Classification non-supervisée de données relationnelles

Jérôme Maloberti[*,**], Shin Ando[**]
Einoshin Suzuki[**]

[*]Université Paris-Sud, Laboratoire de Recherche en Informatique (LRI), Bât 490,
F-91405 Orsay Cedex, France
[**]Electrical and Computer Engineering, Yokohama National University,
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan

1 Introduction

La classification, ou *clustering* (Jain et al., 1999), consiste à associer une classe à chaque élément d'un ensemble, les éléments similaires devant être regroupés dans une classe en n'utilisant que la similarité (ou distance) entre deux éléments ou groupes d'éléments. Le formalisme attributs-valeurs ne permettant pas de représenter les domaines complexes, l'apprentissage en logique du premier ordre, ou Programmation Logique Inductive (PLI), a attiré une attention croissante. Le language utilisé en PLI, DATALOG, est un formalisme relationnel ne permettant pas les fonctions, et dont le test de couverture, la θ-subsomption, est une restriction décidable mais NP-difficile de l'implication logique. Cet article présente une méthode permettant l'utilisation d'algorithmes de clustering sur des données relationnelles, en recherchant préliminairement tous les motifs relationnels existant et en les utilisant pour définir une distance entre des clauses en DATALOG.

2 Présentation de l'algorithme

L'algorithme proposé consiste en trois étapes : la recherche de tous les motifs relationnels de la base, l'élimination des motifs inintéressants et le clustering des clauses DATALOG, en utilisant les motifs pour calculer la distance entre les exemples. La recherche des motifs relationnels est effectuée par JIMI (Maloberti et Suzuki (2003)) qui est une version relationnelle d'un algorithme de recherche en largeur d'itemset fréquents. Chaque exemple est tranformé en un vecteur booléen dont les valeurs correspondent au test de θ-subsomption[1] des motifs contre cet exemple, ces vecteurs permettant d'utiliser les distances existantes. Différents paramètres peuvent être utilisés : différents poids sur les motifs durant le calcul de la distance, tels que la taille des motifs ou l'inverse de la fréquence, utilisation des n premiers niveaux trouvés par JIMI plutôt que tous les niveaux, utilisation d'une partie des motifs (tous les motifs maximaux, i.e. fermés, ou les motifs minimaux).

Notre méthode a été testée sur 2 ensembles de données réelles avec un algorithme de clustering hiérarchique ascendant et une distance euclidienne. Le premier test concerne la détection

[1]La version utilisée vérifie l'Identité d'Objet, toutes les variables sont substituées par des termes différents.