Suppression des Itemsets Clés Non Essentiels en Classification basée sur les Règles d'Association

Viet Phan-Luong

Université de Provence
Laboratoire d'Informatique Fondamentale de Marseille
(LIF - UMR CNRS 6166)
CMI, 39 rue F. Joliot Curie
13453 Marseille, France
viet.phanluong@lif.univ-mrs.fr

Résumé. En classification basée sur les règles d'association, les itemsets clés sont essentiels : la suppression des itemsets non clés n'affecte pas la précision du classifieur en construction. Ce travail montre que parmi ces itemsets clés, on peut s'intéresser seulement à ceux de petites tailles. Plus loin encore, il étudie une généralisation d'une propriété importante des itemsets non clés et montre que parmi les itemsets clés de petites tailles, il y a ceux qui ne sont pas significatifs pour la classification. Ces itemsets clés sont dits non essentiels. Ils sont définis via un test de χ^2. Les expériences menées sur les grands jeux de données montrent que l'optimisation par la suppression de ces itemsets est correcte et efficace.

1 Introduction

Etant donné un ensemble d'objets et un ensemble d'étiquettes de classes, le problème de classification est de chercher une fonction pour attribuer à chaque objet une étiquette de classe. Une telle fonction est appelée un classifieur. Les constructions de ces classifieurs sont en général basées sur les données d'exemples (d'entraînement). Il existe plusieurs méthodes de classification, telles que l'arbre de décision Quinlan (1993), la méthode naïve-Bayes Duda et Hart (1973), les méthodes basées sur les règles Clark et Niblett (1995); Cohen (1995). Ce papier présente une approche à la construction de classifieurs basée sur les règles classe-associations Lent et al. (1997); Liu et al. (1998); Li et al. (2001), en utilisant une structure d'arbre de préfixes pour l'extraction des itemsets fréquents et les règles d'association Agrawal et al. (1993).

Dans les approches telles que CMAR Li et al. (2001), HARMONY Wang et Karypis (2005), par optimisations, les règles d'association sont essentiellement construites sur les itemsets clés Bastide et al. (2000). Ce présent travail montre que parmi ces itemsets clés, on peut s'intéresser seulement à ceux de petites tailles. Ensuite, via un test de χ^2, il montre que parmi ces derniers, il existe encore ceux qui ne sont pas significatifs pour la classification. Ces itemsets clés sont dits non essentiels. Les résultats d'expérimentations sur les grands jeux de données de *UCI* Coenen (2004) montrent que l'optimisation par la suppression de ces itemsets est correcte et efficace.