Représentation condensée de motifs émergents

Arnaud Soulet, Bruno Crémilleux, François Rioult

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
14032 Caen Cedex France
{Prenom.Nom}@info.unicaen.fr

Résumé. Les motifs émergents sont des associations de caractéristiques fortement présentes dans une classe et rares dans les autres. Ils font ressortir les distinctions entre classes et se révèlent particulièrement efficaces pour construire des classifieurs et apporter une aide au diagnostic. À cause de la forte combinatoire du problème, la recherche et la représentation des motifs émergents restent des tâches complexes pour de grandes bases de données. Nous proposons ici une représentation condensée exacte des motifs émergents (i.e., les motifs *et* leurs taux de croissance sont directement obtenus depuis la représentation condensée). L'idée principale est de s'appuyer sur les récents résultats relatifs aux représentations condensées de motifs fermés fréquents. À partir de cette représentation, nous donnons aussi une méthode aisée à mettre en oeuvre pour obtenir les motifs émergents ayant les meilleurs taux de croissance. Ces motifs, appelés motifs émergents forts, ont été exploités avec succès dans une collaboration avec la société Philips.

Mots clés : motifs émergents, représentations condensées, motifs fermés, caractérisation de classes.

1 Introduction

La caractérisation de classes et la classification sont d'importants domaines de recherche en fouille de données et apprentissage. Initialement introduits dans [Dong et Li, 1999], les motifs émergents sont des motifs dont la fréquence varie fortement entre deux classes. Ils caractérisent les classes de manière quantitative et qualitative. De par leur capacité à faire ressortir les distinctions entre classes, les motifs émergents permettent de construire des classifieurs ou de proposer une aide au diagnostic. Ils sont à l'origine de travaux variés et ils sont, entre autres, utilisés dans la réalisation de classifieurs performants [Dong *et al.*, 1999, Li *et al.*, 2000]. Dans un cadre plus applicatif, on peut citer différents travaux sur la caractérisation de propriétés biochimiques ou de données médicales [Li et Wong, 2001].

La recherche de tous les motifs émergents dans les grandes bases de données est une tâche difficile car le nombre de motifs candidats est très élevé. Nous verrons à la section 2.2 que les élagages utilisés par les algorithmes par niveaux [Mannila et Toivonen, 1997] souvent utilisés en fouille de données sont inadaptés. Les méthodes les plus classiques utilisent des manipulations de bordures [Dong et Li, 1999]. Sous un angle plus général,