Vers l'extraction de motifs rares

Laszlo Szathmary*, Sandy Maumus*,**, Pierre Petronin***
Yannick Toussaint*, Amedeo Napoli*

*LORIA, 54506 Vandoeuvre-lès-Nancy
{szathmar, maumus, yannick, napoli}@loria.fr
**INSERM U525, 54000 Nancy
Sandy.Maumus@nancy.inserm.fr
***ENSAI, 35172 Bruz Cedex
pierre.petronin@gmail.com

Résumé. Un certain nombre de travaux en fouille de données se sont intéressés à l'extraction de motifs et à la génération de règles d'association à partir de ces motifs. Cependant, ces travaux se sont jusqu'à présent, centrés sur la notion de motifs fréquents. Le premier algorithme à avoir permis l'extraction de tous les motifs fréquents est Apriori mais d'autres ont été mis au point par la suite, certains n'extrayant que des sous-ensembles de ces motifs (motifs fermés fréquents, motifs fréquents maximaux, générateurs minimaux). Dans cet article, nous nous intéressons aux motifs rares qui peuvent également véhiculer des informations importantes. Les motifs rares correspondent au complémentaire des motifs fréquents. A notre connaissance, ces motifs n'ont pas encore été étudiés, malgré l'intérêt que certains domaines pourraient tirer de ce genre de modèle. C'est en particulier le cas de la médecine, où par exemple, il est important pour un praticien de repérer les symptômes non usuels ou les effets indésirables exceptionnels qui peuvent se déclarer chez un patient pour une pathologie ou un traitement donné.

1 Introduction

La fouille de données a pour objectif d'identifier des relations cachées entre les motifs de grandes bases de données. La recherche de règles d'association est une des tâches les plus importantes de la fouille de données. L'extraction de règles d'association est un domaine de l'extraction de connaissances dans les bases de données (ECBD), qui se définit comme un procédé pour trouver des motifs valides, utiles et compréhensibles dans les données (Fayyad et al., 1996). Une règle d'association est une proposition de la forme "80% des étudiants qui suivent le cours *Introduction à Unix* suivent également *Programmation en C*" (Han et Kamber, 2001).

Jusqu'à présent, la littérature s'est intéressée à la recherche des règles d'association valides *fréquentes* (c'est-à-dire les règles d'association avec un support et une confiance suffisamment élevés). Cela requiert d'abord l'extraction des motifs fréquents de l'ensemble des données. Le problème de l'extraction des motifs fréquents était au départ un sous-problème de la fouille de