Classification supervisée pour de grands nombres de classes à prédire : une approche par co-partitionnement des variables explicatives et à expliquer

Marc Boullé *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Dans la phase de préparation des données du data mining, les méthodes de discrétisation et de groupement de valeurs supervisé possèdent de nombreuses applications : interprétation, estimation de densité conditionnelle, sélection de type filtre des variables, recodage des variables en amont des classifieurs. Ces méthodes supposent habituellement un faible nombre de valeur à expliquer (classes), typiquement moins d'une dizaine, et trouvent leur limite quand leur nombre augmente. Dans cet article, nous introduisons une extension des méthodes de discrétisation et groupement de valeurs, consistant à partitionner d'une part la variable explicative, d'autre part la variable à expliquer. Le meilleur co-partitionnement est recherché au moyen d'une approche Bayesienne de la sélection de modèle. Nous présentons ensuite comment utiliser cette méthode de prétraitement en préparation pour le classifieur Bayesien naïf. Des expérimentations intensives démontrent l'apport de la méthode dans le cas de centaines de classes.

1 Introduction

L'objectif de la classification supervisée est de prédire la valeur d'une variable catégorielle à expliquer connaissant l'ensemble des valeurs des variables explicatives, numériques ou catégorielles. La plupart des problèmes de classification considérés usuellement se limitent à la prédiction d'une valeur booléenne, ou d'une variable comportant un nombre très faible de valeurs, typiquement moins d'une dizaine. On rencontre néanmoins des problèmes où ce nombre de valeurs à expliquer est plus important, comme par exemple la reconnaissance de chiffres manuscrits, la reconnaissance de caractères ou la classification de textes. Les applications émergentes de ciblage publicitaire sur internet sont amenées à considérer le cas du choix d'un bandeau publicitaire parmi plusieurs centaines pour maximiser le taux de clic lors de la navigation des internautes. Les méthodes existantes supposent au moins implicitement un faible nombre de classes, et sont potentiellement moins performantes dans le cas de grands nombres de classes, avec peu d'individus par classe Il s'agit ici d'envisager le problème de classification dans son cadre le plus général sans faire l'hypothèse d'un nombre restreint de