Une méthode d'apprentissage par optimisation multicritère pour le rangement de motifs en fouille de données

Nassim Belmecheri * **, Noureddine Aribi*, Nadjib Lazaar**, Yahia Lebbah*, Samir Loudni***

*Lab. LITIO, Université Oran1, 31000 Oran, Algérie
belmecheri.nassim@edu.univ-oran1.dz,{ylebbah, aribi.noureddine}@gmail.com
**LIRMM, Université de Montpellier, CNRS, Montpellier, France
Nadjib.Lazaar@lirmm.fr
***TASC (LS2N-CNRS), IMT Atlantique, FR – 44307 Nantes, France
samir.loudni@imt-atlantique.fr

Résumé. La découverte de motifs pertinents est une tâche difficile en fouille de données. D'une part, des approches ont été proposées pour apprendre automatiquement des fonctions de rangement de motifs spécifiques à l'utilisateur. Ces approches sont souvent efficaces en qualité, mais très coûteuses en temps d'exécution. D'autre part, de nombreuses mesures d'intérêt sont utilisées pour évaluer l'intérêt des motifs dans le but de se rapprocher le plus possible du rangement de l'utilisateur. Dans cet article, nous formulons le problème d'apprentissage des fonctions de rangement des motifs comme un problème d'optimisation multicritère. L'approche proposée permet d'agréger des mesures d'intérêt en une fonction linéaire pondérée dont les poids sont calculés via la méthode AHP (Analytic Hierarchy Process). Des expérimentations menées sur de nombreux jeux de données montrent que notre approche réduit drastiquement le temps d'exécution, tout en assurant un rangement comparable à celui des approches existantes.

1 Introduction

La découverte des motifs est l'un des problèmes fondamentaux en fouille de données ensemblistes. En raison du grand nombre de motifs découverts dans un ensemble de données, les derniers travaux s'intéressent de plus en plus à l'extraction des motifs les plus pertinents. Cependant, l'utilisation des méthodes classiques d'extraction de motifs pose de nombreuses difficultés en raison de la combinatoire complexe pour garantir à la fois la qualité des motifs et un temps d'exécution acceptable. Les deux principaux problèmes sont : 1) des quantités importantes de motifs sont découverts, avec beaucoup de redondance ; 2) les préférences d'un expert du domaine ne sont pas prises en compte. L'importance de prendre en compte les préférences des utilisateurs a été discutée pour la première fois par Silberschatz et Tuzhilin (1995). L'idée de base est de modéliser les hypothèses des utilisateurs à l'aide de mesures d'intérêt qui se basent sur la structure des données. Toutefois, comme indiqué dans Bie (2011), l'utilisation de mesures d'intérêt est limitée lorsqu'il s'agit de capturer les préférences de l'utilisateur.