Co-classification sous contraintes par la somme des résidus quadratiques

Ruggero G. Pensa*, Jean-François Boulicaut**

*KDD-Lab, ISTI-CNR - Via Giuseppe Moruzzi, 1 - I-56124 Pisa, Italy
ruggero.pensa@isti.cnr.it
**INSA-Lyon, LIRIS CNRS UMR5205, F-69621 Villeurbanne cedex, France
jean-francois.boulicaut@insa-lyon.fr

Résumé. Dans de nombreuses applications, une co-classification est plus facile à interpréter qu'une classification mono-dimensionnelle. Il s'agit de calculer une bi-partition ou collection de co-clusters : chaque co-cluster est un groupe d'objets associé à un groupe d'attributs et les interprétations peuvent s'appuyer naturellement sur ces associations. Pour exploiter la connaissance du domaine et ainsi améliorer la pertinence des partitions, plusieurs méthodes de classification sous contraintes ont été proposées pour le cas mono-dimensionnel, e.g., l'exploitation de contraintes "must-link" et "cannot-link". Nous considérons ici la co-classification sous contraintes avec la gestion de telles contraintes étendues aux dimensions des objets et des attributs, mais aussi l'expression de contraintes de contiguité dans le cas de domaines ordonnés. Nous proposons un algorithme itératif qui minimise la somme des résidus quadratiques et permet l'exploitation active des contraintes spécifiées par les analystes. Nous montrons la valeur ajoutée de ce type d'extraction sur deux applications en analyse du transcriptome.

1 Introduction

Dans de nombreux domaines applicatifs, l'analyste se trouve devant des jeux de données matriciels dans lesquels un certain nombre d'objets sont décrits par un certain nombre d'attributs qui prennent leurs valeurs dans un domaine numérique, éventuellement restreint au domaine 0/1. L'une des techniques phares pour l'étude exploratoire de tels jeux de données est la classification, i.e., le calcul de partitions, soit sur l'ensemble des objets, soit sur l'ensemble des attributs. On peut aussi vouloir faciliter l'interprétation des groupements calculés en développant des méthodes de co-classification. Dans ce cas, les partionnements selon les deux dimensions sont couplés et les algorithmes comme ceux présentés dans Robardet et Feschet (2001); Dhillon et al. (2003); Ritschard et Zighed (2003); Jollois et al. (2003) produisent une bi-partition, i.e., une collection de co-clusters. Chacun des co-clusters est un groupe d'objets associé à un groupe d'attributs et la co-classification apparaît comme une méthode de classification conceptuelle. La co-classification a été particulièrement étudiée dans le contexte de l'analyse du transcriptome (voir, e.g., Cheng et Church (2000); Madeira et Oliveira (2004)). En effet, les technologies à haut débit permettent de construire des matrices d'expression de (tous