Quand les sous-groupes rencontrent les graduels : découverte de sous-groupes identifiant des corrélations exceptionnelles

Mohamed-Ali Hammal*, Céline Robardet*
Marc Plantevit**

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621
**Université de Lyon, CNRS, Université Lyon 1, LIRIS UMR5205, F-69622, France
prénom.nom@liris.cnrs.fr

Résumé. La découverte de sous-groupes permet d'identifier des ensembles d'objets définis en intention qui sont intéressants vis-à-vis d'une mesure de qualité impliquant un ou plusieurs attributs cibles (par exemple motifs discriminants pour une variable de classe). Dans cet article nous proposons une approche pour un nombre quelconque (≥ 2) d'attributs cibles numériques. Pour cela, nous nous appuyons sur l'exploration conjointe de motifs graduels identifiant des corrélations de rang et de sous-groupes afin d'identifier des contextes pour lesquels les corrélations décrites par les motifs graduels sont exceptionnellement fortes par rapport au reste des données. Nous présentons un algorithme d'énumération s'appuyant sur des propriétés d'élagage avec des bornes supérieures. Une étude empirique sur plusieurs jeux de données démontre la pertinence et l'efficacité de notre méthode.

1 Introduction

Parmi les différentes techniques d'analyse exploratoire de données, la découverte de sous-groupes (Klösgen (1996)) vise à identifier des régions dans les données qui se détachent par rapport à une cible. Le principe est d'identifier des ensembles d'objets définis en intention qui sont fortement associés à certaines valeurs de la cible. Dans cet article nous proposons de généraliser cette approche au cas où l'on a plusieurs attributs cibles numériques. On cherche alors à la fois un sous-groupe d'objets défini par une conjonction de restrictions sur un ensemble d'attributs descriptifs et un sous-ensemble d'attributs cibles dont les valeurs sont fortement correlées sur cet ensemble. L'exploration conjointe de l'espace des descriptions et de l'espace des cibles permet de rechercher des corrélations pouvant être expliquées par d'autres variables descriptives de manière complètement non-supervisé.

Pour cela, nous introduisons le problème de **découverte de sous-groupes corrélés sur les rangs** basé sur l'exploration conjointe de motifs graduels identifiant des corrélations de rang et de sous-groupes afin d'identifier des contextes pour lesquels les corrélations sont exceptionnellement fortes par rapport au reste des données. Les motifs recherchés sont composés d'un ensemble D de conditions sur les attributs descriptifs, qu'ils soient numériques ou nominaux, et de C, un modèle de corrélation de rang sur des attributs numériques qui capture des corrélations de rang (positives ou négatives) basées sur une généralisation de τ de Kendall.