Détection de données aberrantes à partir de motifs fréquents sans énumération exhaustive

Arnaud Giacometti, Arnaud Soulet

Université François-Rabelais de Tours, LI EA 6300
Campus de Blois, 41000 Blois
prenom.nom@univ-tours.fr

Résumé. La détection de données aberrantes (outliers) consiste à détecter des observations anormales au sein des données. Durant la dernière décennie, des méthodes de détection d'outliers utilisant les motifs fréquents ont été proposées. Elles extraient dans une première phase tous les motifs fréquents, puis assignent à chaque transaction un score mesurant son degré d'aberration (en fonction du nombre de motifs fréquents qui la couvre). Dans cet article, nous proposons deux nouvelles méthodes pour calculer le score d'aberration fondé sur les motifs fréquents (FPOF). La première méthode retourne le FPOF exact de chaque transaction sans extraire le moindre motif. Cette méthode s'avère en temps polynomial par rapport à la taille du jeu de données. La seconde méthode est une méthode approchée où l'utilisateur final peut contrôler l'erreur maximale sur l'estimation du FPOF. Une étude expérimentale montre l'intérêt des deux méthodes pour les jeux de données volumineux où une approche exhaustive échoue à calculer une solution exacte. Pour un même nombre de motifs, la précision de notre méthode approchée est meilleure que celle de la méthode classique.

1 Introduction

La détection des données aberrantes consiste à détecter les observations anormales au sein d'un jeu de données (Hawkins, 1980). Ce problème de détection des données aberrantes a d'importantes applications telles que la détection de fraudes bancaires ou d'intrusions réseau. Récemment, des méthodes de détection de données aberrantes ont été proposées pour les données catégorielles en utilisant le concept de motifs fréquents (He et al., 2005; Otey et al., 2006; Koufakou et al., 2011). L'idée clé de ces approches est de considérer le nombre de motifs fréquents couvrant chaque observation. Il est peu probable qu'une observation couverte par un grand nombre de motifs fréquents soit une donnée aberrante puisque les motifs fréquents correspondent aux « caractéristiques communes » du jeu de données. Ces méthodes de détection extraient d'abord tous les motifs fréquents du jeu de données et ensuite attribuent à chaque observation un score mesurant le degré d'aberration en comptabilisant les motifs fréquents qu'elle contient. Ces méthodes de détection de données aberrantes suivent donc un schéma en deux étapes : des motifs locaux vers un modèle global.

Les méthodes en deux étapes (Knobbe et al., 2008) visent à extraire exhaustivement tous les motifs locaux d'un jeu de données (première étape) afin de construire des modèles globaux