Extraction d'outliers dans des cubes de données : une aide à la navigation

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

LIRMM, Université Montpellier 2, CNRS,
161 Rue Ada 34392 Montpellier, France
nom.prenom@lirmm.fr

Résumé. La recherche d'algorithmes d'extraction de connaissances à partir de cubes de données est un domaine actuellement très actif qui trouve de très nombreuses applications dans les entrepôts de données disponibles maintenant dans la plupart des entreprises et milieux scientifiques (biologie, santé, etc.). Nous nous intéressons ici à l'extraction de comportements atypiques (dénommés outliers) dans de tels cubes de données quand l'utilisateur veut identifier des séquences anormales. Par exemple, un directeur marketing aimerait savoir quelle zone géographique ne suit pas le même comportement que les autres afin de pouvoir y remédier. Pour ce faire, nous définissons une mesure de similarité capable d'appréhender de telles données complexes et définissons les algorithmes associés que nous avons testés sur différentes bases. Notons que nous considérons des cubes de données très denses, ce qui complexifie le problème de l'extraction.

1 Introduction

Les techniques d'extraction de connaissances apportent une aide non négligeable dans le contexte OLAP où l'utilisateur doit désormais prendre les décisions les mieux adaptées en un minimum de temps. De façon plus précise, la fouille de données constitue une étape clef dans le processus de décision face à de gros volumes de données multidimensionnelles en fournissant des motifs ou règles permettant une autre appréhension des données sources. Nous pouvons citer en particulier les travaux de recherche de motifs dédiés au contexte multidimensionnel (Pinto et al. (2001); Plantevit et al. (2006); Messaoud et al. (2006)). Néanmoins et en particulier lorsque les données sont fortement corrélées, la véritable connaissance n'est pas toujours celle associée aux comportements fréquents. C'est ainsi que les événements rares deviennent plus intéressants et font l'objet du processus même d'extraction. Par exemple, un directeur marketing préférera connaître quels sont les individus qui ne suivent pas les directives plutôt que de savoir que la quasi totalité des représentants suivent ses recommandations. De nombreuses applications ont été développées pour la détection de fraudes, la surveillance des activités criminelles dans le commerce électronique, le suivi des athlétes basées sur cette recherche d'éléments, de motifs atypiques. Mais, à notre connaissance, il n'existe aucune proposition permettant d'extraire des séquences atypiques dans un contexte multidimensionnel. Il devient primordial d'être capable de fournir au décideur ce type de connaissances appelées motifs atypiques, rares ou outliers.

Notre contribution se situe dans ce contexte et propose une méthode de recherche de séquences