Fouille de données sous contraintes et bases de données inductives

Jean-François Boulicaut

Fouille de données sous contraintes et bases de données inductives

Jean-François Boulicaut

LIRIS, CNRS UMR 5205, INSA de Lyon
Bâtiment Blaise Pascal, 69621 Villeurbanne cedex
jean-francois.boulicaut@insa-lyon.fr

Le cadre des bases de données inductives propose de considérer l'extraction de connaissances comme un processus d'interrogation au sens des bases de données. Les requêtes portent alors sur les données (au sens classique, e.g., pour sélectionner un contexte de fouille), sur des abstractions ou généralisations des données, sur des modèles (e.g., pour sélectionner des règles descriptives intéressantes) ou sur les deux composantes (e.g., pour identifier des objets qui satisfont certaines règles). Cette vision a été introduite par Imielinski and Mannila en 1996. De réels progrès ont été réalisés avec d'une part des travaux sur les langages de requêtes pour la fouille de données et d'autre part les nombreuses recherches sur la fouille de données sous contraintes.

En considérant plusieurs processus d'extraction de connaissances typiques (e.g., la recherche d'ensembles fréquents, de règles d'associations a priori intéressantes, de fragments moléculaires caractéristiques ou encore celle de motifs séquentiels), nous présenterons la grille d'analyse proposée par le consortium Européen cInQ (projet IST-FET 2000-26469, Mai 2001-Mai 2004). Elle est basée sur les notions de langage d'hypothèses (e.g., les ensembles ou les séquences), de fonctions d'évaluation (e.g., la fréquence ou la cloture), de contraintes primitives (e.g., imposer des fréquences minimales ou maximales, imposer une structure), de requêtes inductives qui combinent diverses contraintes primitives (e.g., des conjonctions de contraintes primitives ou le cadre général des combinaisons booléennes).

L'évaluation des requêtes inductives implique le développement de solveurs efficaces et si possible complets pour calculer toutes les hypothèses qui satisfont les contraintes exprimées. Le contexte de la fouille de données (très grands espaces de recherche mais aussi très grands volumes de données) pose des problèmes spécifiques qui sont de mieux en mieux maîtrisés dans le cas de l'extraction sous contraintes de motifs locaux [1]. L'application aux modèles ou motifs globaux (comme des calculs de partitions ou de classifieurs) pose encore de nombreux problèmes ouverts. Nous introduirons les directions de travail actuelles.

Références

[1] J.-F. Boulicaut. Inductive databases and multiple uses of frequent itemsets : the cinq approach. In R. Meo et al., editor, *Database support for Data Mining Applications - Discovering Knowledge with Inductive Queries*, LNCS 2682, Springer-Verlag, pages 3–26, 2004.