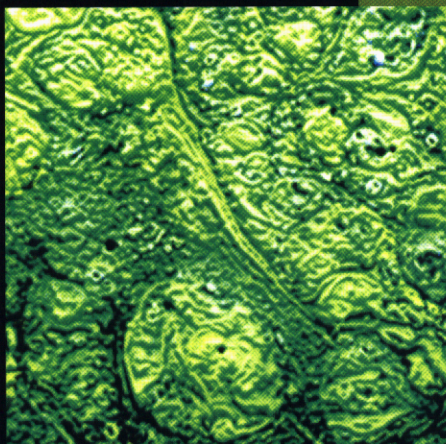


# Le Data Mining

René Lefébure  
Gilles Venturi



 Eyrolles

**INFORMATIQUES**

# Le Data Mining

Collection dirigée par Guy Hervier

7397

# Le Data Mining

PS 2658

René LEFÉBURE

Gilles VENTURI

Troisième tirage 1999

 **Eyrolles**

**INFORMATIQUES**

BIBLIOTHEQUE DU CERIST

# Table des matières

|                                                     |          |
|-----------------------------------------------------|----------|
| <b>Introduction</b>                                 | <b>3</b> |
| Une première définition du Data Mining .....        | 3        |
| Pourquoi cet engouement pour le Data Mining ? ..... | 4        |
| Structure de cet ouvrage .....                      | 5        |
| A qui s'adresse cet ouvrage .....                   | 6        |
| Mythes et réalités du Data Mining .....             | 7        |

## **1 • LES ENJEUX DU DATA MINING**

**11**

|                                                                                                           |    |
|-----------------------------------------------------------------------------------------------------------|----|
| La distinction entre données et connaissance .....                                                        | 11 |
| Trop de données tue l'information .....                                                                   | 14 |
| <b>Des données accessibles aux utilisateurs :</b><br><b>de l'infocentre à l'entrepôt de données</b> ..... | 16 |
| Les premiers infocentres : libérer l'utilisateur .....                                                    | 16 |
| Industrialiser l'infocentre : les entrepôts de données .....                                              | 18 |
| <b>Les systèmes opérationnels et décisionnels</b> .....                                                   | 20 |
| Les systèmes opérationnels .....                                                                          | 20 |
| Les systèmes décisionnels .....                                                                           | 21 |
| <b>Des contraintes et des environnements techniques</b><br><b>nécessairement différents</b> .....         | 22 |
| <b>Les utilisateurs accèdent directement à leurs données</b> .....                                        | 23 |
| <b>L'essor du Data Mining</b> .....                                                                       | 24 |
| Principaux domaines d'application .....                                                                   | 24 |
| Le marché des outils : une croissance de 40 % par an .....                                                | 25 |
| Quelques exemples de retour sur investissement constaté .....                                             | 26 |

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| <b>La notion de distance</b> .....                                                      | 65 |
| Les techniques de classification .....                                                  | 66 |
| <b>La notion d'association</b> .....                                                    | 74 |
| L'association sur des variables quantitatives .....                                     | 74 |
| La notion d'association sur des variables qualitatives .....                            | 81 |
| <b>Quelques notions concernant les probabilités<br/>et les arbres de décision</b> ..... | 84 |
| Principes de calcul .....                                                               | 85 |

## 4 • LES TECHNIQUES DE DATA MINING

89

|                                                         |     |
|---------------------------------------------------------|-----|
| <b>Introduction aux techniques de Data Mining</b> ..... | 89  |
| Data Mining : évolution plus que révolution .....       | 89  |
| Data Mining : un cocktail de techniques .....           | 89  |
| La donnée stockée ou analysée .....                     | 90  |
| L'utilisateur « métier » ou le statisticien .....       | 91  |
| La lisibilité ou la puissance .....                     | 92  |
| <b>Le raisonnement à base de cas</b> .....              | 94  |
| Définition et enjeux .....                              | 94  |
| Principes de construction d'un RBC .....                | 96  |
| Domaines d'application .....                            | 102 |
| Exemple de construction d'un RBC .....                  | 104 |
| Limites et avantages .....                              | 107 |
| <b>Les knowbots ou agents intelligents</b> .....        | 109 |
| Définition et enjeux .....                              | 109 |
| Principes de construction d'un agent .....              | 110 |
| Domaines d'application .....                            | 115 |
| Exemple du site Internet Firefly .....                  | 115 |
| Les limites .....                                       | 117 |
| <b>Les associations</b> .....                           | 118 |
| Définition et enjeux .....                              | 118 |
| Principes de construction des associations .....        | 120 |
| Domaines d'application .....                            | 122 |
| Exemples de recherche du risque avec Strada .....       | 124 |
| Les limites .....                                       | 128 |
| <b>Les arbres de décision</b> .....                     | 131 |

