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.

d'accès hostiles sur le site web "www.slab.dnj.ynu.ac.jp". Les données, dont des résultats ont déjà été publiés dans Narahashi et Suzuki (2003) et Hirose et Suzuki (2005), correspondant à deux ans d'accès et contiennent : $205,590$ requêtes, $32,425$ sessions [2], dont $2,243$ hostiles. Notre méthode a obtenu (sur 10.000 sessions) une précision de 0.991 avec 12 clusters, Narahashi et Suzuki (2003) obtenant 0.981, avec 5 clusters et Hirose et Suzuki (2005) 0.719 avec 2 clusters. Ce problème n'étant pas relationnel, les 2 premiers niveaux ont les meilleurs résultats, l'utilisation de plus de niveaux n'a conduit qu'à la création de plus de clusters. Le second ensemble de données, décrit dans King et al. (1995), concerne la détection de capacité à provoquer des mutations et représente 230 molécules, dont 138 positives et 92 négatives. Les résultats ont été médiocres, une précision de 0.51, car seule la description des atomes et de leurs relations a été utilisée, ce qui est insuffisant pour obtenir des motifs discriminants.

3 Conclusion et perspectives

Nous avons proposé une nouvelle méthode permettant le clustering de données relationnelles et nous avons utilisé ce système sur deux ensembles de données. Les résultats préliminaires montrent que ce système peut égaler les autres algorithmes sur des données non relationnelles, l'expérimentation sur des données relationnelles n'ayant pas permis de conclure. Parmi les perspectives, l'utilisation d'un algorithme de clustering pouvant gérer de grandes dimensions, tel que le subspace clustering, serait intéressante car le grand nombre de motifs rend les distances très instables mathématiquement.

Références

Hirose, N. et E. Suzuki (2005). Detecting hostile accesses to a web site using a visualization method based on probabilistic clustering. In *Proc. 1st WSEAS Intern. Symp. on Datamining*.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

King, R., A. Srinivasan, et M. Stenberg (1995). Relating chemical activity to structure : an examination of ILP successes. *New Generation Computing 13*.

Maloberti, J. et E. Suzuki (2003). Improving efficiency of frequent query discovery by eliminating non-relevant candidates. In *Proc. 6th Inter. Conf. on Discovery Science*.

Narahashi, M. et E. Suzuki (2003). Detecting hostile accesses through incremental subspace clustering. In *IEEE/WIC International Conference on Web Intelligence*, pp. 337–343.

Summary

This paper presents an algorithm for clustering of relational data in DATALOG formalism which searches all relational patterns in the base, then transforms each example in a boolean vector corresponding to the results of its covering tests against the patterns.

[2] Une session est une séquence de requêtes d'un même ordinateur avec délai entre deux requêtes successives inférieur à une heure