Classification de grands ensembles de données avec un nouvel algorithme de SVM

Thanh-Nghi Do*, François Poulet**

*Equipe InSitu, INRIA Futurs, LRI, Bat.490, Université Paris Sud 91405 Orsay Cedex
Thanh-Nghi.Do@lri.fr
http://www.lri.fr/~dtnghi
**ESIEA-Ouest, 38, rue des Docteurs Calmette et Guérin, 53000 Laval
francois.poulet@esiea-ouest.fr
http://visu.egc.free.fr

Résumé. Le nouvel algorithme de boosting de Least-Squares Support Vector Machine (LS-SVM) que nous présentons vise à la classification de très grands ensembles de données sur des machines standard. Les méthodes de SVM et de noyaux permettent d'obtenir de bons résultats en ce qui concerne la précision mais la tâche d'apprentissage pour de grands ensembles de données demande une grande capacité mémoire et un temps relativement long. Nous présentons une extension de l'algorithme de LS-SVM proposé par Suykens et Vandewalle pour le boosting de LS-SVM. A cette fin, nous avons ajouté un terme de régularisation de Tikhonov et utilisé la formule de Sherman-Morrison-Woodbury pour traiter des ensembles de données ayant un grand nombre de dimensions. Nous l'avons ensuite étendu par application du boosting de LS-SVM afin de traiter des données ayant simultanément un grand nombre d'individus et de dimensions. Les performances de l'algorithme sont évaluées sur les ensembles de données de l'UCI, Twonorm, Ringnorm, Reuters-21578 et NDC sur une machine standard (PC-P4, 3GHz, 512 Mo RAM).

1 Introduction

Le volume de données stocké double actuellement tous les 9 mois (Lyman et al, 2003) et donc le besoin d'extraction de connaissances dans les grandes bases de données est de plus en plus important (Fayyad et al, 2004). La fouille de données (Fayyad et al, 1996) est confrontée au challenge de traiter de grands ensembles de données pour identifier des connaissances nouvelles, valides, potentiellement utilisables et compréhensibles. Elle utilise différents algorithmes pour la classification, la régression, le clustering ou les associations.

Nous nous intéressons plus particulièrement ici aux algorithmes de Séparateurs à Vaste Marge (SVM ou Support Vector Machine) proposé par (Vapnik, 1995) car ils se montrent particulièrement efficaces pour la classification, la régression ou la détection de nouveauté. On peut trouver de nombreuses applications des SVM comme la reconnaissance de visages, la catégorisation de textes ou la bioinformatique (Guyon, 1999). L'approche est systématique et motivée par la théorie de l'apprentissage statistique. Les SVM sont les plus connus parmi une classe d'algorithmes utilisant les méthodes de noyau (Cristianini et al, 2000). Les SVM