Sélection de variables non supervisée sous contraintes hiérarchiques

Nhat-Quang Doan, Hanane Azzag, Mustapha Lebbah

Université Paris 13, Sorbonne Paris Cité,
Laboratoire d'Informatique de Paris-Nord (LIPN),
CNRS(UMR 7030),
99, avenue Jean-Baptiste Clément
93430 Villetaneuse, France
prénom.nom@lipn.univ-paris13.fr

Résumé. La sélection des variables a un rôle très important dans la fouille de données lorsqu'un grand nombre de variables est disponible. Ainsi, certaines variables peuvent être peu significatives, corrélées ou non pertinentes. Une méthode de sélection a pour objectif de mesurer la pertinence d'un ensemble utilisant principalement un critère d'évaluation. Nous présentons dans cet article un critère non supervisé permettant de mesurer la pertinence d'un sous-ensemble de variables. Ce dernier repose sur l'utilisation du score Laplacien auquel nous avons ajouté des contraintes hiérarchiques. Travailler dans le cadre non supervisé est un vrai challenge dans ce domaine dû à l'absence des étiquettes de classes. Les résultats obtenus sur plusieurs bases de tests sont très encourageants et prometteurs.

1 Introduction

La sélection de variables joue un rôle très important en classification lorsqu'un grand nombre de variables sont disponibles. Ainsi, certaines variables peuvent être peu significatives, corrélées ou non pertinentes. La sélection de variables permet également d'accélérer l'étape d'apprentissage et de réduire la complexité des algorithmes. Une méthode de sélection repose principalement sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables.

En apprentissage supervisé, la sélection de variables a largement été étudiée car il est connu que la sélection de variables peut améliorer la qualité d'un classificateur (Zhang et al., 2009). Parmi les méthodes supervisées, nous citons le coefficient de corrélation de Pearson (Rodgers et Nicewander, 1988), le score de Fisher (Duda et al., 2000) et le gain de l'information (Cover et Thomas, 2006). La sélection de variables a reçu peu d'attention en apprentissage non supervisé en comparaison au cas supervisé. Le problème devient plus difficile en raison de l'absence des étiquettes des classes pour guider la sélection. Ainsi se pose la question importante, comment évaluer la pertinence d'un sous-ensemble de fonctionnalités sans avoir recours aux étiquettes de classe ? Dans la littérature deux approches sont souvent utilisées pour évaluer la pertinence d'un sous-ensemble de variables sélectionnées, (Kohavi et John, 1997; Yu et