Vers l'exploitation de grandes masses de données

Raphaël Féraud, Marc Boullé, Fabrice Clérot , Françoise Fessant

France Télécom R&D, avenue Pierre Marzin, 22307 Lannion
Contact : raphael.feraud@orange-ftgroup.com

Résumé : Une tendance lourde depuis la fin du siècle dernier est l'augmentation exponentielle du volume des données stockées. Cette augmentation ne se traduit pas nécessairement par une information plus riche puisque la capacité à traiter ces données ne progresse pas aussi rapidement. Avec les technologies actuelles, un difficile compromis doit être trouvé entre le coût de mise en œuvre et la qualité de l'information produite. Nous proposons une approche industrielle permettant d'augmenter considérablement notre capacité à transformer des données en information grâce à l'automatisation des traitements et à la focalisation sur les seules données pertinentes.

Mots clés : fouille de données, grande volumétrie, sélection de variables, sélection d'instances.

1 Introduction

Selon Fayyad et al (1996), le Data Mining est un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données. Plusieurs intervenants industriels ont proposé une formalisation de ce processus, sous la forme d'un guide méthodologique nommé CRISP-DM pour CRoss Industry Standard Process for Data Mining, voir Chapman et al (2000). Le modèle CRISP-DM (FIG 1) propose de découper tout processus Data Mining en six phases:

1. La phase de *recueil des besoins* fixe les objectifs industriels et les critères de succès, évalue les ressources, les contraintes et les hypothèses nécessaires à la réalisation des objectifs, traduit les objectifs et critères industriels en objectifs et critères techniques, et décrit un plan de résolution afin d'atteindre les objectifs techniques.
2. La phase de c*ompréhension des données* réalise la collecte initiale des données, en produit une description, étudie éventuellement quelques hypothèses à l'aide de visualisations et vérifie le niveau de qualité des données.
3. La phase de *préparation des données* consiste en la construction d'une table de données pour modélisation (Pyle, 1999; Chapman et al, 2000). Nous nous y intéressons plus particulièrement par la suite.
4. La phase de *modélisation* procède à la sélection de techniques de modélisation, met en place un protocole de test de la qualité des modèles obtenus, construit les modèles et les évalue selon le protocole de test.
5. La phase *d'évaluation* estime si les objectifs industriels ont été atteints, s'assure que le processus a bien suivi le déroulement escompté et détermine la phase suivante.