

Laboratoire de Recherche en Informatique de Metz

THESE

présentée à

L'UNIVERSITE DE METZ

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE METZ

(mention sciences)

SPECIALITE INFORMATIQUE

par

Franck MARCHETTI

CONTRIBUTION A LA CLASSIFICATION
DE DONNEES BINAIRES ET QUALITATIVES

Soutenue le 15 décembre 1989 devant la commission d'examen

messieurs

G. CELEUX (rapporteur), Chargé de Recherche à l'INRIA

E. DIDAY (rapporteur), Professeur à l'Université de Paris IX

Y. GARDAN (examineur), Professeur à l'Université de Metz

G. GOVAERT (directeur de thèse), Professeur à l'Université de Metz

G. LE CALVE (examineur), Professeur à l'Université de Rennes II

A mes parents

REMERCIEMENTS

Ce travail a été réalisé sous la direction de Monsieur le Professeur G. GOVAERT. Je tiens à lui exprimer toute ma reconnaissance pour son aide, son dynamisme et sa compétence dont il m'a fait profiter tout au long de cette thèse.

Mes remerciements vont également à Monsieur le professeur G. CELEUX qui s'est intéressé de près à ce travail. Je lui exprime toute ma gratitude ainsi qu'à Monsieur le Professeur G. LE CALVÉ pour avoir accepté de rapporter cette thèse, pour les propositions judicieuses qu'ils m'ont suggérées et pour leur participation à la commission d'examen.

Je remercie également Messieurs les Professeurs E. DIDAY et Y. GARDAN pour avoir accepté de participer au jury.

Enfin, j'adresse une pensée particulière à tous les membres du Département Informatique pour leur gentillesse et leur disponibilité à toute épreuve.

