

# 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 (?) 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 pertinente 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'intérêt 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.

## Summary

This paper introduces an improvement of a clustering algorithm (?) that is able to partition objects taking into account simultaneously their relational descriptions given by multiple dissimilarity matrices. These matrices could have been generated using different sets of variables and dissimilarity functions. This method, which is based on the dynamic hard clustering algorithm for relational data, is designed to provided a partition and a prototype for each cluster as well as to learn a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that measures the fit between clusters and their representatives. These relevance weights change at each algorithm iteration and are different from one cluster to another. Moreover, various tools for the partition and cluster interpretation furnished by this new algorithm are also presented. Two experiments demonstrate the usefulness of this clustering method and the merit of the partition and cluster interpretation tools. The first one uses a data set from UCI machine

Algorithme de classification automatique pour données relationnelles multi-vues

learning repository concerning handwritten numbers (digitalized pictures). The second uses a set of reports for which we have an expert classification given a priori.