Détection de changements de distribution dans un flux de données : une approche supervisée

Alexis Bondu*, Marc Boullé**

*EDF R&D ICAME/SOAD, 1 avenue du Général de Gaulle, 92140 Clamart.
alexis.bondu@edf.fr
**Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion.
marc.boulle@orange-ftgroup.com

Résumé. L'analyse de flux de données traite des données massives grâce à des algorithmes en ligne qui évitent le stockage exhaustif des données. La détection de changements dans la distribution d'un flux est une question importante dont les applications potentielles sont nombreuses. Dans cet article, la détection de changement est transposée en un problème d'apprentissage supervisé. Nous avons choisi d'utiliser la méthode de discrétisation supervisée MODL car celle-ci présente des propriétés intéressantes. Notre approche est comparée favorablement à une méthode de l'état-de-l'art sur des flux de données artificiels.

1 Introduction

Ces dernières années, la quantité de données à traiter a considérablement augmenté dans de nombreux domaines applicatifs. Les méthodes d'apprentissage "classiques" passent difficillement à l'échelle sur de tels volumes de données. L'analyse de flux de données (ou *"Data stream mining"*) est une des réponses possibles au traitement des données massives. Le paradigme des "flux de données" prend en compte plusieurs contraintes : i) l'ordre d'arrivé des données (ou tuples [1]) n'est pas contrôlé ; les débits des flux d'entrée sont fluctuants et ne sont pas contrôlés ; les ressources matérielles disponibles sont limitées (mémoire RAM et CPU). Du fait de ces contraintes, les tuples ne peuvent pas être stockés exhaustivement et sont analysés à la volée. Une analyse sur flux de données fournit un résultat en "temps réel" qui évolue continuellement. L'objectif est de maximiser la qualité de l'analyse, compte tenu des contraintes imposées par les flux, et étant donnés les ressources matérielles disponibles. La mise au point de méthodes de détection de changements de distribution qui soient génériques, capables de passer à l'échelle, et pertinentes d'un point de vue statistique est un véritable challenge. La détection de changements dans un flux de données consiste à comparer la distribution des tuples observés sur deux fenêtres temporelles distinctes : la *"référence"* et la *"courante"*. Dans le cadre de cet article, nous traitons le cas de la détection relative à un régime de fonctionnement normal. Dans la littérature, il existe plusieurs manières d'aborder le problème de la détection de changements. Par exemple, un test statistique impliquant deux échantillons de tuples (Dries et Rückert, 2009) peut être utilisé. Le test de *"Wald-Wolfowitz et Smirnov"* a notamment été généralisé au cas des données multidimensionnelles (Friedman et Rafsky, 2006), ce qui est

1. Les données élémentaires émises dans un flux sont appelées des *"tuples"*