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.

2 Sélection d'attributs en spectroscopie proche infra-rouge

La spectroscopie Proche Infra-Rouge (PIR) est une technique non destructrice exploitant l'absorption des rayonnements lumineux d'une substance à des fins d'analyse qualitative et quantitative. Le domaine d'application de nos travaux est la sucrerie de betteraves [2]. Il s'agit de prédire la concentration de saccharose dans les échantillons de betterave. Différents lots, dits respectivement de calibrage, de validation et de prédiction, ont été constitués. Le premier sera utilisé dans la constuction du prédicteur à la phase d'étalonnage. Les deux autres lots, quant à eux, seront exploités à des fins d'analyse de performance du modèle déduit, ceci afin de comparer les prédictions proposées à partir de nouveaux spectres d'absorbance aux concentrations réelles prédéterminées et n'ayant participé à la phase de calibrage.

L'analyse multivariée est un outil statistique très utilisé en instrumentation. Tout particulièrement, la méthode de régression linéaire des moindres carrés partiels (PLS) est appliquée avec succès dans la construction d'un modèle fiable à l'étalonnage. Il est reconnu qu'une sélection préalable des attributs est profitable. Les motivations sont multiples. Concevoir des modèles de calibrage plus simples et robustes, et améliorer la qualité des prédictions.

La détermination d'un ensemble optimal d'attributs par approche enveloppante est souhaitable mais pose un problème de faisabilité, justifié par le nombre élevé d'attributs (*1020*) et donc de sélections candidates (2^{1020}), et le coût d'évaluation d'une sélection donnée (déterminée par régression PLS basée sur un calcul matriciel intensif).

3 Un algorithme génétique coopératif insulaire parallèle

L'algorithme implémenté est constitué d'un modèle coopératif insulaire d'AGs générationnels dont la phase d'évaluation est distribuée suivant le modèle Maître-Esclave. Le choix de l'Algorithme Génétique est motivé par sa tendance à l'exploration et à la diversification, à sa capacité à traiter des espaces décisionnels de grandes dimensions.

Le modèle coopératif insulaire repose sur le déploiement à l'exécution de modèles naturellement parallèles basés sur des évolutions concurrentes et coopératives de populations distribuées dans l'espace. Cela permet une robustesse accrue et l'obtention de meilleurs résultats. Le déploiement comme ici d'A.Gs hétérogènes par leurs opérateurs génétiques de variation accentue encore davantage ce phénomène.

La phase d'évaluation est généralement la plus coûteuse dans l'exécution d'un A.G., et tout particulièrement dans le cadre d'applications issues du monde réel. Succédant à la phase dite de transformation, elle consiste à déterminer la qualité de chaque nouvelle solution produite. Ce calcul étant indépendant du reste de la population, la parallélisation est donc ici naturelle et basée sur un partitionnement des individus. Elle ne modifie en rien le comportement de l'A.G. original, de mêmes résultats sont obtenus plus rapidement.

Le protocole adopté pour le processus d'évaluation d'une sélection est le suivant. Un prédicteur est construit par régression PLS à partir des longueurs d'onde retenues et des données de calibrage. On associe la robustesse de ce dernier à la RMSEV, erreur standard

[2] Ce travail a été réalisé en collaboration avec le Laboratoire de Spectrochimie Infrarouge et Raman (LASIR) de Lille1 (*http://lasir.univ-lille1.fr*)

obtenue à partir des prédictions calculées sur les spectres du lot de validation et des concentrations associées réelles connues. Cette valeur de RMSEV constitue la « performance » d'une sélection candidate lors de son intégration dans la population.

Au terme de l'exécution des A.Gs coopératifs, la meilleure sélection globale (*i.e.* dont la RMSEV est minimale) est extraite. On construit un prédicteur à partir du lot de calibrage, mais on évalue cette fois sa performance sur le troisième et dernier lot d'échantillons dit de prédiction. Celui-ci n'est en effet jamais intervenu dans la phase d'optimisation. L'erreur obtenue constitue la RMSEP ou erreur de prédiction réelle du modèle précédemment généré.

4 Résultats expérimentaux

L'approche proposée a été implémentée en utilisant ParadisEO-CMW [3] (Cahon 2005). Il s'agit d'une plate-forme logicielle dédiée à la conception transparente et réutilisable de métaheuristiques parallèles hybrides sur grilles. Couplée à l'environement Condor MW (Goux et al. 2000), elle permet le déploiement des méthodes développées sur grilles. Elle intègre l'ensemble des algorithmes évolutionnaires et des méthodes à base de recherche locale (algorithmes de descente, recuit simulé, recherche Tabou), la plupart des modèles parallèles et des techniques d'hybridation (Talbi 2002). ParadisEO-CMW est basée sur une séparation conceptuelle claire entre la partie invariante liée aux méthodes de résolution et celle spécifique aux problèmes traités, lui conférant ainsi une grande flexibilité et une réutilisation maximale du code et des modèles.

La sélection générée par l'AG. simple permet d'obtenir un gain en précision de 41,91% par rapport à la précision obtenue avec la méthode PLS sans sélection. Dans le modèle hybride, le phénomène de convergence est manifestement retardé. Ainsi, un gain de 47,78% est alors mesuré. Remarquons néanmoins que cette forme d'hybridation, si elle offre un apport significatif dans la qualité des solutions produites, a un coût au déploiement très important. En effet, la consommation en ressource CPU est multipliée par le nombre d'AGs coopératifs déployés (*i.e.* 16).

