Découverte d'un sous-groupe optimal dans des données purement numériques

Alexandre Millot*, Rémy Cazabet**, Jean-François Boulicaut*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
**Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
{prenom.nom}@insa-lyon.fr, {prenom.nom}@univ-lyon1.fr

Résumé. La découverte de sous-groupes dans des données étiquetées consiste à calculer des motifs dans un espace de description des objets pour faire émerger des ensembles d'objets qui ont une répartition particulière du point de vue des étiquettes, par exemple la surreprésentation d'une valeur. Découvrir des sous-groupes intéressants dans des données purement numériques - attributs et étiquette cible - a été peu traité. Généralement, on exploite des discrétisations qui engendrent une perte d'information et des résultats sous-optimaux. Nous traitons le problème du calcul d'un sous-groupe optimal au regard d'une mesure de qualité dans des données purement numériques. Nous exploitons des concepts de fermetures sur des motifs d'intervalles et des techniques d'élagage sophistiquées. Nous validons empiriquement la pertinence de notre algorithme et décrivons succinctement un cas d'application à l'optimisation de la pousse de végétaux en environnement contrôlé.

1 Introduction

La fouille de données numériques est pertinente dans de nombreux contextes applicatifs. On dispose alors de données sur des objets décrits par des valeurs d'attributs numériques. On peut considérer que l'un de ces attributs est une étiquette cible et vouloir mettre en oeuvre soit de l'apprentissage de modèles prédictifs de la valeur de cette étiquette pour de nouveaux objets soit, ce qui va nous intéresser ici, des méthodes de découverte de sous-groupes (Wrobel (1997)). Cette tâche consiste à chercher des sous-ensembles d'objets, des sous-groupes, démontrant des caractéristiques intéressantes d'après une mesure de qualité calculée sur une étiquette cible. La mesure de qualité doit capturer des différences sur la distribution de l'étiquette cible entre le sous-ensemble d'objets considéré et l'ensemble des objets du jeu de données. Un large éventail de méthodes exhaustives (Atzmueller et Puppe (2006) ; Grosskreutz et Paurat (2011)) et heuristiques (Mampaey et al. (2012) ; Bosc et al. (2017)) ont été proposées. Dans la majorité des cas, les approches considèrent des attributs nominaux et une étiquette binaire. Pour ce qui est du traitement d'attributs numériques, quelques travaux (Grosskreutz et Rüping (2009) ; Nguyen et Vreeken (2016)) présentent des méthodes permettant d'éviter une simple discrétisation des attributs. Malgré tout, aucune des méthodes que nous connaissons ne propose de solution exhaustive (et donc la possibilité d'identifier un optimum global de la mesure