nous appelons *motifs récursifs*, i.e. des motifs de motifs, à l'aide d'un algorithme hiérarchique agglomératif nommé *RepaMiner*. Nous générons non pas un simple ensemble de motifs mais une véritable structure dérivée de dendrogrammes, le *RPgraph*.

1 Introduction

L'extraction de motifs fréquents est une tâche essentielle en fouille de données. Les motifs permettent de résumer un jeu de données de manière intelligible et peuvent être utilisés pour d'autres tâches comme l'analyse d'association, la classification supervisée associative, ou la classification à base de motifs. Des algorithmes efficaces ont été proposés pour extraire des motifs dans différents types de données comme les données transactionnelles, les séquences d'évènements, et les graphes. Le principal inconvénient des techniques d'extraction de motifs est l'abondance des motifs produits, qui résulte de la nature combinatoire des algorithmes en oeuvre. Différentes solutions ont été proposées face à ce problème, comme l'intégration de contraintes dans les algorithmes (Boulicaut et Jeudy, 2005), le filtrage des motifs par des mesures d'intérêt (Blanchard, 2005; Blanchard et al., 2007), et l'extraction de représentations condensées des motifs fréquents, i.e. un sous-ensemble des motifs qui permet de générer la totalité des motifs de manière exacte ou approchée (Calders et al., 2006). Malgré ces efforts, le problème reste à peine atténué, comme le rappelle l'étude récente de Giacometti et al. (2013).

Dans cet article, nous proposons une méthode originale pour extraire un résumé compact, représentatif et intelligible des motifs fréquents dans des données transactionnelles ou séquentielles (par exemple une ou plusieurs séquences d'évènements, du texte, des séquences biologiques). Ce résumé peut être lu et interprété directement, mais il offre aussi la possibilité de générer de manière approchée l'ensemble des motifs fréquents et d'estimer leur support. Dans le détail, notre approche consiste à extraire un nouveau type de motifs que nous appelons *motifs récursifs*, i.e. des motifs de motifs, à l'aide d'un algorithme hiérarchique agglomératif nommé RepaMiner. La nature hiérarchique de ces motifs nous permet de produire non pas