

# Classification automatique des données

---

Environnement statistique  
et informatique

G. Celeux  
E. Diday  
G. Govaert  
Y. Lechevallier  
H. Ralambondrainy

**DUNOD**

informatique

BIBLIOTHEQUE DU CERIST



# Classification automatique des données

BIBLIOTHEQUE DU CERIST

# BIBLIOTHEQUE DU CERIST

# Classification automatique des données

c 2293

GILLES CELEUX

*Chargé de recherche à l'INRIA*

EDWIN DIDAY

*Professeur à l'Université Paris IX*

*Chef de projet à l'INRIA*

GÉRARD GOVAERT

*Professeur à l'IUT de Metz*

YVES LECHEVALLIER

*Directeur de recherche à l'INRIA*

HENRI RALAMBONDRAINY

*Chargé de recherche à l'INRIA*

**DUNOD**  
**informatique**

5795

© BORDAS, Paris, 1989

ISBN : 2-04-018798-7

“ Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement de l'auteur, ou de ses ayants-droit, ou ayants-cause, est illicite (loi du 11 mars 1957, alinéa 1<sup>er</sup> de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait une contrefaçon sanctionnée par les articles 425 et suivants du Code pénal. La loi du 11 mars 1957 n'autorise, aux termes des alinéas 2 et 3 de l'article 41, que les copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective d'une part, et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration ”

## REMERCIEMENTS

Tous les chapitres de ce livre sont illustrés par les commandes du logiciel SICLA (système interactif de classification automatique) qui nous a permis d'acquérir l'expérience nécessaire à l'écriture de cet ouvrage. Ce logiciel est l'oeuvre de toute l'équipe de "classification automatique et reconnaissance des formes" de l'INRIA dont font partie tous les auteurs ainsi que Messieurs C. Désarménien, S. Bochi et Madame Y. Ok que nous tenons à remercier ici.

Nous tenons à exprimer notre reconnaissance à la direction de l'INRIA et notamment à ses présidents successifs Messieurs les professeurs J.L. Lions et A. Bensoussan pour l'appui constant qu'ils nous ont apporté tout au long des activités de notre projet.

Nous tenons à remercier vivement Mesdames M. Cornélis et C. Dubois pour tous les efforts qu'elles ont déployés dans la réalisation technique de cet ouvrage.

# BIBLIOTHEQUE DU CERIST

## PREAMBULE

L'analyse des données cherche à extraire d'une grande masse de données multidimensionnelles les "informations utiles". Cette synthèse peut être effectuée à l'aide de méthodes de visualisation (analyse factorielle) et de méthodes de structuration (classification).

Les méthodes utilisées se sont beaucoup développées ces dernières années et l'informatique y a pris une place très importante. De plus en plus, ce sont des logiciels assez complets, regroupant de nombreuses méthodes, qui sont proposés. Ces logiciels doivent être de véritables boîtes à outils permettant la prise en compte de nombreux types de données, l'enchaînement des différentes méthodes, le traitement conversationnel...

L'utilisation de ces logiciels nécessite une bonne connaissance des méthodes proposées et l'objectif de ce livre est de fournir les bases théoriques nécessaires pour maîtriser ces nouveaux moyens. Nous nous sommes appuyés sur notre expérience de développement et d'utilisation de SICLA, logiciel orienté surtout vers la classification automatique, mais nous ne nous sommes pas limités aux seules méthodes de ce logiciel.

Le premier chapitre est consacré aux étapes préliminaires à une analyse de données. Il s'agit essentiellement de la définition des individus et des variables, de la construction d'un tableau de données et de la description élémentaire de ces données. Nous avons ensuite consacré un paragraphe à la notion d'inertie qui sera très utilisée tout au long de ce livre. Fréquemment les méthodes d'analyse des données ne peuvent pas être utilisées directement sur les données initiales ; les conditions d'application de ces méthodes peuvent nécessiter un recodage des données ou le choix de fonctions de similarité. Les deux derniers paragraphes sont consacrés à ces aspects.

