Mise en œuvre des méthodes de fouille de données spatiales Alternatives et performances

Nadjim Chelghoum, Karine Zeitouni

Laboratoire PRISM, Université de Versailles
45, avenue des Etats-Unis, 78035 Versailles Cedex, France
Nadjim.Chelghoum@prism.uvsq.fr, Karine.Zeitouni@prism.uvsq.fr

Résumé. La fouille de données spatiales nécessite l'analyse des interactions dans l'espace. Ces interactions peuvent être matérialisées dans des tables de distances, ramenant ainsi la fouille de données spatiales à l'analyse multi-tables. Or, les méthodes de fouilles de données traditionnelles considèrent une seule table en entrée où chaque tuple est une observation à analyser. De simples jointures entre ces tables ne résoud pas le problème et fausse les résultats en raison du comptage multiple des observations. Nous proposons trois alternatives de fouille de données multi-tables dans le cadre de la fouille des données spatiales. La première consiste à interroger à la volée les différentes tables et modifie en dur les algorithmes existants. La seconde est une optimisation de la première qui pré-calcule les jointures et adapte les algorithmes existants. La troisième réorganise les données dans une table unique en complétant - et non en joignant- la table d'analyse par les données présentes dans les autres tables, ensuite applique un algorithme standard sans modification. Cet article présente ces trois alternatives. Il décrit leur implémentation pour la classification supervisée et compare leur performance.

1. Positionnement du problème

La fouille de données spatiales (FDS) est aujourd'hui un domaine bien identifié de la fouille de données. Elle est née du besoin d'exploitation dans un but décisionnel de données à caractère spatial produites, importées ou accumulées, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires (Zeitouni 2000). Sa principale caractéristique est qu'elle considère les relations spatiales – qu'on appellera de voisinage – (Egenhofer et al. 1993). Ces relations sont à l'origine implicites et nécessitent des jointures coûteuses sur des critères spatiaux pour être exhibées. Nous avons proposé dans nos travaux antérieurs de les matérialiser en utilisant une structure secondaire appelée index de jointure spatial (Zeitouni et al. 2000). L'idée est de pré-calculer la relation spatiale exacte entre les localisations de deux collections d'objets spatiaux et de la stocker dans une table de type (objet1, relation-spatiale, objet2). Ceci nous permet de pallier le problème du coût des jointures spatiales au moment de l'analyse. Néanmoins, cette organisation ne peut pas être directement analysée par les méthodes de fouilles de données car celles-ci considèrent que les données en entrée sont dans une table unique et que chaque tuple de cette table constitue une observation ou un individu à analyser. On se trouve alors confronté au problème qu'on ne peut exploiter telles quelles les données organisées en plusieurs tables. Il est possible de se ramener à une seule table en joignant les différentes tables initiales. Or, cette jointure peut dupliquer des tuples car les observations à analyser sont en liaison N-M avec les objets