Détection et interprétation visuelle d'outliers dans les grands ensembles de données

Lydia BOUDJELOUD, François POULET

ESIEA – Pôle ECD
38, rue des docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé, 53000 Laval
{boudjeloud | poulet}@esiea-ouest.fr

Résumé. Nous présentons un algorithme hybride de détection d'outliers (individus atypiques) dans de grands ensembles de données, utilisant un algorithme génétique pour la sélection des attributs et une approche basée sur la distance pour la détection de l'élément outlier (atypique) suivant ce sous-ensemble d'attributs. Une fois l'outlier trouvé, nous essayons de l'expliquer : est-ce une erreur, un bruit ou une valeur significativement différente des autres ? Pour ce faire, on utilise des méthodes visuelles telles que les coordonnées parallèles. Nous évaluons les performances de notre méthode sur différents ensembles de données de grandes dimensions et le comparons avec les algorithmes existants.
Mots-clefs. fouille de données, détection d'outliers, visualisation, algorithmes génétiques, coordonnées parallèles, grands ensembles de données.

1. Introduction

Le développement du réseau internet et la baisse des coûts du matériel informatique ont permis à de nombreux organismes de constituer de grandes masses de données trop volumineuses et complexes pour pouvoir être appréhendées par un utilisateur. L'Extraction de Connaissances à partir de Données (ECD) est née de ce besoin, on la définit comme étant l'extraction de nouvelles connaissances potentiellement utiles à partir de grandes quantités de données [Fayyad et al. 1996], le cœur du processus d'ECD est la fouille de données (Data Mining). Dans ce cadre précis (de fouille de données), nous nous intéressons à la recherche d'outliers (individus atypiques). La recherche d'outliers a de nombreuses applications telles que la détection de fraudes, la recherche pharmaceutique, les applications financières, le marketing, etc. Un outlier (individu atypique) est un petit ensemble de données, un point ou une observation qui est considérablement différent, divergent, dissemblable ou distinct du reste des données. Le problème est alors de définir cette dissimilitude entre objets, ce qui caractérise un outlier. Typiquement, celle-ci est estimée par une fonction calculant la distance entre objets, la tâche suivante consiste à déterminer les objets les plus éloignés de la masse. Certaines difficultés apparaissent lorsque l'on est face à des ensembles de données ayant un grand nombre de dimensions en terme d'attributs. En effet, dans les ensembles à grandes dimensions, les données sont rares et la notion de voisinage perd de son sens. La rareté dans les espaces de grandes dimensions implique que tout point est candidat pour être un bon outlier et donc la recherche d'outliers devient complexe et coûteuse en temps de calcul.