Le second chapitre est consacré à la classification automatique et constitue une part importante de ce livre. Après avoir décrit les principales structures de classification utilisées, le principe général de la

méthode des nuées dynamiques (M.N.D) et ses propriétés sont étudiées. Nous analysons ensuite quelques méthodes parmi les plus utilisées, comme la classification adaptative, la classification croisée, la classification d'un tableau de proximités, qui suivent toutes, de façon de plus ou moins proche, le schéma de la M.N.D. Le paragraphe suivant est consacré à la comparaison des problèmes de la recherche des composantes d'un mélange de distributions de probabilité et de la classification. Toutes ces méthodes de classification ne garantissent pas le meilleur résultat, pourtant la classification optimale, que nous étudions ensuite, permet d'atteindre cet objectif lorsque les données sont ordonnées en utilisant le principe de programmation dynamique. Les deux paragraphes suivants sont consacrés aux outils permettant d'interpréter et d'utiliser les résultats fournis par les méthodes de partitionnement. Enfin, nous terminons par l'étude des méthodes hiérarchiques et pyramidales.

Dans le troisième chapitre, nous présentons les méthodes d'analyse factorielle. Ces méthodes constituent un domaine important de l'analyse des données et relèvent de l'analyse linéaire. Elles reposent toutes sur les mêmes bases mathématiques mais ont des domaines d'application différents. Quatre méthodes importantes sont décrites : l'analyse en composantes principales (ACP) pour les tableaux de variables quantitatives, l'analyse factorielle d'un tableau de distances (AFTD), l'analyse factorielle des correspondances (AFC) pour les tableaux de contingence et les tableaux possédant des propriétés similaires, l'analyse factorielle des correspondances multiples (AFCM), généralisation de l'AFC pour l'étude des questionnaires.

Le dernier chapitre est consacré à la présentation de techniques classiques de discrimination : la discrimination linéaire et quadratique, la méthode des k-plus proches voisins, la discrimination par arbre de décision ou segmentation. Dans ce chapitre, nous privilégions l'aspect prévisionnel de la discrimination. Nous optons ainsi pour une présentation bayésienne de ce problème de décision statistique. De même, nous étudions les importants problèmes de sélection de variables et d'évaluation de la qualité de la prévision à l'aide des techniques de rééchantillonnage.

Ce livre ayant pour arrière-plan le logiciel SICLA, le lecteur trouvera, pour chaque domaine étudié, une description succincte des possibilités de traitements offertes par ce logiciel.

# TABLE DES MATIERES

1. LES PREMIERES ETAPES D'UNE ANALYSE DE DONNEES .....	1
<b>1.1 Introduction</b> .....	1
<b>1.2 Les principaux choix de base</b> .....	1
1.2.1 Choix de l'ensemble des individus .....	1
1.2.2 Définition et choix de l'ensemble des variables.....	2
1.2.3 Les différents types de variables.....	3
1.2.4 Le choix du codage .....	4
1.2.5 Choix d'une similarité ou d'une dissimilarité .....	5
<b>1.3 Construction d'un tableau de données</b> .....	6
1.3.1 Définition et notations .....	6
1.3.2 Exemples de tableaux de données .....	6
1.3.3 Les commandes de SICLA pour la saisie des données ...	14
<b>1.4 Description élémentaire d'un tableau de données</b> .....	15
1.4.1 Description élémentaire de variables quantitatives.....	15
1.4.2 Description élémentaire de variables qualitatives.....	20
1.4.3 Les commandes de SICLA pour la description élémentaire .....	21
<b>1.5 Notion d'inertie</b> .....	27
1.5.1 Nuage de points pondérés.....	27
1.5.2 Inertie du nuage par rapport à un point .....	27
1.5.3 Théorème de Huygens .....	28
1.5.4 Inerties associées à une partition.....	28
1.5.5 Tableaux de contingence .....	33
<b>1.6 Changement de variable et codage</b> .....	38
1.6.1 Intérêt du changement de variable .....	38
1.6.2 Formalisation de la notion de changement de variable ..	39
1.6.3 Différents types de changement de variable.....	40
1.6.4 Les commandes de SICLA pour le changement de variable et le codage .....	47
<b>1.7 Similarités et dissimilarités</b> .....	49
1.7.1 Quelques définitions.....	49
1.7.2 Tableaux de variables quantitatives.....	50
1.7.3 Tableaux binaires.....	52
1.7.4 Tableaux de variables qualitatives .....	53

