heig-vd
Haute Ecole d'Ingénierie et de Gestion
du Canton de Vaud

Tél. : 024 557 63 30
Fax : 024 557 64 04

info@heig-vd.ch
http://www.heig-vd.ch

Travail de diplôme 2005

Département Electricité et Informatique

HEIG-VD
Route de Cheseaux 1
CH - 1401 Yverdon-les-Bains

Modules de prétraitement de données dans le cadre du Data Mining

Contexte

Ce projet s'intègre dans un outil d'analyse de données dans le cadre du projet DEPROLO (DEcomposition de PROblèmes LOgistiques).

Description du problème

Dû à la grande taille des bases de données actuelles, les données brutes sont généralement de faible qualité. Elles peuvent être incomplètes (valeurs manquantes ou agrégées), bruitées (valeurs erronées ou aberrantes) ou incohérentes (divergence entre attributs). L'application d'algorithmes de data mining sur de telles données complexifie l'apprentissage et nuit à la performance ainsi qu'à la fiabilité du modèle.

Le prétraitement des données est une étape cruciale dans le processus de découverte de connaissances à partir de grandes bases de données. En effet, il permet d'améliorer la qualité des données soumises par la suite aux algorithmes de data mining. Le prétraitement des données est constitué de différentes étapes successives :

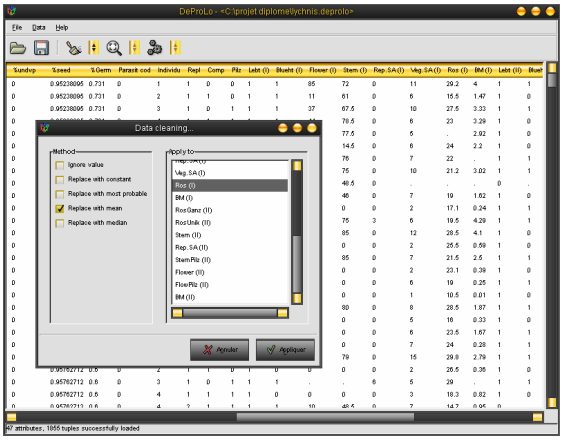
- Nettoyage
- Intégration
- Transformation
- Réduction
- Discrétisation

Réalisation

Différents modules ont étés implémentés dans le langage C. Chacun de ces modules réalise une technique ou un algorithme de prétraitement de données.

Ils offrent les possibilités suivantes :

- Identifier et traiter les données manquantes par remplacement.
- Identifier les données aberrantes et éliminer le bruit des données par lissage, clustering hiérarchique ou régression.
- Identifier les données redondantes par une analyse de corrélation.
- Résumer les données à différents niveaux de granularité.
- Normaliser les données.
- Réduire les données par sélection d'attributs.



Chacun de ces modules peut être utilisé indépendamment des autres, cependant leur combinaison est vivement conseillée afin d'obtenir les meilleurs résultats possibles.

Auteur: **Sandrine Curtet**
Prof. responsable: **Raileanu Laura Elena**

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale