Un critère d'évaluation pour la sélection de variables

Dahbia Semani, Carl Frélicot, Pierre Courtellemont

Laboratoire d'Informatique – Image – Interaction
Université de La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle Cedex, France
{dahbia.semani,carl.frelicot,pierre.courtellemont}@univ-lr.fr

Résumé. Cet article aborde le problème de la sélection de variables dans le cadre de la classification supervisée. Les méthodes de sélection reposent sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables. Nous présentons un nouveau critère d'évaluation fondé sur une mesure d'ambiguïté. Cette mesure est fondée sur une combinaison d'étiquettes représentant le degré de spécificité ou d'appartenance aux classes en présence. Les tests menés sur de nombreux jeux de données réels et artificiels montrent que notre méthode est capable de sélectionner les variables pertinentes et d'augmenter dans la plupart des cas les taux de bon classement.

1 Introduction

En reconnaissance des formes, les données sont des vecteurs réalisations de variables qui correspondent à des mesures réalisées sur un système physique ou à des informations collectées lors d'une observation d'un phénomène. Ces variables ne sont pas toutes aussi informatives : elles peuvent correspondre à du bruit, être peu significatives, corrélées ou non pertinentes pour la tâche à réaliser. La sélection de variables a pour objectif de réduire le nombre de ces variables et donc réduire la taille des informations à traiter. Des traitements plus sophistiqués peuvent alors être utilisés dans des espaces de dimension réduite, l'étape d'apprentissage est facilitée, les performances peuvent augmenter lorsque les variables non pertinentes ou redondantes disparaissent, etc.

Nous traitons, dans cet article, le problème de la sélection de variables dans le cadre de la reconnaissance de formes statistique et plus particulièrement dans le cadre de la classification supervisée (ou classement). Dans ce cas, la sélection de variables a pour objectif de réduire la complexité en sélectionnant le sous-ensemble de variables de taille minimale sans que les performances de la règle de classement diminuent trop voire même augmentent.

Une méthode de sélection repose sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables. Nous nous intéressons aux critères d'évaluation. Ainsi, nous proposons un nouveau critère d'évaluation fondé sur une mesure d'ambiguïté. Cette mesure repose sur la combinaison d'étiquettes représentant le degré de spécificité ou d'appartenance aux classes en présence. Des opérateurs d'agrégation issus de la logique floue sont utilisés pour la combinaison de ces étiquettes.

Cet article est organisé comme suit. Un bref état de l'art sur les algorithmes de sélection de variables et les critères d'évaluation est dressé aux sections 2 et 3. Nous