Grille bivariée pour la détection de changement dans un flux étiqueté

Christophe Salperwyck[*,**], Marc Boullé[*], Vincent Lemaire[*]

[*]Orange Labs
2, Avenue Pierre Marzin 22300 Lannion
prenom.nom@orange.com
[**] LIFL (UMR CNRS 8022) - Université de Lille 3
Domaine Univ. du Pont de Bois - 59653 Villeneuve d'Ascq Cedex

Résumé. Nous présentons une méthode en-ligne de détection de changement de concept dans un flux étiqueté. Notre méthode de détection est basée sur un critère supervisé bivarié qui permet d'identifier si les données de deux fenêtres proviennent ou non de la même distribution. Notre méthode a l'intérêt de n'avoir aucun *a priori* sur la distribution des données, ni sur le type de changement et est capable de détecter des changements de différentes natures (changement dans la moyenne, dans la variance...). Les expérimentations montrent que notre méthode est plus performante et robuste que les méthodes de l'état de l'art testées. De plus, à part la taille des fenêtres, elle ne requiert aucun paramètre utilisateur.

1 Introduction

De nombreux acteurs de l'informatique doivent faire face à l'arrivée massive de données. Les plus connus sont Google et Yahoo avec le traitement des logs pour la publicité en-ligne, Facebook et Twitter qui modélisent les données provenant de leurs centaines de millions d'utilisateurs, les opérateurs téléphoniques pour la gestion de réseaux de télécommunications. La volumétrie de ces données continue de croître rapidement et les quantités ne sont plus compatibles avec l'utilisation de la plupart des méthodes hors-ligne qui supposent de pouvoir accéder à toutes les données. Dans ces conditions, il est préférable de traiter les données à leur passage ce qui impose d'y accéder une seule fois et dans leur ordre d'arrivée. On parle alors d'un accès sous la forme d'un flux de données.

En classification supervisée, on appelle concept $P(C|X)$ la probabilité conditionnelle de la classe C connaissant les données X. Les flux de données peuvent ne pas être stationnaires et comporter des changements de concept si le processus qui génère les données varie au cours du temps. Dans ce cas le modèle de classification supervisée doit être adapté au fur et à mesure que le concept change.

Cet article propose une nouvelle méthode de détection de changement basée sur l'observation des données du flux. Notre méthode utilise deux fenêtres et permet d'identifier si les données de ces deux fenêtres proviennent ou non de la même distribution. Elle est capable de détecter les changements de diverses natures (moyenne, variance...) sur la distribution des