

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 Préliminaires	184
6.3.4.2	Evaluations, Comparaisons et Discussions Complémentaires	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

1 Introduction, Préambule

"Le meilleur (...) ce n'est pas le mal réel qu'on se donne pour accoler le mot au mot, pour entasser brique sur brique ; ce sont les préliminaires, le travail à la bêche que l'on fait en silence en toutes circonstances, que ce soit dans le rêve ou à l'état de veille. Bref, la période de gestation. Personne n'a jamais réussi à jeter sur le papier ce qu'il avait primitivement l'intention de dire (...)"

- Henry Miller -
Sexus, (édition Buchet-Chastel) (1949)

Le traitement automatique de l'information fait appel à différentes ressources techniques, technologiques et théoriques issues de domaines variés tels que l'informatique, l'intelligence artificielle, la statistique, la théorie des probabilités, l'analyse de données, l'optimisation... La fin des années 80 et le début des années 90 ont vu un faisceau de situations favorables¹ à l'émergence d'un domaine de recherche transdisciplinaire s'attachant spécifiquement au traitement de vastes volumes d'information et à leur valorisation sous forme de connaissances : le "Knowledge Discovery in Databases" [PSCF⁺89] (expression que la communauté scientifique francophone a traduit plus tard par Extraction de Connaissances à partir de Données (ECD)).

Une quinzaine d'années plus tard, la communauté mondiale en ECD s'est élargie, structurée et a très largement essaimé dans le monde industriel. L'interaction entre la recherche et l'industrie explique certainement d'ailleurs, la rapidité de cette croissance, tout comme l'enthousiasme et l'activité de cette jeune communauté ont également contribué à cette émergence.

Plus encore, nous pensons que l'agitation, l'ébullition autour du phénomène ECD provient de l'adéquation des problématiques industrielles et académiques : ainsi les promesses de l'ECD en terme de valorisation de l'information ne pouvaient laisser insensibles les acteurs industriels au moment où l'information apparaît comme un élément stratégique déterminant.

1. sur le plan technologique (avancées technologiques en informatique : accroissement des capacités de stockage et de calculs), économique (passage de l'économie de l'ère post-industrielle à l'ère informationnelle)... et peut être épistémologique : le traitement massif et informatisé de l'information souffre alors un peu moins de critiques touchant à la rigueur de cette approche

On peut également chercher des explications non conjoncturelles à ce succès comme la transdisciplinarité d'un domaine qui se fait fort de mettre à profit chacune de ses composantes ainsi que les synergies pouvant exister entre elles. Surtout, l'ECD, comme le rappelle sa définition francophone², est un processus anthropocentré tirant profit d'une interaction entre l'homme et la machine. Placer l'humain au centre du processus relève tout d'abord d'un intérêt pragmatique : les limites actuelles des systèmes automatiques peuvent être repoussées par utilisation de l'intelligence et de l'expertise humaine. On peut également se hasarder à avancer un intérêt "psychologique" : un utilisateur intégré au processus de traitement de l'information, non dépossédé de ses capacités d'analyse tend à mieux accepter, comprendre le processus d'ECD.

La tendance actuelle de la communauté ECD à se pencher sur ses échecs passés (afin de les féconder et de préparer les succès futurs...?) [LMF02] souligne d'ailleurs la grande nécessité d'intégrer l'homme au sein des systèmes et processus d'ECD.

Les travaux présentés dans cette thèse s'inscrivent au sein de la problématique ECD.

Aussi, l'intégration de l'expertise et des connaissances humaines, la définition de méthodes "permettant un échange", une interaction entre homme et machine ainsi que la prise en compte des contraintes d'utilisabilité pour les méthodes développées constituent des exigences fondamentales sous-jacentes aux travaux que nous présentons ici.

