Une méthode de classification supervisée sans paramètre pour l'apprentissage sur les grandes bases de données

Marc Boullé*

*Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Dans ce papier, nous présentons une méthode de classification supervisée sans paramètre permettant d'attaquer les grandes volumétries. La méthode est basée sur des estimateurs de densités univariés optimaux au sens de Bayes, sur un classifieur Bayesien naïf amélioré par une sélection de variables et un moyennage de modèles exploitant un lissage logarithmique de la distribution a posteriori des modèles. Nous analysons en particulier la complexité algorithmique de la méthode et montrons comment elle permet d'analyser des bases de données nettement plus volumineuses que la mémoire vive disponible. Nous présentons enfin les résultats obtenu lors du récent PASCAL Large Scale Learning Challenge, où notre méthode a obtenu des performances prédictives de premier plan avec des temps de calcul raisonnables.

1 Introduction

La phase de préparation des données est particulièrement importante dans le processus data mining (Pyle, 1999). Elle est critique pour la qualité des résultats, et consomme typiquement de l'ordre de 80% du temps d'une étude data mining. Dans le cas d'une entreprise comme France Télécom, le data mining est appliqué dans de nombreux domaines : marketing, données textuelles, données du web, classification de trafic, sociologie, ergonomie. Les données disponibles sont hétérogènes, avec des variables numériques ou catégorielles, des variables cibles comportant de multiples classes, des valeurs manquantes, des distributions bruitées et déséquilibrées, des nombres de variables et d'instances pouvant varier sur plusieurs ordres de grandeurs. Ce contexte industriel impose des contraintes telles que le potentiel des données collectées dans les systèmes d'information est largement sous-utilisé. Cette situation s'aggrave années après années, suite à des vitesses d'évolution divergentes des capacités des systèmes d'information, en augmentation très rapide pour le stockage et "seulement" rapide pour le traitement, des capacités de modélisation des méthodes d'apprentissage statistique, en progression lente, et de la disponibilité des analystes de données, au mieux constante. Dans ce contexte, les solutions actuelles sont impuissantes à répondre à la demande rapidement croissante de l'utilisation de techniques data mining. Les projets, en surnombre, sont abandonnés ou traités sous-optimalement. Pour résoudre ce goulot d'étranglement, nous nous intéressons ici au problème de l'automatisation de la phase de préparation des données du processus data mining.