|                                                            |     |
|------------------------------------------------------------|-----|
| Les vendeurs de matériels .....                            | 211 |
| Les startups.....                                          | 211 |
| Les intégrateurs .....                                     | 212 |
| <b>Quelques critères de choix d'un logiciel</b> .....      | 212 |
| Les gammes de prix et de puissance .....                   | 213 |
| Les caractéristiques informatiques .....                   | 214 |
| Adéquation de l'outil avec la complexité du problème ..... | 215 |
| Le niveau de compétence requis .....                       | 217 |
| L'expérimentation reste le meilleur critère de choix.....  | 218 |
| Un exemple de grille d'évaluation .....                    | 218 |
| <b>Intelligent Miner d'IBM</b> .....                       | 220 |
| <b>Clementine de ISL</b> .....                             | 222 |
| <b>SAS Enterprise Miner de SAS</b> .....                   | 223 |
| <b>4Thought de Cognos</b> .....                            | 225 |
| <b>Predict de NeuralWare</b> .....                         | 227 |
| <b>Previa de Elseware</b> .....                            | 229 |
| <b>Saxon de Pmsi</b> .....                                 | 231 |
| <b>Strada de Complex System</b> .....                      | 233 |
| <b>Scenario de Cognos</b> .....                            | 235 |
| <b>Alice de Isoft</b> .....                                | 236 |
| <b>Knowledge Seeker de Angoss</b> .....                    | 238 |
| <b>Datamind</b> .....                                      | 240 |
| <b>Wizwhy de Wizsoft</b> .....                             | 241 |
| <b>SPAD de CISIA</b> .....                                 | 243 |
| <b>Quelques références en matière de prix</b> .....        | 245 |
| Arbres de décision .....                                   | 245 |
| Réseaux de neurones .....                                  | 246 |
| Les intégrés.....                                          | 248 |
| <b>Présentation du cas</b> .....                           | 249 |

## 6 • ÉTUDE DE CAS

**249**

|                                                 |     |
|-------------------------------------------------|-----|
| <b>Phase 1 : poser le problème</b> .....        | 249 |
| <b>Phase 2 : la recherche des données</b> ..... | 250 |
| <b>Phase 3 : la sélection des données</b> ..... | 252 |
| <b>Phase 4 : le nettoyage des données</b> ..... | 253 |

Les valeurs aberrantes



**8 • UTILISATION DU CD-ROM 291**

|                                                      |     |
|------------------------------------------------------|-----|
| <b>Contenu</b> .....                                 | 291 |
| <b>Installation de Classpad</b> .....                | 291 |
| Sous Windows 3.1x ou Windows NT 3.5x .....           | 292 |
| Sous Windows 95 ou Windows NT 4.0 .....              | 292 |
| <b>Exploitation de Classpad</b> .....                | 292 |
| <b>Les données de l'exemple</b> .....                | 293 |
| <b>Le chargement des données</b> .....               | 294 |
| <b>La construction d'un réseau de neurones</b> ..... | 295 |
| <b>La construction d'un arbre de décision</b> .....  | 301 |
| <b>Et après...</b> .....                             | 305 |

**9 • RÉPERTOIRE DES FOURNISSEURS D'OUTILS 307**

|                                         |     |
|-----------------------------------------|-----|
| <b>Raisonnement à base de cas</b> ..... | 307 |
| <b>Visualisation graphique</b> .....    | 307 |
| <b>Logique floue</b> .....              | 308 |
| <b>Knowledge discovery</b> .....        | 309 |
| <b>Réseaux neuronaux</b> .....          | 310 |

**Bibliographie 313**

|                                                |     |
|------------------------------------------------|-----|
| <b>Quelques sites web</b> .....                | 313 |
| <b>Revue traitant du Data Mining</b> .....     | 314 |
| <b>Ouvrages et articles intéressants</b> ..... | 314 |

**Glossaire 321****Index 327**