Détection de groupes outliers en classification non supervisée

Amine Chaibi[*,**], Mustapha Lebbah[*], Hanane Azzag[*],

[*] {prenom.nom}@lipn.univ-paris13.fr
[*]LIPN-UMR 7030
Université Paris 13 - CNRS
[**] Anticipeo
4 bis, impasse Courteline 94800 Villejuif, France

Résumé. Nous proposons dans ce papier une nouvelle méthode de détection de groupes outliers. Notre mesure nommée GOF (Group Outlier Factor) est estimée par l'apprentissage non-supervisé. Nous l'avons intégré dans l'apprentissage des cartes topologiques. Notre approche est basée sur la densité relative de chaque groupe de données, et fournit simultanément un partitionnement des données et un indicateur quantitatif (GOF) sur "la particularité" de chaque cluster ou groupe. Les résultats obtenus sont très encourageants et prometteurs pour continuer dans cette optique.

1 Introduction

Un outlier est un petit ensemble de données, une observation ou un point qui est différent du reste des données. Souvent, les outliers contiennent des informations précieuses sur le processus d'enregistrement et de la collecte de données. Ils peuvent aussi être problématiques, car ils risquent de biaiser les résultats, notamment pour les méthodes basées sur les distances entre individus, c'est pour cela que ces points devraient être étudiés attentivement.

L'analyse des outliers dans la littérature scientifique remonte à longtemps. En effet, les statisticiens se sont intéressés à ce genre de problématique afin de rendre les modèles mieux adaptés à leurs besoins. La détection des outliers dans les séries chronologiques représente un axe important dans ce genre de problématique, c'est ainsi que (Box et Tiao, 1965) ont étudié le changement de niveau dans les séries chronologiques. Le changement de la variance du premier ordre des modèles auto-régressifs a été examiné par (Wichern et al., 1976). L'approche Local Outlier Factor (LOF) apparue dans les années 2000 par (Breunig et al., 2000) reste la plus utilisée dans les modèles qui se basent sur la densité. L'avantage de cette méthode est qu'elle ne fait aucune hypothèse sur la distribution des données. Les auteurs de (Alhasan et al., 2009) ont donné une définition simplifiée de l'approche LOF. En effet, cette méthode consiste à comparer la densité locale d'une observation avec la densité moyenne de ses k-plus proches voisins (*k-ppv*). (Zengyou et al., 2003) ont utilisé LOF au niveau des clusters pour donner de l'importance aux données au niveau local. Le modèle utilisé permet d'affecter une mesure pour chaque cluster (CBLOF) afin d'identifier les outliers.

Les cartes auto-organisatrices aussi appelées carte de Kohonen ou carte SOM sont souvent