Classification incrémentale supervisée : un panel introductif

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

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

Résumé. Les dix dernières années ont été témoin du grand progrès réalisé dans le domaine de l'apprentissage statistique et de la fouille de données. Il est possible à présent de trouver des algorithmes d'apprentissage efficaces et automatiques. Historiquement les méthodes d'apprentissage faisaient l'hypothèse que toutes les données étaient disponibles et pouvaient être chargées en mémoire pour réaliser l'apprentissage. Mais de nouveaux domaines d'application de la fouille de données émergent telles que : la gestion de réseaux de télécommunications, la modélisation des utilisateurs au sein d'un réseau social, le web mining... La volumétrie des données explose et il est nécessaire d'utiliser des algorithmes d'apprentissage incrémentaux. Cet article a pour but de présenter les principales approches de classification supervisée incrémentale recensées dans la littérature. Il a pour vocation de donner à un lecteur débutant des indications de lecture sur ce sujet; sujet qui connaît déjà des applications industrielles.

1 Introduction

Les dix dernières années ont été témoin des grands progrès réalisés dans le domaine de l'apprentissage automatique et de la fouille de données. Ces techniques ont montré leurs capacités à traiter des volumétries importantes de données et ce sur des problèmes réels (Guyon et al., 2009; Féraud et al., 2010). Néanmoins le plus important effort a été réalisé pour des analyses de données homogènes et stationnaires et à l'aide d'algorithmes centralisés. La plupart des approches d'apprentissage automatique supposent que les ressources sont illimitées, par exemple que les données tiennent en mémoire vive. Dans ce contexte les algorithmes classiques d'apprentissage utilisent des bases d'apprentissage de taille finie et produisent des modèles statiques. Cependant la volumétrie des données continue de croître et ce plus vite que les capacités de traitement.

De nouveaux domaines d'application de la fouille de données émergent où les données ne sont plus sous la forme de tableaux de données persistants mais plutôt sous la forme de données "passagères". Parmi ces domaines on citera : la gestion de réseaux de télécommunications, la modélisation des utilisateurs au sein d'un réseau social, le web mining... Le défi scientifique principal est alors d'automatiser l'ensemble des étapes du processus d'apprentissage mais aussi