Sélection rapide en apprentissage supervisé

Pierre-Emmanuel JOUVE *, Gaëlle LEGRAND*, Nicolas NICOLOYANNIS *

*LABORATOIRE ERIC, Université Lumière - Lyon2
Bâtiment L, 5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
{pierre.jouve, gaelle.legrand}@eric.univ-lyon2.fr, nicolas.nicoloyannis@univ-lyon2.fr
http://eric.univ-lyon2.fr

Résumé. La sélection de variables (SdV) permet de réduire l'espace de représentation des données. Ce processus est de plus en plus critique en raison de l'augmentation de la taille des bases de données. Traditionnellement, les méthodes de SdV nécessitent plusieurs accès au jeu de données, ce qui peut représenter une part relativement importante du temps d'execution de ces algorithmes. Nous proposons une nouvelle méthode efficiente et rapide (ne nécessitant qu'un unique accès aux données). Cette méthode utilise les algorithmes génétiques ainsi que des mesures de validité de classification non supervisée (cns).

1 Introduction

La taille des bases de données étant de plus en plus importante, l'amélioration de la qualité de l'espace de représentation des données (ERD) est devenue un problème majeur de l'ECD. L'une des difficultés majeures liée à l'ERD est la dimension des données (le nombre de variables descriptives caractérisant chacun des objets). Ce problème peut se résumer par la phrase de Liu et Motoda [Liu et Motoda, 1998] "Less is more." qui signifie que 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 (SdV) 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 et/ou redondantes, autorisant ainsi l'accélération et l'amélioration de la précision prédictive des processus d'apprentissage. Il existe deux familles d'algorithmes de SdV : les méthodes "Enveloppe" [Kohavi et John, 1997] et les méthodes "Filtre"[Kira et Rendell, 1992]. La différence fondamentale entre ces deux familles réside dans le fait que la première est liée à l'algorithme d'induction utilisé (ce qui lui confère un coût calculatoire bien souvent trop important) alors que la seconde est totalement indépendante.

Les approches filtre sont de 4 types : exhaustive, heuristique, probabiliste et sélection en un seul parcours de base. **Les méthodes exhaustives**(MDLM [Sheinvald *et al.*, 1990], FOCUS [Almuallim et Dietterich, 1991]...) testent tous les sous-ensembles possibles de variables, ces algorithmes sont donc le plus souvent impossible à appliquer du fait de leur coût calculatoire trop élevé. **Les méthodes heuristiques** sont très nombreuses, la plus connue est RELIEF[Kira et Rendell, 1992]. Sa complexité est linéaire selon le nombre d'objets et le nombre d'itérations effectuées. Il existe également des méthodes