Détection de groupes atypiques pour une variable cible quantitative

Sylvie Guillaume*, Florian Guillochon*, Michel Schneider*

* Laboratoire LIMOS, UMR 6158 CNRS, Université Blaise Pascal
Complexe scientifique des Cézeaux, 63177 Aubière Cedex - France
sylvie.guillaume@isima.fr, flo.guillochon@orange.fr, michel.schneider@isima.fr

Résumé. Une tâche importante en analyse des données est la compréhension de comportements inattendus ou atypiques de groupes d'individus. Quelles sont les catégories d'individus qui gagnent de particulièrement forts salaires ou au contraire, quelles sont celles qui ont de très faibles salaires ? Nous présentons le problème d'extraction de tels groupes atypiques vis-à-vis d'une variable cible quantitative, comme par exemple la variable ″salaire″, et plus particulièrement pour les faibles et fortes valeurs d'un intervalle déterminé par l'utilisateur. Il s'agit donc de rechercher des conjonctions de variables dont la distribution diffère significativement de celle de l'ensemble d'apprentissage pour les faibles et fortes valeurs de l'intervalle de cette variable cible. Une adaptation d'une mesure statistique existante, l'intensité d'inclination, nous permet de découvrir de tels groupes atypiques. Cette mesure nous libère de l'étape de transformation des variables quantitatives, à savoir l'étape de discrétisation suivie d'un codage disjonctif complet. Nous proposons donc un algorithme d'extraction de tels groupes avec des règles d'élagage pour réduire la complexité du problème. Cet algorithme a été développé et intégré au logiciel d'extraction de connaissances WEKA. Nous terminons par un exemple d'extraction sur la base de données IPUMS du bureau de recensement américain.

1 Introduction

Un problème important en analyse des données est la compréhension de comportements inattendus ou atypiques de groupes d'individus. Quelles sont les catégories d'individus qui gagnent de particulièrement forts salaires ou au contraire, quelles sont celles qui ont de très faibles salaires ?

Notre but est de détecter automatiquement tous les groupes d'individus ayant un comportement différent de celui de l'ensemble d'apprentissage pour une variable quantitative donnée et plus particulièrement pour les faibles et les fortes valeurs d'un intervalle déterminé par l'utilisateur. Nous recherchons donc les motifs ou conjonctions de variables dont la distribution diffère significativement de celle de l'ensemble d'apprentissage pour les faibles et fortes valeurs de l'intervalle de cette variable cible.