PoBOC : un algorithme de "soft-clustering". Applications à l'apprentissage de règles et au traitement de données textuelles.

Guillaume Cleuziou, Lionel Martin, Christel Vrain
LIFO, Laboratoire d'Informatique Fondamentale d'Orléans
Rue Léonard de Vinci B.P. 6759
45067 Orléans cedex 2 - FRANCE
{cleuziou,martin,cv}@lifo.univ-orleans.fr
http ://www.univ-orleans.fr/SCIENCES/LIFO/

Résumé. Nous décrivons l'algorithme PoBOC (Pole-Based Overlapping Clustering) qui génère un ensemble de clusters non-disjoints (ou "soft-clusters") présentés sous forme d'une hiérarchie de concepts à partir de la seule matrice de similarités sur les données considérées. Nous évaluons l'approche sur deux situations d'apprentissage : la classification par apprentissage de règles et l'organisation de données plus complexes et peu structurées telles que les données textuelles.
La validation des méthodes de clustering est une étape difficile résolue le plus souvent par une évaluation d'experts. Les deux applications proposées permettent de valider la méthode d'organisation selon deux points de vue : d'une part quantitativement en évaluant l'influence de la méthode pour la classification, d'autre part en permettant une analyse "humaine" du résultat dans le cas des données textuelles. Nous mettons en évidence l'intérêt de PoBOC comparativement à d'autres approches d'apprentissage non-supervisé.

1 Introduction

Le clustering consiste à organiser les données de manière à regrouper les objets les plus similaires et à séparer ceux qui se ressemblent le moins. De nombreux algorithmes ont été proposés dans divers domaines d'application : la reconnaissance de formes [Jain *et al.*, 2000], la classification par apprentissage non-supervisé [Agrawal *et al.*, 1992], la segmentation d'images [Pham et Prince, 1998] ou encore le regroupement de données textuelles [Baker et McCallum, 1998]. Dans cette étude nous distinguons les algorithmes de regroupement "dur" (ou *hard-clustering*) qui renvoient un ensemble de groupes disjoints, des méthodes de regroupement "flou" (ou *fuzzy-clustering*) pour lesquelles sont renvoyés un ensemble de foyers (points dans l'espace ou objets réels) ainsi qu'une matrice d'appartenance floue. Pour ces deux types d'approches, on parle d'algorithmes de regroupement "objet" lorsque les données sont décrites par des attributs, par opposition au regroupement "relationnel" basé sur une matrice de similarités. De nombreux algorithmes ont été proposés dont certainement le plus connu est c-means (algorithme de partitionnement objet dur) et ses variantes : c-medoids (variante relationnelle), fuzzy-c-means (variante floue) et fuzzy-c-medoids (variante relationnelle floue) [MacQueen, 1967]. Ces algorithmes, avec les méthodes agglomératives hiérarchiques