Une nouvelle approche pour la sélection de variables basée sur une métrique d'estimation de la qualité

Jean-Charles Lamirel*, Pascal Cuxac**, Kafil Hajlaoui**

*SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, France.
jean-charles.lamirel@loria.fr,
http://www.loria.fr
**CNRS-Inist, Vandoeuvre-lès-Nancy, France.
pascal.cuxac@inist.fr
http://www.inist.fr

Résumé. La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée (clustering) qui favorise les clusters ayant une valeur maximale de F-mesure d'étiquetage. Dans cet article, nous montrons qu'une adaptation de cette métrique dans le cadre de la classification supervisée permet de réaliser une sélection de variables et de calculer pour chacune d'elles une fonction de contraste. La méthode est expérimentée sur différents types de données textuelles. Dans ce contexte, nous montrons que cette technique améliore les performances des méthodes de classification de façon très significative par rapport à l'état de l'art des techniques de sélection de variables, notamment dans le cas de la classification de données textuelles déséquilibrées, fortement multidimensionnelles et bruitées.

1 Introduction

Depuis les années 1990, les progrès de l'informatique et des capacités de stockage permettent la manipulation de très gros volumes de données : il n'est pas rare d'avoir des espaces de description de plusieurs milliers, voire de dizaines de milliers de variables. On pourrait penser que les algorithmes de classification sont plus efficaces avec un grand nombre de variables, mais la situation n'est pas aussi simple que cela. Le premier problème qui se pose est l'augmentation du temps de calcul. En outre, le fait qu'un nombre important de variables soit redondant ou non pertinent pour la tâche de classification perturbe considérablement le fonctionnement des classifieurs. De plus, la plupart des algorithmes d'apprentissage exploitent des probabilités dont les distributions peuvent être difficiles à estimer en présence d'un très grand nombre de variables. L'intégration d'un processus de sélection de variables dans le cadre de la classification des données de grande dimension devient donc un enjeu central. Dans la littérature, trois types d'approches pour la sélection de variables sont principalement proposés : les approches directement intégrées aux méthodes de classification, dites «embedded», les méthodes basées sur des techniques d'optimisation, dites «wrapper», et finalement, les approches de filtrage. Des états de l'art exhaustifs ont été réalisés par de nombreux auteurs, comme Ladha