

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Institut National de formation en Informatique
OUED SMAR

Mémoire pour l'obtention du diplôme
d'ingénieur d'état en informatique
option

Système Informatique

THEME

Conception et réalisation d'un outil d'Extraction
de Connaissances à partir des Données
(DataMiner1.0)

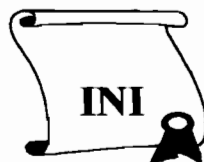
Réalisé par :

ADJIRI Réda

ZEGHICHI Réda

Encadré par :

M^{me} S. LOUNI (C.E.R.I.S.T)



Promotion 1998/1999



Résumé :

De nos jours, les organisations, les entreprises et autres institutions accumulent de plus en plus de données. Cette importante masse de données renferme des informations inconnues auparavant bien qu'elles soient stratégiques et très utiles pour l'entreprise. Cette sous-exploitation de l'information stockée a focalisé l'attention des chercheurs ce qui donner naissance à un nouveau domaine de recherche se situant au carrefour de trois domaines habituellement séparés, à savoir : les bases de données (BD), l'apprentissage automatique et les statistiques. On parle dans ce cas d'Extraction automatique de connaissances à partir de données ou encore de Data Mining.

Ce mémoire présente notre contribution à la conception et à la réalisation d'un outil prototype de Data Mining, appelé « *DataMiner1.0* » qui, afin d'extraire des connaissances à partir de données, utilise les techniques suivantes :

- Il utilise, d'une part, une technique adaptée de l'algorithme Bottom-up permettant d'extraire des Dépendances Fonctionnelles dans le but d'aider les concepteurs de bases de données durant la phase de conception « Design » du schéma de la base et pour permettre le contrôle de l'inférence dans les bases de données durant le « Design time ».
- Il utilise, d'autre part, deux techniques d'apprentissage automatique inductif, à savoir les Arbres de décision et avec les Réseaux de neurones qui ont chacune pour résultats un arbre de décision avec un ensemble de règles, et un réseau de neurones pour faire des prédictions.

Mot clés :

KDD, Data Mining, Apprentissage, Dépendances Fonctionnelles (DF), Arbres de décision, Réseaux de neurones, Algorithme Bottom-up, propagation, retro-propagation, Data warehouse.

Sommaire

Introduction générale

1. Introduction	I
2. Problématique	I
3. Objectifs	II
4. Organisation de ce mémoire	II

Chapitre I : Background

1. Introduction	1
2. Les concepts de base des Bases de Données	1
2.1. Définitions d'une BD	1
2.2. Définition d'un Système de Gestion de Bases de Données (SGBD)	2
2.3. Quelques définitions	2
2.4. Les dépendances fonctionnelles	3
2.4.1. Définition	3
2.4.2. Propriétés des DF	3
2.4.3. Les DF élémentaires	3
2.4.4. Les DF directes	3
2.4.5. Couverture minimale	4
3. Les concepts de bases de l'apprentissage	4
3.1. Définition de l'IA	4
3.2. L'intelligence artificielle et l'apprentissage	4
3.3. L'apprentissage	4
3.3.1. L'induction	5
3.3.2. L'apprentissage inductif	5

3.3.3. L'apprentissage supervisé et l'apprentissage non supervisé	6
4. Les statistiques	7
4.1. La régression	7
4.2. La classification	7
5. Conclusion	8

Chapitre II : Le KDD et le Data Mining

1. Introduction	9
2. Différence entre données et connaissances	9
3. Le KDD	10
3.1. Définitions du KDD	10
3.2. Le processus du KDD	11
4. Relation entre le KDD et le Data Mining	16
5. Définitions du Data Mining	16
6. Quelques applications du DataMining	18
7. Conclusion	19

Chapitre III : Les techniques du Data Mining

1. Introduction	20
2. Les techniques de Data Mining	21
2.1. L'extraction de dépendances fonctionnelles	21
2.1.1. L'algorithme naïf (naïve algorithm)	21
2.1.2. L'algorithme amélioré (improved algorithm)	21
2.1.3. L'algorithme Hypergraph Transversal	22
2.1.4. L'algorithme basé sur le tri (a sort-based algorithm)	22
2.1.5. L'approche SQL	22
2.1.6. L'algorithme Bottom-up	23
2.1.7. Choix d'un algorithme	23
2.2. Les techniques d'apprentissage	24

