Motifs séquentiels flous : un peu, beaucoup, passionnément

Céline Fiot, Anne Laurent, Maguelonne Teisseire

LIRMM - UM II, 161 rue Ada, 34392 Montpellier Cedex 5
fiot, laurent, teisseire@lirmm.fr

Résumé. La plupart des bases de données issues du monde réel sont constituées de données numériques et historisées (données de capteurs, données scientifiques, données démographiques). Dans ce cadre, les algorithmes d'extraction de motifs séquentiels, s'ils sont adaptés au caractère temporel des données, ne permettent pas le traitement de données numériques. Les données sont alors pré-traitées pour les transformer en données binaires, ce qui entraîne une perte d'information. Des algorithmes ont donc été proposés pour traiter les données numériques sous forme d'intervalles et d'intervalles flous notamment. En ce qui concerne la recherche de motifs séquentiels fondée sur des intervalles flous, les deux méthodes de la littérature ne sont pas satisfaisantes car incomplètes soit dans le traitement des séquences soit dans le calcul du support. Dans cet article, nous proposons donc trois méthodes d'extraction de motifs séquentiels flous (SpeedyFuzzy, MiniFuzzy et TotallyFuzzy) et en détaillons les algorithmes sous-jacents en soulignant les différents niveaux de fuzzification. Ces algorithmes sont implémentés et évalués à travers différentes expérimentations menées sur des jeux de tests synthétiques.

1 Introduction

La plupart des bases de données issues du monde réel sont constituées de données numériques et historisées (données de capteurs, données démographiques, ...). Dans le cadre de la fouille de grandes bases de données, peu de travaux ont été réalisés pour traiter cette problématique et la majorité des propositions sont restreintes aux règles d'association (Fu et al., 1998; Kuok et al., 1998; Srikant and Agrawal, 1996). Une première proposition (Srikant and Agrawal, 1996) traite les données quantitatives pour la recherche de règles d'association grâce à un découpage des attributs en intervalles discrets. Toutefois, ce découpage peut dissimuler des associations fréquentes en raison des bornes trop restrictives des différents intervalles. Plus récemment, l'utilisation de la théorie des sous-ensembles flous a permis des coupures moins brutales entre les intervalles, ce qui conduit à l'obtention de règles plus pertinentes.

Contrairement aux approches basées sur les règles d'association, les algorithmes d'extraction de motifs séquentiels permettent de prendre en compte le caractère temporel des données (suivi, évolution de phénomènes, concepts émergeants, détection ...) et sont donc adaptés aux données historisées. Néanmoins, ils ne permettent pas le traitement des données numériques. En effet, de telles données doivent être pré-traitées afin