Vis-SVM : approche coopérative en fouille de données

Thanh-Nghi Do, François Poulet

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

Résumé. La compréhension des résultats en sortie d'un algorithme de fouille de données est aussi importante que d'obtenir de bons taux de précision. Malheureusement, les modèles obtenus par les algorithmes de support vector machines ou séparateurs à vaste marge (SVM) fournissent seulement les vecteurs support qui sont utilisés comme une « boîte noire » pour classifier efficacement les données avec de bons taux de précision. Il est donc indispensable d'améliorer la compréhensibilité des modèles de SVM. Cet article présente différentes coopérations entre des méthodes de visualisation et des algorithmes de SVM en fouille de données. En post-traitement d'algorithmes de SVM, nous présentons une approche coopérative graphique interactive pour interpréter des résultats de classification, régression et détection d'individus atypiques. Nous étendons l'approche d'interprétation graphique pour améliorer les résultats obtenus par la classification de SVM. Nous présentons ensuite une approche coopérative permettant d'impliquer plus significativement l'utilisateur dans la tâche de classification à l'aide de SVM. Ce type d'approche présente notamment comme avantage la possibilité d'utiliser les capacités humaines en reconnaissance de formes par le biais de méthodes de visualisation. L'utilisateur a une meilleure compréhension du modèle construit et une meilleure confiance dans ce modèle parce qu'il a participé activement à sa construction. Nous montrons comment l'utilisateur peut utiliser des outils coopératifs pour construire des modèles de SVM. Une étape de prétraitement est également utilisée dans notre outil coopératif pour pouvoir traiter de grands ensembles de données. Nous évaluons les performances de la nouvelle approche coopérative sur les ensembles de données de l'UCI, Delve, Statlog et biomédicales.

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. D'après [Fayyad et al., 1996], la définition de l'ECD est : « un processus non trivial d'identification de connaissances inconnues, valides, potentiellement exploitables et compréhensibles dans les données ». Ce processus est complexe, il vise à exploiter des techniques venant de différents domaines de recherches (intelligence artificielle, apprentissage automatique, statistique, analyse de données, visualisation d'informations, bases de données) pour l'extraction de connaissances. Parmi elles, on trouve les arbres de