2.2.1. Raisonnement à base de cas	24
2.2.2. La recherche d'associations	24
2.2.3. Les réseaux bayésiens	25
2.2.4. Les arbres de décision	25
2.2.5. Les réseaux de neurones	25
2.2.6. Comparaison entre les techniques d'apprentissage	26
3. Conclusion	27

Chapitre IV : Les techniques de Data Mining choisies

1. Introduction	28
2. Les arbres de décision	28
2.1. Introduction	28
2.1.1. Base d'exemples	28
2.1.2. Induction logique et induction incertaine	29
2.1.3. Arbre d'induction	29
2.2. Exemple de construction d'un arbre d'induction	29
2.3. Critère de sélection d'une variable : "la notion d'information"	32
2.3.1. Heuristique de la notion d'information	32
2.3.2. Information d'une variable	33
2.3.3. L'information conditionnée par une variable	33
2.3.4. Le gain d'information	34
2.4. Conditions d'arrêt lors de la construction d'un arbre en induction incertaine	34
2.4.1. Faible nombre de cas	34
2.4.2. Faible information apportée par une variable	35
3. Les Réseaux de neurones	35
3.1. Introduction	35
3.2. Avantages de l'approche connexionniste	36
3.3. L'analogie avec le cerveau	36
3.3.1. La structure des neurones	36

3.3.2. Fonctionnement des neurones	37
3.4. Modélisation	38
3.4.1. Le neurone formel	38
3.4.2. L'organisation en couches	39
3.4.3. L'auto-apprentissage	39
3.5. L'apprentissage	39
3.5.1. La rétro-propagation	40
3.5.2. Les modes d'apprentissage	41
3.6. Les paramètres du réseau	42
4. L'extraction de Dépendances Fonctionnelles	43
4.1. Introduction	43
4.2. Les Dépendances Fonctionnelles	43
4.2.1. Le principe du Bottom-up	43
4.2.2. Quelques définitions	44
4.2.3. L'algorithme	45
4.2.4. Les DF incertaines	48
5. Conclusion	49

Chapitre V : Un outil prototype pour l'extraction de connaissances (DataMiner1.0)

1. Introduction	50
2. Notre système de Data Mining	50
3. Le DataMiner1.0	51
3.1. le module de création	53
3.2. Le module de nettoyage	53
3.3. Extraction de DF	54
3.3.1. Calcul de l'ensemble des DF invalides	56
3.3.2. Construction de l'ensemble des DF valides à 100%	61
3.3.3. Calcul de l'incertitude pour les DF non valides à 100%	65
3.4. Apprentissage avec les arbres de décision	66

3.4.1. Entrée du processus de construction de l'arbre de décision	66
3.4.2. Construction de l'arbre	66
3.4.3. L'utilisation de l'arbre de décision	67
3.5. Apprentissage avec les réseaux de neurones	69
3.5.1. Le codage des entrées et La normalisation de la sortie	69
3.5.2. La définition de l'architecture	71
3.5.3. Apprentissage	71
3.5.4. L'utilisation du réseau de neurones	73
3.6. le module de la prédiction	73
3.7. Le module de visualisation	75
4. Conclusion	75

Chapitre VI : Tests et évaluations des résultats

1. Introduction	75
2. L'environnement de développement	75
3. Schéma de réalisation	75
3.1. La création d'une Base d'exemple	77
3.2. Modification de la Base	77
3.3. La partie extraction de Dépendances Fonctionnelles	78
3.3.1. L'interface	78
3.3.2. Test d'exécution	80
3.4. La partie Apprentissage avec les arbres de décision	85
3.4.1. L'interface	85
3.4.2. Tests et validation	87
3.5. La partie Apprentissage par réseaux de neurones	93
3.5.1. L'interface	93
3.5.2. Tests et validation	94
4. Conclusion	98

Conclusion Générale 99

Références Bibliographiques 101

Annexes

Annexe 01: Les principales structures de données

Annexe 02: Une présentation du Data Warehouse

Annexe 03: Le C++ Builder

Annexe 04: Un rappel sur les graphes