Apprentissage d'ensemble d'opérateurs de projection orthogonale pour la détection de nouveauté

Fatma Hamdi*, Younès Bennani*

* LIPN-UMR 7030, Université Paris 13,
99, av. J-B Clément, 93430 Villetaneuse, France,
Prénom.Nom@lipn.univ-paris13.fr

Résumé. Dans ce papier, nous proposons une approche de détection de nouveauté fondée sur les opérateurs de projection orthogonale et l'idée de double bootstrap (bi- bootstrap). Notre approche appelée Random Subspace Novelty Detection Filter (RS-NDF), combine une technique de rééchantillonnage et l'idée d'apprentissage d'ensemble. RS-NDF est un ensemble de filtres NDF (Novelty Detection Filter), induits à partir d'échantillons bootstrap des données d'apprentissage, en utilisant une sélection aléatoire des variables pour l'apprentissage des filtres. RS-NDF utilise donc un double bootstrap, c'est à dire un rééchantillonnage avec remise sur les observations et un rééchantillonnage sans remise sur les variables. La prédiction est faite par l'agrégation des prédictions de l'ensemble des filtres. RS-NDF présente généralement une importante amélioration des performances par rapport au modèle de base NDF unique. Grâce à son algorithme d'apprentissage en ligne, l'approche RS-NDF est également en mesure de suivre les changements dans les données au fil du temps. Plusieurs métriques de performance montrent que l'approche proposée est plus efficace, robuste et offre de meilleures performances pour la détection de nouveauté comparée aux autres techniques existantes.

1 Introduction

Plusieurs travaux de recherche ont été proposées pour le problème de la détection de nouveauté (Markou et Singh, 2003a,b) et (Markou et Singh, 2003c) avec une grande variété d'applications et méthodes. Le but essentiel de la détection de nouveauté conciste à apprendre un modèle ou un ensemble de modèles sur des données disponibles, et l'utiliser aprés pour identifier les données nouvelles (nouveauté). Les applications typiques de ce problème sont la détection des fraudes, la maintenance préventive, la détection des intrusions dans le réseau, le diagnostic de maladies rares et de nombreux autres domaines. La détection de nouveauté est particulièment utile quand une classe importante est sous représentée dans les données. L'exemple typique de ce problème est la détection des fraudes oú il peut y avoir un intervalle de plusieurs heures entre deux transactions frauduleuses. On distingue trois grandes familles d'approches de détection de nouveauté :

– Les méthodes qui déterminent la nouveauté sans aucune connaissance préalable sur les données. Il s'agit essentiellement d'approches d'apprentissage analogues à la classifica-