Sélection d'attributs en fouille de données sur grilles [1]

Sébastien Cahon, Nouredine Melab et El-Ghazali Talbi
Laboratoire d'Informatique Fondamentale de Lille
UMR CNRS 8022, Cité scientifique
INRIA Futurs – DOLPHIN
59655 Villeneuve d'Ascq cedex
<prenom>.<nom>@lifl.fr
http://www.lifl.fr/OPAC/

1 Introduction

En Data Mining, les données manipulées sont généralement larges et denses. Aussi, leur exploitation se révèle difficile en pratique. Le Data Mining Hautes Performances (DMHP) (Zaki 1999) s'applique à l'analyse efficace de telles masses de données. Différentes approches combinent la mise en oeuvre de méthodes performantes et extensibles (heuristiques), et le déploiement d'algorithmes sur architectures parallèles ou distribuées. A l'instar des techniques d'échantillonnage et de discrétisation, la sélection d'attributs constitue un troisième aspect, orienté « données », du DMHP. En effet, selon l'objet de l'étude, un certain nombre d'attributs s'avèrent non pertinents, signifiant que leur valeurs n'affectent en rien la procédure de traitement. D'autres, également inutiles, sont dits redondants *i.e.* fortement corrélés à d'autres champs de la structure n'apportant que peu d'information utile. Ceci justifie une sélection préalable des attributs, afin de réduire le coût de l'analyse de ces données.

On distingue généralement deux classes de méthodes selon que la sélection tienne compte ou non des résultats mesurés en phase d'apprentissage (Kohavi et al. 1996). Dans la première approche, dite « filtrante », la sélection se réalise une et une seule fois, avant le traitement et se base généralement sur une mesure de distance entre les enregistrements ou de similitude entre les attributs. Au contraire, l'approche « enveloppante » procède par cycles, composé chacun d'une étape de sélection puis d'exploitation des enregistrements réduits. On réitère le procédé où chaque nouvelle sélection générée est optimisée en tenant compte de la qualité du précédent modèle déduit. Cette approche est reconnue plus rigoureuse et la sélection est adaptée au processus d'extraction de connaissances, mais également plus coûteuse, puisqu'il convient d'appliquer tout un processus d'apprentissage pour chacune des sélections candidates. L'exploitation des grilles (Foster et al. 1999) permet, outre la distribution des calculs, le déploiement de modèles de résolution robustes basés sur l'hybridation d'algorithmes (Talbi 2002).

Ce chapitre est organisé ainsi : nous présentons d'abord le problème de sélection d'attributs en spectroscopie proche infra-rouge. Puis, nous proposons un algorithme génétique coopératif parallèle pour la résolution du problème. Enfin, avant de conclure, nous présentons les résultats expérimentaux obtenus sur une grille de 122 machines en utilisant la plate-forme ParadisEO-CMW dédiée à la conception de métaheuristiques parallèles hybrides sur grilles.

[1] Ce travail a été réalisé dans le cadre du projet Géno-Médicale (GGM) de l'ACI Masse de données.