Un algorithme de classification automatique pour des données relationnelles multi-vues

Francisco de A. T. de Carvalho*, Filipe M. de Melo*, Yves Lechevallier**, Thierry Despeyroux**

*Centro de Informatica -CIn/UFPE - Av. Prof. Luiz Freire, s/n -Cidade Universitaria - CEP 50740-540, Recife-PE, Brésil
{fatc,fmm}@cin.ufpe.br
**INRIA, Paris-Rocquencourt - 78153 Le Chesnay cedex, France
{Yves.Lechevallier,Thierry.Despeyroux}@inria.fr

Résumé. classification automatique (De Carvalho et al., 2012) capable de partitionner des objets en prenant en compte de manière simultanée plusieurs matrices de dissimilarité qui les décrivent. Ces matrices peuvent avoir été générées en utilisant différents ensembles de variables et de fonctions de dissimilarité. Cette méthode, basée sur l'algorithme de nuées dynamiques est conçu pour fournir une partition et un prototype pour chaque classe tout en découvrant une pondération pertinante pour chaque matrice de dissimilarité en optimisant un critère d'adéquation entre les classes et leurs représentants. Ces pondérations changent à chaque itération de l'algorithme et sont différentes pour chacune des classes. Nous présentons aussi plusieurs outils d'aide à l'interprétation des groupes et de la partition fournie par cette nouvelle méthode. Deux exemples illustrent l'interêt de la méthode. Le premier utilise des données concernant des chiffres manuscrits (0 à 9) numérisés en images binaires provenant de l'UCI. Le second utilise un ensemble de rapports dont nous connaissons une classification experte donnée à priori.

1 Introduction

La classification est une activité courante en extraction de connaissances. Le but de la classification est d'organiser un ensemble d'objets en sous-ensembles appelés clusters ou classes de telle façon que les objets d'une même classe se ressemblent. Pour quantifier cette ressemblance il est important de bien définir cette dissimilarité entre deux objets à partir de leurs représentations.

Dans le cas où la représentation d'un objet n'est pas unique, ces données sont appelées multi-vues. Elles sont présentes dans plusieurs domaines tels que la bioinformatique, le marketing, etc. (Cleuziou et al., 2009). Dans les documents structurés, par exemple les documents XML, il existe plusieurs blocs ou sections, chacun peut être interprété comme une vue.

L'idée générale est que chaque matrice de dissimilarité ait un rôle collaboratif (Pedrycz, 2002) dans le but d'arriver à un consensus sur une partition (Leclerc et Cucumel, 1987). Ces