Détection de changements dans des flots de données qualitatives

Dino Ienco[*,***], Albert Bifet[**],
Bernhard Pfahringer[***], Pascal Poncelet[*]

[*]Irstea, UMR TETIS, Montpellier, France
LIRMM, Montpellier, France
{dino.ienco@irstea.fr, pascal.poncelet@lirmm.fr}

[**]Yahoo! Research Barcelona, Catalonia, Spain
abifet@yahoo-inc.com

[***]University of Waikato, Hamilton, New Zealand
bernhard@cs.waikato.ac.nz

Résumé. Pour mieux analyser et extraire de la connaissance de flots de données, des approches spécifiques ont été proposées ces dernières années. L'un des challenges auquel elles doivent faire face est la détection de changement dans les données. Alors que de plus en plus de données qualitatives sont générées, peu de travaux de recherche se sont intéressés à la détection de changement dans ce contexte et les travaux existants se sont principalement focalisés sur la qualité d'un modèle appris plutôt qu'au réel changement dans les données. Dans cet article nous proposons une nouvelle méthode de détection de changement non supervisée, appelée *CDCStream* (*Change Detection in Categorical Data Streams*), adaptée aux flux de données qualitatives.

1 Introduction

De nombreux domaines d'application génèrent en permanence d'énormes quantités de données. L'un des défis essentiels auquel les approches de fouilles de flots doivent faire face est la détection de changement dans ces données. En effet, l'information disponible dans les flots change et évolue au fil du temps et les connaissances acquises au préalable peuvent s'avérer non représentatives des nouvelles données. Dans un contexte d'apprentissage, cela se traduit par le fait que des classes ou des concepts sous représentés (resp. surreprésentés) peuvent apparaître surreprésentés (resp. sous représentés) après une période plus longue. Savoir détecter le plus tôt possible les réels changements dans le flot permet alors de pouvoir réévaluer automatiquement les apprentissages précédents et surtout garantir que la connaissance extraite à un moment donné est vraiment représentative des données disponibles sur le flot. Dans cet article, nous proposons une nouvelle méthode de détection de changement définie pour traiter des donnés qualitatives : *CDCStream* (*Change Detection in Categorical data Streams*). L'une