Détection d'objets atypiques dans un flot de données : une approche multi-résolution

Alice Marascu et Florent Masseglia

INRIA Sophia Antipolis, 2004 route des lucioles - BP 93, FR-06902 Sophia Antipolis
Email: First.Last@sophia.inria.fr

1 Introduction

Les éléments atypiques (ou outliers) peuvent fournir des connaissances précieuses dans les domaines liés à la sécurité (*e.g.* détection de fraudes aux cartes de crédit, cyber securité ou sécurité des systèmes critiques. En général, l'atypicité dépend du degré d'isolation d'un (groupe d') enregistrement(s) en comparaison du reste des données. Pour découvrir les outliers, une méthode consiste à i) appliquer une technique de segmentation sur les données (afin d'obtenir des clusters) et ii) identifier les clusters qui correspondent à la notion d'atypicité selon un critère choisi (*e.g.* éloignement aux autres clusters, faible taille, grande densité...). À notre connaissance, les méthodes existantes pour la détection d'outlier reposent toujours sur un paramètre qui situe le degré d'atypicité au delà duquel les enregistrements doivent être considérés comme inhabituels (Knorr et Ng (1998)). Dans cet article, nous proposons DOO (Détection d'Outliers par les Ondelettes), une méhode sans paramètre destinée à l'extraction automatique d'outliers dans les résultats d'un algorithme de clustering. Notre méthode s'adapte à tous les résultats d'un algorithme de segmentation et toutes les caractéristiques peuvent être utilisées (distances entre objets, densité, taille des clusters).Dans un flot de données, les données sont générées à une vitesse et dans des quantités qui interdisent toute opération bloquante. Dans ce contexte, demander un paramètre tel que k, pour les top-k outliers, ou x, un pourcentage de clusters en queue de distribution, doit être évité. Premièrement, parce que l'utilisateur n'a pas assez de temps pour tester plusieurs paramètres. Deuxièmement, parce qu'une valeur choisie à un instant t dans le flot sera probablement inadapté au temps $t + n$. En effet, d'une fenêtre d'observation sur le flot, à l'autre, les résultats de la segmentation évoluent et la distribution des clusters change, ainsi que le nombre ou pourcentage d'outliers. Notre solution se base sur une analyse de la distribution des clusters, après les avoir triés par taille croissante. Une distribution classique est illustrée par la figure 1 (capture d'écran réalisée avec nos données réelles). L'idée de DOO est d'utiliser la transformée en ondelettes (Young (1995)) de cette distribution pour trouver la meilleure séparation. Du point de vue mathématique, la transformée en ondelettes continue est définie par :

$$T^{wav} f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\psi^*(\frac{x-b}{a})dx$$

où z^* dénote le nombre complexe conjugué de z, $\psi^*(x)$ est l'ondelette, a (> 0) est le facteur de mise à l'échelle et b est le paramètre de translation. On garde alors les deux coefficients les plus significatifs et les autres sont mis à zéro.