Un Outil d'extraction de connaissances à partir des données : DataMiner 1.0

Mme Louni S., Me Adjiri R. et Me Zeghichi R.

Laboratoire des Logiciels de base, CERIST.
3, Rue des frères Aissiou Ben Aknoun Alger, Algérie.
Email : louni@tassili.cerist.dz

1. Introduction

Les faibles coûts des machines en terme de stockage et de puissance de calcul et le développement des techniques informatiques ont encouragé les entreprises et les organisations à accumuler de grandes masses de données qui sont souvent sous exploitées alors qu'elles peuvent renfermer des connaissances stratégiques que des experts peuvent ignorer.

Ces réservoirs de données représentent une importante mine d'informations que les entreprises doivent exploiter et explorer pour découvrir des informations pertinentes et utiles à des fins de prédiction et de prise de décisions. Les mécanismes de Data Mining (DM) ou encore de Knowledge Discovery in Databases (KDD) sont alors mis en place pour répondre à ce besoin à travers un nouveau domaine de recherche qui se situe au carrefour de plusieurs disciplines à savoir : les bases de données, l'apprentissage automatique et les statistiques.

2. Problématique

Le problème de recherche ou de découverte de relations, de liens et de règles générales à partir d'instances de données est devenu un problème très populaire et fait en général référence aux machines d'apprentissage (learning machine) [Kodratoff 93] et aux techniques de découverte de Dépendances Fonctionnelles [Mannila 94]. La différence essentielle entre ces deux méthodes, est le domaine d'application. L'apprentissage (avec par exemple les arbres de décision et les réseaux de neurones) classifie les exemples et offre ainsi un moyen très important pour faire de la prédiction et de la prise de décision, alors que les techniques d'extraction de dépendances fonctionnelles infèrent des dépendances fonctionnelles valides ou non valides, certaines ou incertaines à partir de relations et offre ainsi un moyen efficace d'aide dans la conception du schéma de bases de données ou encore peuvent être utilisées dans le domaine du reverse engineering.

3. Objectif et motivation

Les outils de datamining construisent des modèles de manière plus ou moins interactive avec l'utilisateur. A l'extrême, on trouve des produits presse-bouton qui s'adressent à des non-spécialistes. Les produits intermédiaires proposent généralement une certaine interaction avec l'utilisateur tant dans le paramétrage de l'apprentissage que pendant la recherche du modèle. A l'autre extrême, les techniques statistiques requièrent un maniement par des statisticiens professionnels, bien que certains outils commencent à évoluer vers une meilleure convivialité et une assistance à l'utilisateur accrue (Clementine [Clementine], DBMiner2.0 [DBMiner]).

La lisibilité ou la puissance ? Notre objectif essentiel a été d'offrir un modèle présentant un bon pouvoir de prédiction et une bonne lisibilité des résultats. Il existe un compromis entre clarté du modèle et pouvoir prédictif. Plus un modèle est simple, plus il sera facile à comprendre, mais moins il sera

capable de prendre en compte des dépendances subtiles ou trop variées (non linéaires). La figure 1 [Lefébure 98] illustre ce compromis. Les arbres de décision sont très faciles à interpréter. Néanmoins, ces techniques ne reconnaissent que des frontières nettes de discrimination. L'existence de relations d'interdépendances entre les variables conduit à une diminution de la performance du modèle. Les réseaux de neurones, par leur capacité à intégrer les relations entre les variables, présentent un pouvoir prédictif élevé. Néanmoins, ce progrès entraîne une perte de lisibilité, compte tenu de la complexité du modèle mathématique sous-jacent.

Cette relative antinomie entre lisibilité et puissance a un impact fort sur le type d'utilisateurs. Ainsi, les arbres de décision, de par leur forte lisibilité, s'adressent davantage à des utilisateurs métier. Au contraire, les réseaux de neurones ou bayésiens nécessitent des experts de la modélisation.

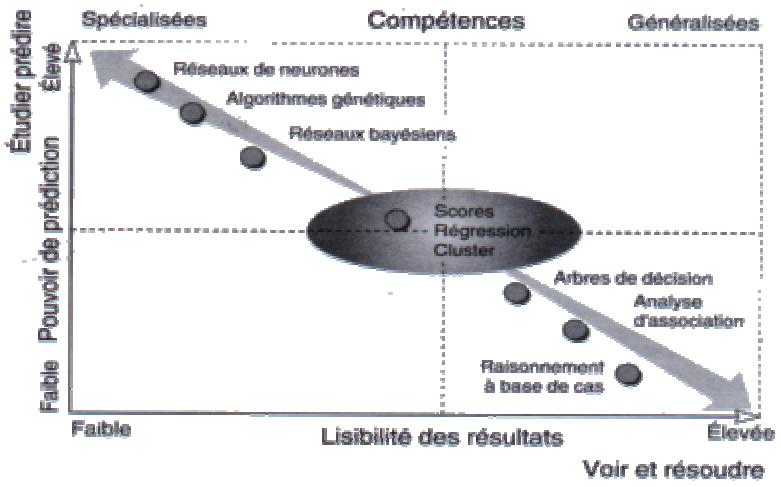

Figure 1 : Le Compromis entre lisibilité et prédiction

Il est important à noter qu'il est indispensable de disposer dans un outil de Data Mining de plusieurs techniques d'extraction de relations et de connaissances afin de réussir à modéliser le plus grand nombre de problèmes qui puissent se présenter. Pour ce qui est de l'extraction de connaissances par apprentissage, un compromis a été pris d'offrir aussi bien un modèle de réseaux de neurones qu'un modèle à base d'arbre de décision.

Ainsi, notre objectif a été de concevoir et de réaliser un outil prototype d'extraction de connaissances à partir de bases de données relationnelles, pouvant être utilisé aussi bien dans le domaine de la prédiction et de la prise de décision que dans le domaine de la conception de base de données.

Le système offre deux grandes fonctionnalités :

- L'aptitude d'extraire des dépendances fonctionnelles (DFs) valides et non valides, qui peuvent être utilisées par les concepteurs de bases de données durant la conception du schéma conceptuel de bases de données ou encore dans le domaine du « reverse engineering » par l'analyse de bases de données existantes, ou enfin dans le domaine de la sécurité des bases de données pour le contrôle de l'inférence dans les bases de données[ldb1][1]. L'algorithme Bottom_up [Savnik 93] a été amélioré et mis en œuvre.
- L'aptitude à établir des prédictions à partir d'apprentissages automatiques à base d'arbres de décision [Kodratoff 93] et de réseaux de neurones [Davalo 93].

Notre motivation essentielle a été de mettre en pratique nos connaissances dans ce nouveau domaine et les algorithmes étudiés et proposés.

4. Description de DataMiner1.0