TABLE DES MATIERES

INTRODUCTION.....	1
-------------------	---

CHAPITRE 1

ETUDE DE LA DISTANCE L_1 POUR DES

DONNEES BINAIRES ET QUALITATIVES.....	7
---------------------------------------	---

1. Introduction.....	7
----------------------	---

2. La distance en valeurs absolues.....	8
---	---

2.1 Définition.....	8
---------------------	---

2.2 Caractéristiques d'une variable réelle.....	9
---	---

2.2.1 Caractéristique de valeur centrale : la médiane.....	9
---	---

2.2.2 Caractéristique de dispersion : l'écart moyen.....	10
---	----

2.2.3 Exemple.....	11
--------------------	----

2.3 Caractéristiques d'un nuage de points.....	11
--	----

2.3.1 Caractéristique de valeur centrale du nuage le centre médian.....	11
--	----

2.3.2 Caractéristique de dispersion du nuage l'inertie.....	12
--	----

3. Description d'un tableau binaire.....	12
--	----

3.1 Notations.....	13
--------------------	----

3.2 Caractéristiques d'une variable binaire.....	13
--	----

3.2.1 La médiane.....	14
-----------------------	----

3.2.2 L'écart moyen.....	15
--------------------------	----

3.2.3 Exemple d'application.....	16
----------------------------------	----

3.2.4 Propriété.....	16
----------------------	----

3.3 Caractéristiques d'un nuage de points.....	17
--	----

3.3.1 Centre médian du nuage.....	17
-----------------------------------	----

3.3.2 Inertie du nuage.....	17
-----------------------------	----

3.3.3 Exemple d'application.....	17
----------------------------------	----

3.3.4 Propriété.....	18
----------------------	----

3.4 Indépendance vis à vis du codage.....	19
---	----

3.4.1 Caractéristiques d'une variable binaire.....	19
--	----

3.4.2 Caractéristiques d'un nuage de points.....	19
--	----

3.4.3 Généralisation.....	20
---------------------------	----

3.5 Conclusion.....	21
---------------------	----

4. Description d'un tableau de codage additif.....	21
--	----

4.1 Notations.....	22
--------------------	----

4.1.1 Le tableau de modalités.....	22
------------------------------------	----

4.1.2 Le codage binaire additif.....	22
--------------------------------------	----

4.1.3 Le tableau binaire de codage.....	23
---	----

4.2 Propriété.....	24
--------------------	----

4.3	Caractéristiques d'une variable qualitative ordinale	25
4.3.1	Etude de l'échantillon d'une variable	25
4.3.2	Etude de l'échantillon transformé par le codage binaire additif	26
4.3.3	Exemples illustratifs	28
4.4	Caractéristiques d'un nuage de points	29
4.4.1	Propriété et caractéristiques	29
4.4.2	Exemple d'application	30
4.5	Conclusion	30
5.	Description d'un tableau disjonctif complet	31
5.1	Notations	31
5.1.1	Tableau de modalités	31
5.1.2	Le codage disjonctif complet	32
5.1.3	Le tableau de codage	33
5.2	Propriété	34
5.3	Caractéristiques d'une variable qualitative nominale	35
5.3.1	Etude de l'échantillon d'une variable	35
5.3.2	Etude de l'échantillon transformé par le codage disjonctif complet	36
5.4	Caractéristiques d'un nuage de points	37
5.4.1	Cas du nuage associé au tableau de modalités	38
5.4.2	Cas du nuage associé au tableau de codage	38
5.4.3	Exemple d'application	39
5.5	Conclusion	39
CHAPITRE 2		
CLASSIFICATION SUR TABLEAU DE VARIABLES BINAIRES.....		41
1.	Introduction	41
2.	La méthode des Nuées Dynamiques	42
3.	Application au tableau de variables binaires.....	42
3.1	Notations	43
3.2	Le problème	43
3.3	L'algorithme	44
3.4	Expression du critère à la convergence	45
3.5	Autres expressions du critère et problèmes équivalents.....	46
3.6	Indices d'aide à l'interprétation	47
3.7	Exemple simple d'application	48
3.8	Remarques	49
3.9	Classification de données binaires et modèle.....	50
4.	Extension à la famille de distances de Minkowski	50
4.1	La famille de distances de Minkowski	50
4.2	Les problèmes	51
4.3	Première étude	51
4.4	Seconde étude.....	52
5.	Etude comparative.....	52
5.1	Etude comparative des critères.....	53
5.2	Exemple comparatif	54

6. Programme et applications	55
6.1 Présentation du programme	55
6.2 Applications	55
6.3 Applications à des tableaux construits suivant un modèle	63

CHAPITRE 3**CLASSIFICATION SUR TABLEAU DE CODAGE**

BINAIRE ADDITIF	71
------------------------------	-----------

1. Introduction	71
------------------------------	-----------

2. La méthode de classification	72
--	-----------

2.1 Rappel des notations	72
2.2 Le problème	73
2.3 L'algorithme	73
2.4 Interprétation des noyaux	74
2.5 Expression du critère à la convergence	75
2.6 Méthode de classification pour tableau de modalités	75
2.7 Indices d'aide à l'interprétation	76
2.8 Exemple simple d'application	77

3. Application d'une méthode pour variables quantitatives	79
--	-----------

3.1 Application de la méthode au tableau de modalités	79
3.1.1 La méthode	79
3.1.2 La méthode avec contrainte sur les noyaux	80
3.2 Application de la méthode au tableau de codage	80
3.2.1 La méthode	80
3.2.2 Interprétation des noyaux	81
3.2.3 Exemple d'application	83
3.2.4 La méthode avec contrainte de noyaux de modalités	83

4. Etude comparative	84
-----------------------------------	-----------

4.1 Comparaison des deux applications	84
4.2 Comparaison des méthodes	85
4.3 Application illustrative	85

5. Programme et application	87
--	-----------

5.1 Présentation du programme	87
5.2 Application de la méthode	87

CHAPITRE 4**CLASSIFICATION SUR TABLEAU DE CODAGE**

DISJONCTIF COMPLET	99
---------------------------------	-----------

1. Introduction	99
------------------------------	-----------

2. La méthode de classification	100
--	------------

2.1 Rappel des notations	100
2.2 Le problème	101
2.3 L'algorithme	101
2.4 Expression du critère à la convergence	102
2.5 La méthode sur tableau de modalités	103

2.6	Indices d'aides à l'interprétation.....	104
2.7	Exemple simple d'application.....	104
3.	Classification sur les profils des individus.....	106
3.1	Notations et définitions.....	106
3.2	La distance du Khi2.....	107
3.3	La méthode.....	107
3.4	La méthode avec contrainte.....	108
3.4.1	La méthode.....	108
3.4.2	Expression du critère à la convergence.....	110
3.4.3	Interprétation des noyaux.....	110
3.4.4	Version de l'algorithme utilisant le tableau de modalités.....	111
3.4.5	Exemple simple d'application.....	112
4.	Comparaison des méthodes.....	113
4.1	Comparaison des critères.....	113
4.2	Application illustrative.....	113
5.	Programmes et application.....	114
5.1	Programmes.....	114
5.2	Application.....	115
CHAPITRE 5		
INERTIE SUR L'ESPACE BINAIRE ET APPLICATION		
A LA CLASSIFICATION.....125		
1.	Introduction.....	125
2.	Inertie sur l'espace binaire.....	126
2.1	La première approche.....	126
2.2	Extension de cette approche.....	127
2.2.1	La nouvelle pondération.....	127
2.2.2	Centre médian.....	127
2.2.3	Pondération associée au centre médian.....	127
2.2.4	Propriété de conservation du centre médian.....	128
2.3	Inertie sur l'espace binaire.....	129
2.4	Pseudo-théorème de Huyghens.....	130
2.5	Relation de décomposition de l'inertie.....	131
3.	La méthode de classification MNDBIN.....	132
3.1	La méthode MNDBIN.....	132
3.2	La nouvelle approche.....	132
3.3	Généralisation de la méthode.....	133
3.4	Indices de description d'une partition.....	134
3.4.1	Notations.....	134
3.4.2	Définition des indices.....	135
3.4.3	Remarque sur les indices.....	137
3.4.4	Exemple d'application.....	137
3.4.5	Programme.....	138
3.5	Influence de la transformation du tableau initial.....	139
4.	La méthode de classification croisée CROBIN.....	140
4.1	Le principe de la classification croisée (G. Govaert 1983).....	140

