Sélection de variables et agrégation d'opinions

Gaëlle Legrand et Nicolas Nicoloyannis*

*Laboratoire ERIC
Université Lumière Lyon 2
Batiment L
5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
glegrand@eric.univ-lyon2.fr ; nicolas.nicoloyannis@univ-lyon2.fr

Résumé. La taille des bases de données étant de plus en plus importante, le processus de sélection de variables devient essentiel. Nous proposons une méthode de sélection, pour les variables qualitatives, basée sur l'agrégation d'opinion. Le résultat, sous forme d'un préordre de variables, est fourni par l'agrégation des résultats obtenus par plusieurs méthodes myopes de sélection de variables.

1 Introduction

La taille des bases de données étant de plus en plus importante, l'amélioration de la qualité de représentation des données est devenue un problème majeur de l'extraction des connaissances à partir des données. L'une des difficultés principales liée à la représentation des données est la dimension des données.

Le problème de la dimension des données concerne le nombre de variables descriptives caractérisant chacun des individus. Parmi ces variables, certaines peuvent être non pertinentes, inutiles et/ou redondantes. Donc, si l'on désire extraire de l'information utile et compréhensible à partir de nos données, il convient en premier lieu de retirer les parties non pertinentes.

La sélection de variables permet de résoudre ce problème. C'est un processus choisissant un sous-ensemble optimal de variables selon un critère particulier. Il permet l'élimination de variables inutiles, non pertinentes et redondantes ainsi que l'élimination du bruit généré par certaines variables. Le processus d'apprentissage est accéléré et la précision prédictive des algorithmes d'induction peut être améliorée.
Il existe deux familles d'algorithmes de sélection de variables : les méthodes "enveloppe" [John *et al.*, 1994] et les méthodes "filtre" [Kira et Rendell, 1992a]. La différence fondamentale entre ces deux familles réside dans le fait que la première est liée à l'algorithme d'induction utilisée alors que la seconde est totalement indépendante.

1.1 Approches Enveloppe

Les méthodes de type enveloppe prennent en compte l'influence du sous-ensemble de variables sélectionné sur les performances de l'algorithme d'induction. Elles utilisent l'algorithme d'apprentissage comme fonction d'évaluation pour tester les différents sous-ensembles de variables générés. Cependant, leur coût calculatoire est bien souvent trop important.