

Université Lumière Lyon2

Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE

le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur

Gilles Venturini, Rapporteur

Mohand-Saïd Hacid, Examineur

Michel Lamure, Examineur

Gilbert Ritschard, Examineur

Nicolas Nicoloyannis, Directeur de thèse

Professeur, Facultés Universitaires N.D. de la Paix, Namur

Professeur, Université de Tours

Professeur, Université Claude Bernard-Lyon 1

Professeur, Université Claude Bernard-Lyon 1

Professeur, Université de Genève

Professeur, Université Lumière-Lyon 2

Table des matières

1	Introduction, Préambule	1
2	Concepts, Notions et Notations Utiles	7
2.1	Données Catégorielles	7
2.1.1	Domaines et Attributs Catégoriels	8
2.1.2	Objets Catégoriels	9
2.1.2.1	Similarités, Dissimilarités entre Objets Catégoriels	10
2.1.3	Ensemble d'Objets Catégoriels	11
2.1.3.1	Mode d'un Ensemble d'Objets Catégoriels	11
2.1.3.2	Similarités et Dissimilarités entre Ensembles d'Objets Catégoriels	12
2.1.3.3	Similarités et Dissimilarités au sein d'un Ensemble d'Objets Catégoriels	12
2.1.3.4	Voisinage d'une Partition d'un Ensemble d'Objets Catégoriels	13
2.2	Le Nouveau Critère de Condorcet	13
3	Classification Non Supervisée	15
3.1	Introduction	15
3.1.1	Méthodologie Générale de la Classification Non Supervisée	16
3.1.2	Applications de la Classification Non Supervisée	16
3.1.3	Taxonomies des Méthodes de Classification Non Supervisée	17
3.1.4	Méthodes de Classification Non Supervisée pour Données Catégorielles	19
3.1.5	Challenges Actuels en Classification Non Supervisée	22
3.2	Une Nouvelle Méthode de Classification Non Supervisée "Orientée Utilisateur"	24
3.2.1	Critère d'Évaluation de l'Aspect Naturel d'une Partition d'Objets	24
3.2.2	La Méthode de Classification Non Supervisée "Orientée Utilisateur"	26
3.2.2.1	Travaux Liés et Spécificités du Travail	26
3.2.2.2	L'Algorithme de Classification Non Supervisée	27

3.2.2.3	Complexité de l'Algorithme	29
3.2.2.4	Qualités de la Méthode pour l'Utilisateur . . .	30
3.2.2.5	Illustration du Fonctionnement de l'Algorithme	30
3.2.3	Evaluation de l'Algorithme de Classification non Supervisée	31
3.2.3.1	Evaluation de la Validité des Classifications . .	31
3.2.3.2	Evaluation de la Stabilité	37
3.2.3.3	Evaluation de l'Efficacité Algorithmique	40
3.2.4	Eléments Additionnels	42
3.2.4.1	Valeurs Spécifiques pour le Domaine des Variables Catégorielles	42
3.2.4.2	Gestion des Valeurs Manquantes :	44
3.2.4.3	Introduction de Contraintes :	44
3.2.4.4	De l'Apprentissage Non Supervisé à l'Apprentissage Supervisé : l'Apprentissage Non Supervisé sous Contraintes	50
3.3	Conclusion	54
4	Validité en Apprentissage Non Supervisé	57
4.1	Validité d'une Classification Non Supervisée : Définition et Evaluation	58
4.1.1	Mode d'Evaluation par Critères Externes	59
4.1.1.1	Méthode de Monte Carlo	59
4.1.1.2	Mesures Statistiques	60
4.1.2	Mode d'Evaluation par Critères Internes	61
4.1.3	Modes d'Evaluation Relatifs	63
4.1.3.1	Cas 1 : Le nombre final de classes, nc , n'est pas contenu dans P_{alg}	63
4.1.3.2	Cas 2 : Le nombre final de classes, nc , est contenu dans P_{alg}	64
4.1.3.3	Indices	64
4.1.4	Autres Modes d'Evaluation	67
4.2	Nouveaux Indices et Nouvelle Méthodologie pour l'Evaluation et la Comparaison de la Validité de Classifications Non Supervisées	68
4.2.1	Concepts et Formalismes Introductifs	69
4.2.1.1	Evaluation de l'homogénéité interne des classes d'une cns	71
4.2.1.2	Evaluation de la séparation entre classes d'une cns (ou hétérogénéité entre classes) 72	
4.2.1.3	Notions Additionnelles	73
4.2.1.4	Remarques importantes concernant l'aspect calculatoire	73

