Découverte de sous-groupes avec les arbres de recherche de Monte Carlo

Guillaume Bosc*, Jean-François Boulicaut* Chedy Raïssi**
Mehdi Kaytoue*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
**INRIA Nancy - Grand Est, Villers-lès-Nancy, F-54600, France
prenom.nom@insa-lyon.fr, chedy.raissi@inria.fr

Résumé. Découvrir des règles qui distinguent clairement une classe d'une autre reste un problème difficile. De tels motifs permettent de suggérer des hypothèses pouvant expliquer une classe. La découverte de sous-groupes (Subgroup Discovery, SD), un cadre qui définit formellement cette tâche d'extraction de motifs, est toujours confrontée à deux problèmes majeurs: (i) définir des mesures de qualité appropriées qui caractérisent la singularité d'un motif et (ii) choisir une heuristique d'exploration de l'espace de recherche correcte lorsqu'une énumération complète est irréalisable. À ce jour, les algorithmes de SD les plus efficaces sont basés sur une recherche en faisceau (Beam Search, BS). La collection de motifs extraits manque cependant de diversité en raison de la nature gloutonne de l'exploration. Nous proposons ici d'utiliser une technique d'exploration récente, la recherche arborescente de Monte Carlo (Monte Carlo Tree Search, MCTS). Le compromis entre l'exploitation et l'exploration ainsi que la puissance de la recherche aléatoire permettent d'obtenir une solution disponible à tout moment et de surpasser généralement les approches de type BS. Notre étude empirique, avec plusieurs mesures de qualité, sur divers jeux de données de référence et du monde réel démontre la qualité de notre approche.

1 Introduction

L'extraction de groupes d'objets caractéristiques d'un attribut de classe a été intensément étudiée en fouille de données (Novak et al., 2009). La découverte de sous-groupes (Subgroup Discovery, SD) est une instance de ce problème (Wrobel, 1997). Étant donné un ensemble d'objets décrits par des attributs et associés à un ou plusieurs labels de l'attribut de classe, un sous groupe est un sous ensemble d'objets respectant une description sur les attributs. Le caractère discriminant d'un sous groupe est évalué par une mesure de qualité (F1 mesure, précision, etc.). Jusqu'à présent, puisque la taille de l'espace de recherche est exponentielle, les algorithmes les plus efficaces en SD sont basés sur une recherche en faisceau (Beam Search, BS) (van Leeuwen et Knobbe, 2012; Meeng et al., 2014; Duivesteijn et al., 2016).

Les problèmes principaux des approches heuristiques en SD sont (i) le manque de diversité des motifs extraits et (ii) la redondance : (i) une faible partie des optimums locaux de l'espace de recherche sont détectés, et (ii) plusieurs sous-groupes sont similaires à un même