Fouille de Grands Ensembles de Données avec un Boosting de Proximal SVM

Thanh-Nghi Do*, François Poulet*

*ESIEA Recherche
38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval - Changé
53000 Laval
(dothanh, poulet)@esiea-ouest.fr

Résumé. Les SVM (support vector machines) ont montré leur efficacité dans plusieurs domaines d'application. L'apprentissage des SVM se ramène à résoudre un programme quadratique, dont la mise en œuvre est en général coûteuse en temps. Une reformulation plus récente des SVM (proximal SVM), proposée par Fung et Mangasarian, ne nécessite que la résolution d'un système linéaire, cet algorithme de PSVM est plus efficace et permet de traiter des données dont le nombre d'individus est très important (10^9) et le nombre d'attributs plus restreint (10^4). Nous proposons d'utiliser la formule de Sherman-Morrison-Woodbury pour adapter le PSVM à la fouille d'ensembles de données dont le nombre d'attributs est très important et le nombre d'individus plus restreint sur un matériel standard. Puis nous présentons un algorithme de boosting de PSVM pour classifier des données de très grandes tailles en nombre d'individus et d'attributs. Nous évaluons les performances du nouvel algorithme sur les ensembles de données de l'UCI, Twonorm, Ringnorm, Reuters-21578 et Ndc.

1. Introduction

La fouille de données est un domaine récent de l'informatique dont le développement est lié aux masses de données de plus en plus importantes qui sont stockées à l'heure actuelle. Son but est de pouvoir extraire des informations intéressantes de ces bases de données. De nombreux algorithmes venant de différents domaines de recherche sont utilisés pour l'extraction de connaissances (intelligence artificielle, statistique, analyse de données, bases de données…). Parmi eux, on trouve les arbres de décision, les règles d'association, les SVM. Nous nous intéressons particulièrement à une classe d'algorithmes d'apprentissage supervisé : les SVM ou Séparateurs à Vaste Marge. En fournissant des outils performants de classification, régression et détection de nouveauté [Bennett et Campbell, 2000], les algorithmes de SVM ont été utilisés dans plusieurs applications comme : la reconnaissance de visages, la reconnaissance de chiffres manuscrits, la classification de textes ou la bioinformatique [Guyon, 1999]. En général, ils donnent de bons taux de précision. L'apprentissage des SVM se ramène à résoudre un programme quadratique, la mise en œuvre d'un algorithme de SVM est donc coûteuse en temps.

Récemment, Fung et Mangasarian ont proposé un algorithme de proximal SVM qui n'a besoin que de la résolution d'un système linéaire. Le PSVM offre des particularités intéressantes :