La sélection quasi-optimale obtenue par le modèle en tore se compose d'un ensemble de 122 attributs parmi les 1020 longueurs d'onde candidates initiales. L'élimination de près de 88% des fréquences permet une accélération très importante du traitement d'analyse de chaque échantillon : 394 μ s. contre 2930 initialement (PC cadencé à 1 Ghz).

Le caractère naturellement hétérogène et dynamique des grilles rend difficile les mesures de performance à l'exécution. Les métriques communément utilisées en calcul haute performance (*e.g.* l'accélération parallèle) ne peuvent convenir ici puisqu'elles ne s'appliquent qu'à un ensemble de ressources homogènes et dédiées. En normalisant le temps CPU nécessaire à la réalisation d'une tâche avec la performance d'un Noeud Travailleur (NT) correspondant, ParadisEO agrège les temps d'exécution et établit des statistiques viables sur différentes exécutions. Ce facteur de normalisation peut être basé sur une information matérielle, si disponible (*e.g.* MIPS, FLOPS, …). Alternativement, ParadisEO déploie des benchmarks sur les NTs participants afin de quantifier leur performance.

[3] PARAllel and DIStributed Evolving Objects on Condor MasterWorker (*http://www.lifl.fr/OPAC/paradiseo/*)

Dans (Goux et al. 2000), différentes métriques de performance ont été définies formellement, et tout particulièrement l'efficacité parallèle sur grilles, composés de ressources de calculs hétérogènes et dynamiques dans l'espace et le temps.

En considérant les informations suivantes,

- *U(i)*, le temps de disponibilité cumulée du NT *i*,
- *t(j)*, le temps utilisateur passé par le NT *w(j)* à accomplir la tâche *j*,

nous pouvons définir l'efficacité parallèle ainsi

$$\mu = \frac{\sum_{j \in J} t(j)}{\sum_{i \in I} U(i)}$$

Le tableau TAB 1 établit quelques statistiques mesurées à l'exécution de l'AG. coopératif insulaire (de topologie torique) parallèle sur un réseau de 122 stations de travail interconnectées dans le cadre universitaire, initialement dédiées à l'enseignement doc non dédiées, et globalement sous-exploitées.

Nombre de NTs	122
Durée d'exécution réelle	36953 s. (10 heures)
Durée d'exécution cumulée	2363571 s. (27 jours)
Nombre moyen de NTs actifs	115
Nombre moyen de NTs participants	78
Performance parallèle	0.82

TAB 1 - *Statistiques mesurées à l'exécution de l'A.G. coopératif parallèle.*

5 Conclusion

La sélection d'attributs en Data Mining réduit l'information intiale, en identifiant les attributs redondants ou ceux dont la valeur ne contribue pas à la phase d'exploitation des données. Les deux principales approches de ce processus, dites filtrante et enveloppante, se distinguent par un coût à la mise en oeuvre et une exactitude antagonistes. De l'approche enveloppante, préférable car adaptée au processus d'analyse des données, émergent deux problématiques majeures : le caractère NP-difficile de ce problème d'optimisation combinatoire, et « l'évaluation » généralement très gourmande en ressources d'une sélection candidate. Aussi, l'emploi de méthodes approchées apparaît encore insuffisant.

Le développement d'heuristiques doit être accompagné d'un déploiement sur environnements d'exécution parallèles à grande échelle que sont les grilles de calcul. L'exploitation de plate-formes logicielles telles ParadisEO-CMW réduit considérablement l'effort en développement, masquant la forte complexité inhérente à la mise en oeuvre d'applications parallèles sur grilles. Elle a contribué au développement de méthodes efficaces et performantes face à une application industrielle en spectroscopie PIR, difficile en pratique. Les premiers résultats obtenus se montrent très satisfaisants tant en terme de qualité des solutions obtenues que des performances mesurées à l'exécution.

Références

Cahon S. et Melab N. et E-G. Talbi (2005), An Enabling Framework for Parallel Optimization on the Computational Grid, in the Proc. of the 5th IEEE/ACM Intl. Symposium on Cluster Computing and the Grid (CCGRID'2005), Cardiff, UK., 2005.

Foster I. et Kesselman C. (1999), The Grid: Blueprint for a New Computing Infrastructure, Morgan-Kaufmann, 1999.

Goux J-P. et Kulkarni S. et Yoder M. et Linderoth J., An Enabling Framework for Master-Worker Applications on the Computational Grid, in HPDC '00: Proceedings of the Ninth IEEE International Symposium on High Performance Distributed Computing (HPDC'00), IEEE Computer Society, 2000.

Kohavi R. et John G.H. (1996), Wrappers for feature subset selection, AIJ special issue on relevance, 1996.

Talbi E-G. (2002), A Taxonomy of Hybrid Metaheuristics, Journal of Heuristics, Kluwer Academic Publishers, Vol. 8, pp 541-564, 2002.

Zaki M. (1999), Parallel and Distributed Data Mining: An Introduction, Large-Scale Parallel Data Mining, pp 1-23, 1999.

Summary

Feature selection in Data Mining reduces the cost of the data processing and aims to find simpler and robust models of knowledge. This NP-hard problem is practically difficult as it processes large and dense databases. Hence, its resolution requires not only the development of approached methods, but the design of hybrid techniques and the deployment on large-scale parallel platforms too. In this paper, we present a wrapper feature selection approach for specroscopic data mining. The approach has been implemented using ParadisEO-CMW a software framework dedicated to the design and the deployment of parallel and hybrid metaheuristics on grids. The results are convincing both in terms of quality of provided solutions and efficiency at execution.