Intégration efficace des arbres de décision dans les SGBD : utilisation des index bitmap

Cécile Favre, Fadila Bentayeb

Laboratoire ERIC, Lyon 2
5 Avenue Pierre Mendès France
69676 Bron CEDEX

{cfavre,bentayeb}@eric.univ-lyon2.fr,

Résumé. Nous présentons dans cet article une nouvelle approche de fouille qui permet d'appliquer des algorithmes de construction d'arbres de décision en répondant à deux objectifs : (1) traiter des bases volumineuses, (2) en des temps de traitement acceptables. Le premier objectif est atteint en intégrant ces algorithmes au cœur des SGBD, en utilisant uniquement les outils fournis par ces derniers. Toutefois, les temps de traitement demeurent longs, en raison des nombreuses lectures de la base. Nous montrons que, grâce aux index bitmap, nous réduisons à la fois la taille de la base d'apprentissage et les temps de traitements. Pour valider notre approche, nous avons implémenté la méthode ID3 sous forme d'une procédure stockée dans le SGBD Oracle.

Mots clés : Index bitmap, bases de données, fouille de données, arbres de décision, performance, complexité.

1 Introduction

L'application efficace de méthodes de fouille sur des bases de données volumineuses devient un enjeu de recherche de plus en plus important. Les algorithmes traditionnels de fouille de données s'appliquent sur des tableaux attributs/valeurs (Zighed et Rakotomalala 2000). La volumétrie des bases étant croissante, les algorithmes classiques se heurtent au problème de la limitation de la taille de la mémoire centrale dans laquelle les données sont traitées. La "scalabilité" (capacité de maintenir des performances malgré un accroissement du volume de données), peut alors être assurée en optimisant soit les algorithmes (Agrawal et al. 1996, Gehrke et al. 1998), soit l'accès aux données (Ramesh et al. 2001, Dunkel et Soparkar 1999). Une autre issue au problème consiste à réduire la volumétrie des données à traiter. Pour cela, une phase de prétraitement est généralement appliquée sur les données : l'échantillonnage (Ttoivonen 1996, Chauchat et Rakotomalala 2000) ou la sélection d'attributs (Lia et Motoda 1998).

Récemment, une nouvelle approche de fouille de données est apparue pour pallier au problème de limitation de la taille de la mémoire. Il s'agit d'intégrer les méthodes de fouille de données au cœur des Systèmes de Gestion de Bases de Données (SGBD) (Chaudhuri 1998). Ainsi, le volume des données traitées n'est plus limité par la taille de la mémoire. Cette piste de recherche est conjointement liée à l'avènement des entrepôts de données et de l'analyse en ligne (OLAP) plus particulièrement (Codd 1993).