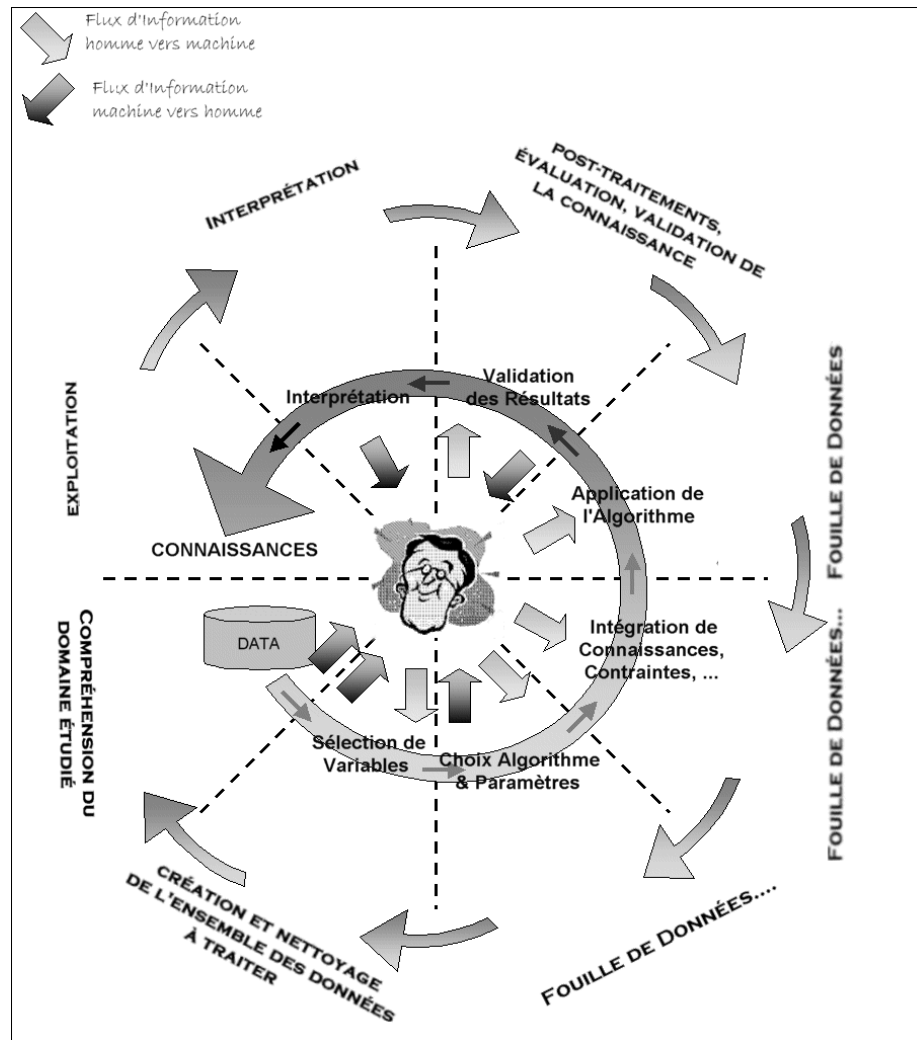
Préalablement évoquée, la relation forte entre recherche et industrie permet et nécessite la prise en compte des besoins issus de la pratique (tels que la limitation des coûts calculatoires, la limitation des coûts de stockage, la conformité à des exigences issues de la distribution de l'information, etc...). Proposer des solutions ne dérogeant pas non plus à ces nécessités constitue la deuxième exigence que nous imposons à nos travaux.

Plus précisément, il s'agit dans ce document de présenter un ensemble de contributions pour l'apprentissage non supervisé. De manière plus détaillée, nous tentons d'apporter ici un ensemble de nouvelles solutions pour l'intégration de l'apprentissage non supervisé au sein d'un processus ECD³.

2. L'ECD est définie par la communauté scientifique francophone comme le processus itératif, anthropocentré (interactif), non trivial d'identification de connaissances valides, nouvelles, potentiellement utiles et intelligibles au sein d'un ensemble de données.

3. Le processus ECD est schématisé sur la figure 1.1. Ce processus comprend classiquement les étapes suivantes :

1. compréhension du domaine étudié
2. création et nettoyage de l'ensemble des données à traiter (Sélection, Construction de Variables....)

FIG. 1.1 –: *Eléments du Processus ECD*

Ces solutions concernent les différentes étapes clés de ce processus, à savoir, la sélection de variables (chapitre 5), l'application de la méthode d'apprentissage non supervisé 3, la validation / l'estimation de la qualité d'un modèle d'apprentissage non supervisé 4. Un chapitre est également consacré à l'agrégation de modèles d'apprentissage non supervisé et aux différents intérêts que

3. extraction des régularités cachées dans les données et formulation des connaissances mises à jour sous forme de modèles ou de règles (cette étape dans le processus global d'ECD est habituellement désignée sous le nom de fouille des données)
4. post-traitements, évaluation, validation de la connaissance découverte
5. interprétation des résultats
6. exploitation des résultats

cela revêt dans le cadre de la prise en compte des deux niveaux d'exigence évoqués plus tôt.

