SVM incrémental, parallèle et distribué pour le traitement de grandes quantités de données

Thanh-Nghi Do*, François Poulet**

*College of Information Technology, Cantho University
1 Ly Tu Trong street, Cantho City, Vietnam
dtnghi@cit.ctu.edu.vn
**ESIEA - Pôle ECD
38, rue des Docteurs Calmette et Guérin, 53000 Laval - France
poulet@esiea-ouest.fr

Résumé. Nous présentons un nouvel algorithme de SVM (Support Vector Machine ou Séparateur à Vaste Marge) linéaire et non-linéaire, parallèle et distribué permettant le traitement de grands ensembles de données dans un temps restreint sur du matériel standard. A partir de l'algorithme de Newton-GSVM proposé par Mangasarian, nous avons construit un algorithme incrémental, parallèle et distribué permettant d'améliorer les performances en temps d'exécution et mémoire en s'exécutant sur un groupe d'ordinateurs. Ce nouvel algorithme a la capacité de classifier un million d'individus en 20 dimensions et deux classes en quelques secondes sur un ensemble de dix PC.

1 Introduction

A l'heure actuelle, les données arrivent plus vite que la capacité de traitement des algorithmes de fouille de données ne permet de les traiter. L'amélioration des performances des algorithmes de fouille de données est indispensable pour traiter de grands ensembles de données. Nous nous intéressons au cas de la classification supervisée et plus particulièrement à une classe d'algorithmes : les SVM [Vapnik, 1995]. En règle générale, ils donnent de bons taux de précision mais, l'apprentissage des SVM se ramène à résoudre un programme quadratique et est donc coûteux en temps et mémoire. Pour remédier à ce problème, les méthodes de décomposition [Platt, 1999], [Chang et Lin, 2003] travaillent sur des sous-ensembles arbitraires de données, on utilise alors des heuristiques [Do et Poulet, 2005] permettant de choisir les sous-ensembles de données. D'autres travaux visent à construire des algorithmes incrémentaux [Fung et Mangasarian, 2002] dont le principe est de ne charger qu'un petit bloc de données en mémoire à la fois, de construire un modèle partiel et de le mettre à jour en chargeant consécutivement des blocs de données. Les SVMs parallèles et distribués [Poulet et Do, 2004] utilisent un réseau de machines pour améliorer les performances. Nous présentons un nouvel algorithme de SVM linéaire et non-linéaire pour traiter de grands ensembles de données dans un temps restreint sur du matériel standard. A partir de l'algorithme de Newton-GSVM [Mangasarian, 2001], nous avons construit un algorithme incrémental, parallèle et distribué permettant d'améliorer les performances en temps d'exécution et mémoire en s'exécutant sur un groupe d'ordinateurs. Les résultats