Un algorithme de génération des itemsets fermés pour la fouille de données

Huaiguo Fu, Engelbert Mephu Nguifo

CRIL-CNRS FRE2499, Université d'Artois-IUT de Lens
Rue de l'université SP 16, 62307 Lens cedex. France
{fu,mephu}@cril.univ-artois.fr

Résumé. Le traitement de grand volume de données est un problème pour l'extraction de connaissances. La fouille de données nécessite des méthodes de résolution efficaces. Le treillis de concepts (treillis de Galois) est un outil utile pour l'analyse de données. Des travaux en classification et sur les règles d'association ont permis d'accroître son intérêt. Plusieurs algorithmes de génération ont été proposés, parmi lesquels NextClosure est l'un des meilleurs pour traiter des données de grande taille. Mais la complexité de NextClosure reste malgré tout très élévé. Aussi nous proposons un nouvel algorithme efficace nommé **ScalingNextClosure**, et basé sur une méthode de partitionnement de données pour générer de manière indépendante les itemsets fermés de chaque partition. Les résultats expérimentaux montrent que cette technique de partitionnement améliore efficacement NextClosure.

1 Introduction

Dans le domaine des bases de données et de l'analyse de données, les données de grande taille reste difficile à analyser. Par exemple, dans le cas de la fouille de règles d'association [Agrawal *et al.*, 1996], trouver tous les itemsets fréquents est un problème de complexité exponentielle par rapport au nombre d'items. La génération des itemsets fréquents est l'étape la plus importante dans le processus de recherche des règles d'association. Il est donc nécessaire de développer des algorithmes efficaces pour traiter cette étape. La structure de treillis de concepts est un outil intéressant pour la génération des itemsets fermés fréquents qui est l'étape la plus importante dans la recherche de règles d'association. Ses fondements théoriques repose sur la théorie mathématique des treillis [Birkhoff, 1967]. Elle est constituée de concepts formels générés à partir d'un contexte de données pour montrer les relations entre les objets et les attributs de ce contexte. Le treillis des itemsets fermés est inclus dans le treillis des concepts. Le problème de génération des itemsets fréquents peut être réduit à la génération des itemsets fermés fréquents avec le treillis des itemsets fermés. Et il est possible d'élaguer le nombre de règles générées sans perte d'information à partir du treillis des itemsets fermés. Un itemset fermé est l'intension d'un concept formel, et est un itemset maximal pour la recherche des règles d'association [Pasquier *et al.*, 1999].

Plusieurs algorithmes ont été proposés pour générer les concepts formels et/ou le treillis de concepts formels. Les comparaisons expérimentales des temps de calcul de ces algorithmes montrent que l'algorithme Nextclosure [Ganter et Wille, 1999] est