Détection de nouveautés en utilisant un nouveau score de détection de ”groupes-outliers“

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

* {prenom.nom}@lipn.univ-paris13.fr
*Université Paris 13, Sorbonne Paris Cité - CNRS
LIPN-UMR 7030
99, av. J-B Clément - F-93430 Villetaneuse

Résumé. Dans cet article, nous introduisons une nouvelle mesure pour qualifier “l'outlier-ness“ de chaque groupe/cluster. Cette mesure, nommée GOF, est intégrée et estimée dans un processus d'apprentissage non supervisé en utilisant les cartes topologiques. Ceci permet d'apprendre la structure des données tout en fournissant un nouveau score (GOF). Ce paramètre est basé sur la densité et quantifie ainsi la particularité de chaque groupe (cluster) : plus la valeur est grande, plus le groupe est susceptible d'être un ”groupe-outlier“. GOF est utilisé par la suite comme classifieur pour le problème de détection de nouveautés.

1 Introduction

Avec la quantité croissante des données recueillies, il devient plus important et difficile de repérer les observations inhabituelles ou inattendues. Un tel comportement inattendu peut être soit non désirée (par exemple, la détection d'intrusion réseau, la surveillance des maladies), nécessitant une intervention de l'utilisateur, ou intéressant (par exemple en astronomie), ce qui conduit à une meilleure compréhension du système. La tâche de détection d'outliers joue un rôle important, puisque dans la plupart des cas, elle permet de prévenir ou d'atténuer les effets d'une situation indésirable.

Les données sont généralement un ensemble d'enregistrements décrite par un ensemble d'attributs (ou caractéristiques). D'une manière générale, les outliers peuvent être soit des outliers individuels (un seul enregistrement) ou des “groupes-outliers“, aussi appelés outliers collectifs (correspondant à des groupes d'enregistrements). Dans le cas de la détection d'outlier individuel, une approche standard consiste à créer un modèle de données normales, et de comparer les enregistrements de la base de test. Cependant, dans le cas étudié dans ce papier qui concerne les ”groupes-outliers”, plutôt que de trouver des comportements individuels anormaux (bruit ou des erreurs dans les données), nous nous sommes intéressés plus par la détection de l'émergence de nouvelles observations, qui ne peuvent être expliquées par un précédent modèle. En général, ces comportements donnent lieu à des enregistrements multiples dans un même ensemble de données formant un groupe dense et significativement isolé. Notre objectif dans ces travaux est d'utiliser la présence de ces multiples cas afin de mieux détecter