Algorithmes rapides de boosting de SVM

Thanh-Nghi Do*, Jean-Daniel Fekete*, François Poulet**

*Equipe Aviz, INRIA Futurs, LRI
Bât.490, Université Paris Sud 91405 Orsay Cedex
{dtnghi@lri.fr | Jean-Daniel.Fekete@inria.fr}
{http://www.lri.fr/~dtnghi | http://www.lri.fr/~fekete}
**IRISA TexMex, Université de Rennes I
Campus de Beaulieu, 35042 Rennes Cedex
francois.poulet@irisa.fr
http://www.irisa.fr/texmex/people/poulet/index_fr.php

Résumé. Les algorithmes de boosting de Newton Support Vector Machine (NSVM), Proximal Support Vector Machine (PSVM) et Least-Squares Support Vector Machine (LS-SVM) que nous présentons visent à la classification de très grands ensembles de données sur des machines standard. Nous présentons une extension des algorithmes de NSVM, PSVM et LS-SVM, pour construire des algorithmes de boosting. A cette fin, nous avons utilisé un terme de régularisation de Tikhonov et le théorème Sherman-Morrison-Woodbury pour adapter ces algorithmes au traitement d'ensembles de données ayant un grand nombre de dimensions. Nous les avons ensuite étendus par construction d'algorithmes de boosting de NSVM, PSVM et LS-SVM afin de traiter des données ayant simultanément un grand nombre d'individus et de dimensions. Les performances des algorithmes sont évaluées sur des grands ensembles de données de l'UCI comme Adult, KDDCup 1999, Forest Covertype, Reuters-21578 et RCV1-binary sur une machine standard (PC-P4, 2,4 GHz, 1024 Mo RAM).

1 Introduction

Les algorithmes de Séparateurs à Vaste Marge proposés par (Vapnik, 1995) et les méthodes de noyaux permettent de construire des modèles précis et deviennent des outils de classification de données de plus en plus populaires. On peut trouver de nombreuses applications des SVM (réf. http://www.clopinet.com/isabelle/Projects/SVM/applist.html) comme la reconnaissance de visages, la catégorisation de textes ou la bioinformatique. Cependant, les SVM demandent la résolution d'un programme quadratique dont le coût de calcul est au moins d'une complexité égale au carré du nombre d'individus de l'ensemble d'apprentissage et la quantité de mémoire nécessaire les rend impossible à utiliser sur de grands ensembles de données à l'heure actuelle (Lyman et al., 2003). Il y a besoin de permettre le passage à l'échelle des SVM pour traiter de grands ensembles de données sur des machines standard. Une heuristique possible pour améliorer l'apprentissage des SVM est de décomposer le programme quadratique en une série de plus petits problèmes (Boser et al, 1992), (Chang et al, 2003), (Osuna et al, 1997), (Platt, 1999). Au niveau de la mise en œuvre,