1.7.5	Dissimilarités entre groupes d'individus.....	54
1.7.6	Les commandes de SICLA pour la création de tableaux de distances.....	57
2.	<b>CLASSIFICATION AUTOMATIQUE.....</b>	<b>59</b>
2.1	<b>Introduction.....</b>	<b>59</b>
2.2	<b>Les principaux espaces de classification .....</b>	<b>61</b>
2.2.1	Introduction.....	61
2.2.2	Définition des principaux espaces de classification.....	61
2.2.3	Formalisation de la notion d'espace de classification ...	65
2.3	<b>La méthode des nuées dynamiques.....</b>	<b>67</b>
2.3.1	Introduction.....	67
2.3.2	Un exemple simple.....	68
2.3.3	Les principaux aspects de la méthode .....	70
2.3.4	Tableau des principaux modes de représentation utilisés .....	72
2.3.5	Les notions de base de la méthode des nuées dynamiques .....	72
2.3.6	Etude d'une famille importante.....	73
2.3.7	Initialisation et nombre de classes.....	77
2.4	<b>La méthode des nuées dynamiques dans le cas des centres de gravité.....</b>	<b>78</b>
2.4.1	Introduction.....	78
2.4.2	Les principales étapes .....	78
2.4.3	Expressions du critère .....	80
2.4.4	Les centres mobiles dans SICLA .....	86
2.4.5	Méthodes voisines .....	86
2.5	<b>Classification avec distances adaptatives.....</b>	<b>87</b>
2.5.1	Introduction.....	87
2.5.2	Une distance adaptative unique.....	88
2.5.3	Une distance adaptative par classe.....	92
2.5.4	Considérations pratiques sur les deux méthodes.....	96
2.5.5	Les distances adaptatives dans SICLA .....	97
2.6	<b>La classification croisée.....</b>	<b>97</b>
2.6.1	Introduction.....	97
2.6.2	Un exemple.....	98
2.6.3	Principe général .....	100
2.6.4	Tableaux de contingence .....	101
2.6.5	Tableaux de variables qualitatives .....	108
2.6.6	Tableaux de variables quantitatives.....	109

2.6.7	Tableaux binaires.....	117
2.6.8	La classification croisée dans SICLA .....	124
<b>2.7</b>	<b>Classification d'un tableau de proximités .....</b>	<b>127</b>
2.7.1	Introduction.....	127
2.7.2	Méthode de classification d'un tableau de proximités... ..	128
2.7.3	Méthode optimisant un critère indépendant du nombre de classes.....	131
2.7.4	La classification sur tableaux de proximités dans SICLA.....	136
<b>2.8</b>	<b>Mélange de distributions de probabilité et classification .....</b>	<b>137</b>
2.8.1	Introduction.....	137
2.8.2	L'approche classification .....	138
2.8.3	L'approche estimation .....	143
2.8.4	Les mélanges dans SICLA .....	147
<b>2.9</b>	<b>Classification optimale.....</b>	<b>148</b>
2.9.1	Introduction.....	148
2.9.2	La programmation dynamique.....	150
2.9.3	Partitionnement sous contrainte d'ordre total.....	151
2.9.4	Applications.....	153
2.9.5	La classification optimale dans SICLA .....	154
<b>2.10</b>	<b>Interprétation d'une partition .....</b>	<b>154</b>
2.10.1	Notations et définitions .....	155
2.10.2	Indice général.....	155
2.10.3	Contributions des variables.....	156
2.10.4	Description des classes .....	157
2.10.5	Description des classes par variable. ....	157
2.10.6	Indices complémentaires .....	158
2.10.7	Cas particulier des variables qualitatives.....	159
2.10.8	L'interprétation d'une partition dans SICLA.....	160
<b>2.11</b>	<b>Analyse d'une multi-partition.....</b>	<b>164</b>
2.11.1	Définition d'une multi-partition .....	164
2.11.2	Représentation d'une multi-partition.....	164
2.11.3	Les formes fortes .....	165
2.11.4	L'algorithme des connexités descendantes.....	166
2.11.5	L'arbre de longueur minimum sur les formes fortes .	167
2.11.6	Partition centrale associée à une multi-partition.....	167
2.11.7	L'analyse d'une multi-partition dans SICLA .....	171
<b>2.12</b>	<b>La classification hiérarchique .....</b>	<b>171</b>
2.12.1	Introduction .....	171
2.12.2	Définition d'une hiérarchie indicée.....	173
2.12.3	Indice d'agrégation entre groupes d'individus .....	173