4.2.2	La nouvelle méthodologie pour l'évaluation et la comparaison de validité de cns	75
4.2.2.1	Caractérisation statistique des valeurs de: <i>LM</i> et <i>NLD</i>	76
4.2.2.2	Méthodologie	77
4.2.2.3	Expérimentations	82
4.2.2.4	Expérimentations sur le jeu de données Small Soybean Disease	82
4.2.3	Expériences sur le jeu de données Mushrooms	92
4.2.3.1	Description	92
4.2.3.2	Analyse des Résultats	95
4.2.4	Résumé et Informations Supplémentaires	96
5	Sélection de Variables, Contributions pour l'apprentissage supervisé et non supervisé	105
5.1	Sélection de Variables pour l'Apprentissage Supervisé	107
5.1.1	Caractéristiques de la Sélection de Variables	107
5.1.2	Les Types de Méthodes	107
5.1.3	Directions de Recherche	108
5.1.3.1	Forward Selection (FS) (Ajout de variables)	108
5.1.3.2	Backward Elimination (BE) (Suppression de variables)	109
5.1.3.3	Méthodes Bidirectionnelles	109
5.1.4	Stratégie de Recherche	109
5.1.5	Fonction d'Evaluation	110
5.1.6	Critère d'Arrêt	111
5.1.7	Approches Filtres	111
5.1.8	Approches Enveloppes	114
5.1.9	Autres Approches	115
5.2	Contribution à la Sélection de Variables pour l'Apprentissage Supervisé: Une Nouvelle Méthode Efficace et Rapide	118
5.2.1	Hypothèses et Idées Fondamentales	118
5.2.2	Evaluation de la Validité d'une Partition dans un Sous-Espace de l'ERD	119
5.2.3	La Nouvelle Méthode de Sélections de Variables	120
5.2.3.1	La Méthode de Base: une Méthode Exhaustive	121
5.2.3.2	Réduction de la Complexité par Introduction d'un AG	124
5.2.4	Evaluation Expérimentale	126
5.2.4.1	Présentation de l'Evaluation Expérimentale	126
5.2.4.2	Analyse de l'Evaluation Expérimentale	127
5.2.5	Conclusion	131
5.3	Contribution à la Sélection de Variables pour l'Apprentissage Non Supervisé: Une Nouvelle Méthode Efficace et Rapide	143

5.3.1	Evaluation de l'Adéquation entre deux Ensembles de Variables	144
5.3.2	Remarques Importantes Concernant l'Aspect Calculatoire 145	
5.3.3	Evaluation de l'adéquation entre EV un Ensemble de Variables et EV_★ un Sous Ensemble de EV ($\mathbf{EV}_\star \subseteq \mathbf{EV}$)	146
5.3.4	Evaluation/Comparaison de l'Adéquation entre un Ensemble de Variables (EV) et des Sous Ensembles de EV	148
5.3.5	La Nouvelle Méthode de Sélection de Variables	148
5.3.6	Evaluations Expérimentales	149
5.3.6.1	Expérience #1 : Evaluation expérimentale sur jeux de données synthétiques	149
5.3.6.2	Expérience #2 : Evaluation Expérimentale sur Jeux de Données de l'UCI	154
5.3.7	Conclusion	162
6	Agrégation de Classifications Non Supervisées : La Problématique "Cluster Ensembles"	165
6.1	Introduction	165
6.1.1	Illustration de la Problématique "Cluster Ensembles" . .	167
6.1.2	Motivations, Objectifs de la Problématique "Cluster Ensembles"	168
6.1.2.1	Réutilisation de Connaissances	169
6.1.2.2	Calcul Distribué pour la cns	169
6.1.3	Travaux Liés	172
6.1.4	Principaux Challenges pour la Problématique "Cluster Ensembles"	175
6.2	Mesures d'Adéquation	175
6.2.1	Adéquation entre Classifications Non Supervisées . . .	176
6.2.2	Adéquation pour un Couple de Classification Non Supervisée	176
6.2.3	Adéquation entre une Classification Non Supervisée et un Ensemble de Classifications Non Supervisées	177
6.3	Contribution à la Problématique "Cluster Ensembles" : Trois Méthodes pour l'Agrégation de Classifications Non Supervisées . .	177
6.3.1	Première Méthode pour l'Agrégation de cns: Une Méthode Intuitive	178
6.3.2	Seconde Méthode pour l'Agrégation de Classifications Non Supervisées : Utilisation de la Méthode KEROUAC .	179
6.3.2.1	Utilisation de KEROUAC pour la cns en considérant des Méta-Variables	179
6.3.2.2	Relation entre P_\star and P_β	180
6.3.2.3	Conclusion	181
6.3.2.4	Illustration	181

