Classification hiérarchique de variables discrètes fondée sur l'information mutuelle en pré-traitement d'un algorithme de sélection de variables pertinentes

Hélène Daviet*,**, Ivan Kojadinovic* et Pascale Kuntz*

*LINA CNRS FRE 2729, Site Polytech Nantes, rue Christian Pauc, 44306 Nantes, France
{prenom.nom}@polytech.univ-nantes.fr
**PerformanSe SAS, Atlanpole La Fleuriaye, 44470 Carquefou, France
{prenom.nom}@performanse.fr

Résumé. Le travail présenté a pour contexte la sélection de variables pertinentes dans les problèmes de discrimination caractérisés par un grand nombre de variables potentiellement discriminantes toutes discrètes ou nominales. Dans ce cadre, nous proposons une procédure de sélection fondée sur une troncature k-additive de l'information mutuelle et utilisant une classification ascendante hiérarchique des variables potentiellement discriminantes afin de réduire le nombre de sous-ensembles dont la pertinence est estimée.

1 Introduction

Le problème de la sélection de variables en discrimination se rencontre généralement lorsque le nombre de variables, pouvant être utilisées pour expliquer la classe d'un individu, est très élevé. Le rôle de la procédure de sélection de variables consiste alors à sélectionner un sous-ensemble de variables *potentiellement discriminantes* permettant d'expliquer la classe de façon optimale ou quasi-optimale. La nécessité de ce traitement préalable est essentiellement due au fait que, généralement, l'utilisation d'un nombre de variables discriminantes trop élevé dans un modèle de discrimination détériore grandement sa capacité de *généralisation* et la compréhension de la relation modélisée.

D'un point de vue structurel, une procédure de sélection de variables peut être vue comme composée de deux éléments fondamentaux (Liu et Motoda, 1998) : une *mesure de pertinence*, utilisée pour mesurer l'*influence* d'un sous-ensemble de variables potentiellement discriminantes sur la variable qualitative à expliquer, et un *algorithme de recherche*, dont le rôle est de *parcourir* l'ensemble des sous-ensembles de variables à la recherche d'un sous-ensemble optimal ou quasi-optimal au sens de la mesure de pertinence. Du point de vue de la définition de la mesure de pertinence, les procédures de sélection de variables peuvent être essentiellement regroupées en deux classes (Liu et Motoda, 1998) : les procédures *filtres* et les procédures *modèle-dépendantes*. Dans le cas des procédures filtres, la sélection de variables est totalement indépendante du modèle de discrimination choisi et s'effectue en tant que traitement préalable à la phase d'estimation. Parmi les procédures filtres, citons le travail de Fleuret (2004), proche du notre. En revanche, dans le cas des procédures modèle-dépendantes, la mesure de pertinence