4.2	La méthode CROBIN (G. Govaert 1983).....	141
4.3	La mesure d'information	141
4.4	Tableau associé à un couple de partitions.....	142
4.5	La nouvelle approche	144
4.5.1	Les deux algorithmes intermédiaires.....	144
4.5.2	La méthode CROBIN et la mesure d'information.....	145
4.6	L'algorithme généralisé.....	146
5.	Classification ascendante hiérarchique sur données binaires	146
5.1	Notations.....	147
5.2	Les limites de l'analogie.....	147
5.2.1	Indice analogue à l'indice de Ward	147
5.2.2	L'indice de la distance entre centres médians.....	148
5.3	Indice de l'inertie	149
5.4	Un nouvel indice	150
5.5	Programme et applications	151
CHAPITRE 6		
CLASSIFICATION ET ANALYSE EN COMPOSANTES PRINCIPALES		
POUR DONNEES BINAIRES.....		
		157
1.	Introduction	157
2.	Notations	159
2.1	Les données.....	159
2.2	L'espace.....	159
2.3	Vecteurs binaires et opérations.....	159
2.4	Notion de base pour l'espace binaire.....	160
3.	Vecteurs binaires et sous-espaces binaires.....	160
3.1	Vecteur binaire et sous-ensemble associé.....	160
3.2	Axe binaire.....	161
3.3	Axes orthogonaux.....	161
3.4	Système d'axes binaires et sous-espace engendré.....	161
4.	Projection sur un sous-espace binaire.....	162
4.1	Projection sur un axe binaire.....	162
4.2	Image d'un nuage par la projection sur un axe.....	164
4.3	Projection sur un système d'axes.....	165
4.4	Cas d'un système d'axes orthogonaux	166
4.4.1	Propriété	166
4.4.2	Vecteur de pondérations associé au projeté d'un point	167
4.4.3	Projection du nuage des individus	167
4.4.4	Remarques.....	168
4.5	Cas d'un système d'axes quelconques.....	168
4.5.1	Le problème de la projection.....	168
4.5.2	Un algorithme	169
4.5.3	Remarque	170
5.	Inertie par rapport à un sous-espace binaire.....	171
5.1	Inertie d'un nuage par rapport à un axe.....	171
5.1.1	Définition	171
5.1.2	Propriétés	171

5.1.3	Axe d'inertie minimale	172
5.1.4	Sous-tableau homogène	173
5.1.5	Inertie et mesure d'information.....	174
5.2	Inertie d'un nuage par rapport à un système d'axes orthogonaux.....	174
5.2.1	Définition	174
5.2.2	Propriétés	174
5.2.3	Sous-espace d'inertie minimale	176
5.2.4	Sous-tableaux homogènes	176
5.2.5	Inertie et mesure d'information.....	177
5.3	Inertie d'un nuage par rapport à un système d'axes quelconques.....	177
5.3.1	Suppression de la contrainte.....	178
5.3.2	Le problème d'optimisation.....	179
5.3.3	L'algorithme.....	179
5.3.4	Remarques.....	180
6.	Analyse en composantes principales avec contrainte d'axes orthogonaux.....	181
6.1	Analyse en composantes principales et classification	181
6.2	Remarques sur la méthode	182
6.3	Interprétation des résultats.....	183
6.4	Exemple d'application	184
6.5	Influence de la transformation du tableau initial	186
6.6	Les limites de cette approche.....	187
7.	Analyse en composantes principales sans contrainte d'axes orthogonaux.....	188
7.1	La méthode.....	188
7.2	Remarque sur la méthode.....	189
7.3	L'analyse factorielle booléenne	189
7.3.1	Principe de la méthode	189
7.3.2	La méthode	190
7.4	Lien entre les deux méthodes	191
7.5	Un exemple simple d'application.....	192
8.	Une conclusion sur les méthodes pour tableau de variables binaires	194
8.1	Modèle matriciel associé au méthodes	194
8.2	Utilisation des méthodes.....	196
9.	Programmes et Applications.....	197
	CONCLUSION	205
	BIBLIOGRAPHIE	207