Caractérisation automatique des classes découvertes en classification non supervisée

Nistor Grozavu*, Younès Bennani*
Mustapha Lebbah*

*LIPN UMR CNRS 7030, Université Paris 13,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr

Résumé. Dans cet article, nous proposons une nouvelle approche de classification et de pondération des variables durant un processus d'apprentissage non supervisé. Cette approche est basée sur le modèle des cartes auto-organisatrices. L'apprentissage de ces cartes topologiques est combiné à un mécanisme d'estimation de pertinences des différentes variables sous forme de poids d'influence sur la qualité de la classification. Nous proposons deux types de pondérations adaptatives : une pondération des observations et une pondération des distances entre observations. L'apprentissage simultané des pondérations et des prototypes utilisés pour la partition des observations permet d'obtenir une classification optimisée des données. Un test statistique est ensuite utilisé sur ces pondérations pour élaguer les variables non pertinentes. Ce processus de sélection de variables permet enfin, grâce à la localité des pondérations, d'exhiber un sous ensemble de variables propre à chaque groupe (cluster) offrant ainsi sa caractérisation. L'approche proposée a été validé sur plusieurs bases de données et les résultats expérimentaux ont montré des performances très prometteuses.

1 Introduction

La classification automatique - clustering - est une étape importante du processus d'extraction de connaissances à partir de données. Elle vise à découvrir la structure intrinsèque d'un ensemble d'objets en formant des regroupements - clusters - qui partagent des caractéristiques similaires (Fisher, 1996; Cheeseman et al., 1988). La complexité de cette tâche s'est fortement accrue ces deux dernières décennies lorsque les masses de données disponibles ont vu leur volume exploser. En effet, le nombre d'objets présents dans les bases de données a fortement augmenté mais également la taille de leur description. L'augmentation de la dimension des données a des conséquences non négligeables sur les traitements classiquement mis en oeuvre : outre l'augmentation naturelle des temps de traitements, les approches classiques s'avèrent parfois inadaptées en présence de bruit ou de redondance.

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'observations. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est