Intégration efficace de méthodes de fouille de données dans les SGBD

Cédric Udréa, Fadila Bentayeb
Jérôme Darmont, Omar Boussaid

ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès-France – 69676 Bron Cedex – France
{cudrea | bentayeb | jdarmont | boussaid}@eric.univ-lyon2.fr

Résumé. Cet article présente une nouvelle approche permettant d'appliquer des algorithmes de fouille, en particulier d'apprentissage supervisé, à de grandes bases de données et en des temps de traitement acceptables. Cet objectif est atteint en intégrant ces algorithmes dans un SGBD. Ainsi, nous ne sommes limités que par la taille du disque et plus par celle de la mémoire. Cependant, les entrées-sorties nécessaires pour accéder à la base engendrent des temps de traitement longs. Nous proposons donc dans cet article une méthode originale pour réduire la taille de la base d'apprentissage en construisant sa table de contingence. Les algorithmes d'apprentissage sont alors adaptés pour s'appliquer à la table de contingence. Afin de valider notre approche, nous avons implémenté la méthode de construction d'arbre de décision ID3 et montré que l'utilisation de la table de contingence permet d'obtenir des temps de traitements équivalents à ceux des logiciels classiques.

Mots Clés: Intégration, Bases de données, Fouille de données, Arbres de décision, Vues relationnelles, Table de contingence, Apprentissage supervisé, Performance.

1 Introduction

L'application d'opérateurs de fouille de données sur de grandes bases de données est un enjeu intéressant. Cependant, les algorithmes de fouille de données ne peuvent opérer que sur des structures en mémoire de type tableau attributs-valeurs, ce qui limite la taille des bases à traiter. De ce fait, les méthodes classiques de fouille de données utilisent des méthodes de pré-traitement sur les données, telles que la sélection de variables [Lia et Motoda, 1998] ou l'échantillonnage [Chauchat, 2002].

Afin d'appliquer des algorithmes de fouille de données sur de grandes bases de données, de nouvelles voies de recherche sont apparues ces dernières années. Elles consistent à intégrer des méthodes de fouille dans les Systèmes de Gestion de Bases de Données (SGBD) [Chaudhuri, 1998]. L'une des premières avancées dans le domaine de l'intégration de méthodes d'analyse des données dans les SGBD a été amorcée par l'avènement des entrepôts de données et de l'analyse en ligne (OLAP) en particulier [Codd, 1993]. D'autres travaux de recherche ont concerné l'intégration des méthodes de règles d'association et de leur généralisation [Meo *et al.*, 1996] [Sarawagi *et al.*, 1998]. En revanche, il existe peu de travaux d'intégration dans les SGBD de méthodes classiques de fouille