SVM et Visualisation pour la Fouille de Grands Ensembles de Données

Thanh-Nghi Do*, François Poulet*

*ESIEA Recherche, BP 0339, 53003 Laval-France
(dothanh, poulet)@esiea-ouest.fr

Résumé. Nous présentons un algorithme de SVM et des méthodes graphiques pour le traitement de grands ensembles de données. Pour pouvoir traiter de tels ensembles de données, nous utilisons une représentation des données de plus haut niveau (sous forme symbolique). L'algorithme de séparateur à vaste marge (SVM) est adapté pour pouvoir traiter ce nouveau type de données. Nous construisons un nouveau noyau RBF (Radial Basis Function) que l'algorithme utilise à la fois pour la classification, la régression et la détection d'individus atypiques dans des données de type intervalle. Nous utilisons ensuite des méthodes de visualisation interactive (elles aussi adaptées au cas des variables de type intervalle) pour expliquer les résultats obtenus par les SVM. La méthode est évaluée sur des ensembles de données symboliques existant ou créés artificiellement.

1 Introduction

Le début des années 2000 voit la quantité d'information stockée dans le monde croître de manière très importante. On estime qu'elle augmente de deux exa (10^{18}) octets tous les ans (Lyman et al. 2003). Une telle masse d'information est trop complexe pour pouvoir être appréhendée simplement par un utilisateur. L'extraction de connaissances à partir de données (ECD) s'est développée pour pouvoir découvrir des connaissances à partir de très grandes quantités d'information. Le processus d'ECD (Fayyad et al. 1996) est un processus non trivial permettant d'identifier des structures inconnues, valides et potentiellement exploitables dans les bases de données.

Dans cette problématique, nous nous sommes plus particulièrement intéressés à une méthode récente de fouille de données à l'aide de SVM (Vapnik 1995). Les SVM et les méthodes de noyaux sont reconnues comme une méthodologie efficace pour la résolution de plusieurs problèmes : la classification supervisée, la régression, la détection d'individus atypiques, le clustering. Les SVM donnent de bons résultats dans la pratique en ce qui concerne le taux de précision, mais ils nécessitent la résolution d'un programme quadratique dont la mise en œuvre est coûteuse en temps d'exécution et mémoire. Un autre inconvénient est que les SVM fournissent très peu d'informations en sortie, ils ne retournent que les vecteurs support pour construire la frontière de séparation des données. L'utilisateur peut se servir de cette frontière pour classifier ses données avec de bons taux de précision mais il ne peut pas expliquer le modèle obtenu. Or la compréhensibilité des résultats est elle aussi importante même si elle n'apparaît pratiquement jamais dans l'évaluation des algorithmes de fouille de données. L'interprétation des résultats de SVM est nécessaire pour permettre à l'utilisateur de comprendre les résultats et cela augmente sa confiance dans ces résultats.