Notons enfin que le choix de l'apprentissage non supervisé s'explique, d'une part, par l'intégration de ces travaux au sein du projet de recherche universitaire BC3⁴ dont l'un des objectifs était la mise au point de méthodes de fusion de données (l'apprentissage non supervisé constituant une piste envisagée), et d'autre part car nous pensons que l'apprentissage non supervisé occupe une place particulière et centrale au sein du processus ECD. Ainsi, la présentation de nos travaux sur l'apprentissage non supervisé est complétée par différentes contributions pour l'apprentissage supervisé exploitant justement nos contributions pour l'apprentissage non supervisé.

Ce document s'organise de la manière suivante : le premier chapitre introduit les éléments nécessaires pour la lecture du reste du document ; les chapitres 3 (présentation d'une nouvelle méthode d'apprentissage supervisé) et 4 (présentation d'une nouvelle méthodologie pour l'évaluation de la validité d'un modèle d'apprentissage non supervisé) peuvent être abordés de manière indépendante des autres alors que la lecture du chapitre 5 (présentation de méthodes de sélection de variables pour l'apprentissage supervisé et non supervisé) et du chapitre 6 (présentation de méthodes d'agrégation de modèles d'apprentissage non supervisé) nécessite respectivement la lecture préalable des chapitres 4 et 3. Notons enfin, que contrairement à la majorité des thèses,

4. Le projet BC3 (Projet Base de Connaissances Cœur-Cerveau, [http : //kbbrain.free.fr](http://kbbrain.free.fr)) initié par différentes équipes universitaires et hospitalières vise à la création d'une base de connaissances (BDC) sur les pathologies organiques touchant deux organes vitaux : le cœur et le cerveau. Une telle BDC a pour objectif premier de servir de support pour le recueil et la conservation de données médicales, permettant ainsi la capitalisation de l'expérience et des connaissances de chercheurs et cliniciens des domaines concernés. La métaphore d'une encyclopédie médicale numérique ayant pour sujet les 2 organes vitaux que constituent le cœur et le cerveau peut dans un premier temps être adoptée pour décrire cet outil. Les travaux que nous menons dans le cadre de ce projet impliquent toutefois d'étendre cette métaphore à celle d'une encyclopédie stockant non seulement de l'information mais se voulant également "génératrice" de connaissances. Nous proposons en effet d'intégrer des méthodes de traitement de l'information à la BDC de manière à permettre l'exploitation de l'information stockée dans cette base et la découverte de nouvelles connaissances.

Equipes Universitaires :

- CREATIS : Centre de Recherche et d'Applications en Traitement de l'Image et du Signal, Université Claude Bernard Lyon 1, Hôpital Cardiologique L.Pradel Service de Radiologie
- ERIC : Equipe de Recherche en Ingénierie des Connaissances, Université Lumière Lyon2
- ISC : Institut des Sciences Cognitives, Université Claude Bernard Lyon1
- LASS : Laboratoire d'Analyse des Systèmes de Santé, Université Claude Bernard Lyon1
- TIMC : Techniques en Imagerie, Modélisation et Cognition, Faculté de Médecine de Grenoble

Equipe hospitalière : Centre de Neuropsychologie de l'hôpital de la Pitié-Salpêtrière (Paris)

le début de ce document n'est pas consacré à un état de l'art général car chacun des chapitres le constituant est amorcé par un état de l'art spécifique.

Nous tenons également à signaler que les différentes expérimentations proposées dans cette dissertation ont été possibles grâce à l'utilisation des logiciels libres : Sipina développé au laboratoire ERIC⁵, WEKA de l'Université de Waikato en Nouvelle-Zélande⁶ et d'un logiciel mis au point au cours de cette thèse.

5. Sipina est disponible au téléchargement à l'adresse <http://eric.univ-lyon2.fr/%7Ericco/sipina.html>

6. WEKA est disponible au téléchargement à l'adresse <http://www.cs.waikato.ac.nz/ml/weka>