6.3.2.5	Propriétés de la Méthode	182
6.3.3	Troisième Méthode pour l'Agrégation de Classifications Non Supervisées : Utilisation de la Méthode K-Modes . .	183
6.3.3.1	Illustration	184
6.3.3.2	Propriétés de la Méthode	184
6.3.4	Evaluations Expérimentales	184
6.3.4.1	Evaluations, Comparaisons et Discussions Pré- liminaires	184
6.3.4.2	Evaluations, Comparaisons et Discussions Com- plémentaires	191
6.3.4.3	Comportement de la méthode KEROUAC face à des cns à agréger possédant des nombre de classes très différents	204
6.4	Conclusion	207
7	Conclusion	211
8	Données Utilisées pour les Expérimentations	217
8.1	Jeu de Données ADULT	217
8.2	Jeu de Données MUSHROOMS	218
8.3	Jeu de Données BREAST CANCER	220
8.4	Jeu de Données CAR	222
8.5	Jeu de Données : ADULT	224
8.6	Jeu de Données Contraceptive Method Choice	225
8.7	Jeu de Données FLAGS	226
8.8	Jeu de Données GERMAN	227
8.9	Jeu de Données HOUSE VOTES 84	229
8.10	Jeu de Données IONOSPHERE	230
8.11	Jeu de Données MONKS	231
8.12	Jeu de Données NURSERY	232
8.13	Jeu de Données PIMA	234
8.14	Jeu de Données SICK	235
8.15	Jeu de Données SMALL SOYBEAN DISEASES	236
8.16	Jeu de Données VEHICLE	237
8.17	Jeu de Données WINE	240
8.18	Jeu de Données SPAM	241
	Bibliographie	243
	Table des figures	254
	Liste des tableaux	257

Remerciements

Ce travail a commencé –et continuera je l’espère– avec Nicolas NICOLOYAN-NIS. Le hasard a bien voulu que je le rencontre il y a maintenant plus de 5 ans. De cette rencontre et de différentes circonstances a germé l’idée puis le projet de mes études doctorales auxquelles je ne me destinais pas forcément. Ses qualités humaines et scientifiques, son soutien, son amitié, ..., m’ont permis de mener à bien et avec grand plaisir et liberté ces travaux et je l’en remercie très chaleureusement.

MERCI NICOLAS !

Bien d’autres personnes ont contribué à rendre possible ce projet, je pense notamment ici à Ricco RAKOTOMALALA et Djamel ZIGHED qui ont guidé mes premiers pas au laboratoire ERIC ; Michel LAMURE et les membres du département Recherche et Technologie de la région Rhône-Alpes, le premier pour m’avoir fait confiance et intégré au sein d’un projet de recherche soutenu par la région et les seconds pour m’avoir accordé un financement de thèse.

De ces années, je retiendrai, outre le plaisir de la recherche, le bonheur d’avoir rencontré de nouveaux amis : Gaëlle et Laurent à qui ce travail doit énormément –enfin surtout à Gaëlle parce que Laurent...–.

GAELLE, LAURENT MERCI !

Je tiens également à souligner combien il fut agréable de parcourir ce chemin au sein du laboratoire ERIC dont je remercie l’ensemble des membres, et plus particulièrement les adeptes de discussions footballistiques, les buveuses et buveurs de thé ou café, ainsi que Astrid, Valérie et Lydie qui m’ont facilité certaines démarches.

D’un point de vue scientifique et humain, je voudrais rappeler le plaisir et l’honneur que m’ont fait d’accepter d’être membres de mon jury de thèse les Professeurs Mohand-Saïd HACID, Michel LAMURE, Gilbert RITSCHARD, Jean-Paul RASSON et Gilles VENTURINI. Je remercie les rapporteurs Jean-Paul RASSON et Gilles VENTURINI pour l’œil à la fois critique et bienveillant qu’ils ont bien voulu porter sur mes travaux. Je tiens à exprimer tout particulièrement ma reconnaissance envers :

- Jean-Paul RASSON, pour la précision de sa lecture de mon travail et la justesse des remarques qu’il m’a transmises ;

- Gilbert RITSCHARD, pour l'extrême et stupéfiante précision de sa lecture de mon travail ainsi que pour la justesse des remarques qu'il m'a transmises bien qu'il n'ait pas été rapporteur.

Vos remarques m'ont permis de poursuivre ma réflexion et d'améliorer ce document, même s'il est certain qu'il me faudrait un temps considérable pour pouvoir exploiter entièrement votre travail.

Je voudrais maintenant remercier de tout mon cœur Maman, Papa, Philippe et Emilie puisque je leur dois ce que je suis et bien plus encore...
MERCI, MERCI, MERCI ET ENCORE MERCI !

Bien que n'ayant pas véritablement contribué à un avancement raisonné de mes travaux, je remercie l'ensemble de mes amis pour avoir assuré "la préservation de ma santé mentale" (bien que...) et permis de vivre d'excellents moments. MERCI A TOUS !

Je voudrais également avoir ici une pensée des plus affectueuses pour mes grands-parents.

Enfin, terminons avec grâce et beauté : Karen ; Karen je te remercie le plus amoureuxment possible pour tout ton amour, tout ce que tu m'apportes et m'apprends... KAREN, MERCI !

Je tiens finalement à demander des excuses pour ceux que j'oublie ou ne cite pas nommément, pour mon incapacité à remercier en mesure de ce qui m'a été donné, et pour ceux dont la lecture de cette dissertation ne constituera pas un moment agréable ou utile...

Voilà, Merci à tous, et aux autres !

Pierre,

Grand Croix, Janvier 2004