Une méthode optimale d'évaluation bivariée pour la classification supervisée

Marc Boullé

France Télécom R&D, 2, avenue Pierre Marzin, 23300 Lannion
marc.boulle@orange-ft.com

Résumé. En préparation des données pour la classification supervisée, les méthodes filtres usuellement utilisées pour la sélection de variables sont efficaces en temps de calcul. Néanmoins, leur nature univariée ne permet pas de détecter les redondances ou les interactions constructives entre variables. Cet article présente une nouvelle méthode permettant d'évaluer l'importance prédictive jointe d'une paire de variables de façon automatique, rapide et fiable. Elle est basée sur un partitionnement de chaque variable exogène, en intervalles dans le cas numérique et groupes de valeurs dans le cas catégoriel. La grille de données exogène résultante permet alors d'évaluer la corrélation entre la paire de variables exogènes et la variable endogène. Le meilleur partitionnement bivarié est recherché au moyen d'une approche Bayesienne de la sélection de modèle. Les expérimentations démontrent les apports de la méthode, notamment une amélioration significative des performances en classification.

1 Introduction

Dans un projet de fouille de données, la phase de préparation des données vise à extraire une table de données pour la phase de modélisation (Pyle, 1999; Chapman et al, 2000). La préparation des données est non seulement coûteuse en temps d'étude, mais également critique pour la qualité des résultats escomptés. La préparation repose essentiellement sur la recherche d'une représentation pertinente pour le problème à modéliser, recherche qui se base sur une sélection de variables. L'objectif de la sélection de variable est triple: améliorer la performance prédictive des classifieurs, le temps d'apprentissage et de déploiement des modèles, et leur interprétabilité (Guyon et Elisseeff, 2003). Deux approches principales, filtre et enveloppe (Kohavi et John, 1997), ont été proposées dans la littérature. Les méthodes filtres évaluent la corrélation entre les variables exogènes et la variable endogène, indépendamment de la méthode de classification utilisée. Les méthodes enveloppes recherchent pour un modèle donné le meilleur sous-ensemble de variables. Les méthodes enveloppes, très coûteuses en temps de calcul, sont plutôt adaptées à la phase de modélisation. Parmi les méthodes filtres, les méthodes procédant par analyse univariée permettent d'ordonner les variables exogènes par importance prédictive décroissante. Elles sont classiquement utilisées en phase de préparation des données pour rapidement extraire un sous-ensemble de variables pertinent pour la modélisation à partir d'un ensemble de variables candidates potentiellement de grande taille. Dans cet article, nous nous focalisons sur l'approche filtre.