2.12.4	Construction de hiérarchies indicées .....	174
2.12.5	La formule de récurrence de Lance et Williams.....	178
2.12.6	Hiérarchies indicées et ultramétriques .....	178
2.12.7	L'ultramétrie sous-dominante.....	180
2.12.8	Ultramétriques sur-dominantes .....	183
2.12.9	Hiérarchies basées sur l'inertie .....	185
2.12.10	Partitions ou hiérarchies ? .....	185
2.12.11	Les hiérarchies dans SICLA.....	187
<b>2.13</b>	<b>Les pyramides.....</b>	<b>187</b>
2.13.1	Introduction .....	187
2.13.2	Des ultramétriques aux dissimilarités pyramidales ..	188
2.13.3	Quelques propriétés des pyramides .....	190
2.13.4	Construction d'une pyramide .....	191
2.13.5	Hiérarchies et pyramides .....	192
2.13.6	Les Pyramides dans SICLA.....	193
<b>3.</b>	<b>L'ANALYSE FACTORIELLE.....</b>	<b>195</b>
<b>3.1</b>	<b>Introduction.....</b>	<b>195</b>
<b>3.2</b>	<b>L'analyse en composantes principales (ACP).....</b>	<b>196</b>
3.2.1	Domaine d'application.....	196
3.2.2	But et cadre de l'ACP.....	196
3.2.3	Inerties .....	199
3.2.4	Formulation du problème de l'ACP .....	201
3.2.5	Résolution du problème .....	203
3.2.6	Facteurs associés aux axes factoriels.....	206
3.2.7	Composantes principales .....	206
3.2.8	Représentation des individus.....	208
3.2.9	Représentation des variables .....	209
3.2.10	Choix de la métrique M.....	210
3.2.11	Les éléments illustratifs.....	211
<b>3.3</b>	<b>L'analyse factorielle sur tableau de distances (AFTD) .....</b>	<b>213</b>
3.3.1	But de l'AFTD.....	213
3.3.2	Théorème fondamental .....	214
3.3.3	Résolution du problème .....	215
<b>3.4</b>	<b>L'analyse factorielle des correspondances (AFC) .....</b>	<b>217</b>
3.4.1	Domaine d'application.....	217
3.4.2	Notations, définitions, résultats préliminaires .....	218
3.4.3	Résolution du problème de l'AFC.....	219
3.4.4	Les formules de transition .....	223
3.4.5	Représentation en AFC .....	224

<b>3.5 L'analyse canonique (AC)</b> .....	225
3.5.1 Cadre et but de l'AC .....	225
3.5.2 Formulation du problème.....	226
3.5.3 Résolution du problème .....	226
3.5.4 Les facteurs canoniques.....	228
3.5.5 Propriétés et représentations.....	228
3.5.6 Analyse canonique généralisée .....	229
3.5.7 L'analyse des correspondances comme analyse canonique .....	230
<b>3.6 L'analyse des correspondances multiple (ACM)</b> .....	231
3.6.1 Domaine d'application .....	231
3.6.2 Notations et définitions .....	231
3.6.3 L'analyse dans le cas de deux questions.....	232
3.6.4 L'analyse dans le cas général.....	233
3.6.5 Propriétés de l'ACM.....	234
<b>3.7 L'analyse factorielle à partir de SICLA</b> .....	235
<b>4. L'ANALYSE DISCRIMINANTE</b> .....	237
<b>4.1 Le cadre de l'analyse discriminante</b> .....	237
4.1.1 Exemples .....	237
4.1.2 Description du chapitre.....	238
<b>4.2 Le modèle de base</b> .....	239
4.2.1 Notations et définitions .....	239
4.2.2 L'approche bayésienne .....	240
4.2.3 Recherche de la règle de décision de Bayes.....	241
4.2.4 Etude du cas gaussien.....	244
4.2.5 Pratique de l'approche bayésienne .....	245
<b>4.3 Méthodes métriques</b> .....	247
4.3.1 L'analyse discriminante linéaire .....	247
4.3.2 Discrimination quadratique .....	249
<b>4.4 L'analyse factorielle discriminante</b> .....	250
4.4.1 Le problème.....	251
4.4.2 Propriétés .....	252
4.4.3 Représentations factorielles .....	253
<b>4.5 Sélection des variables, méthodes pas à pas</b> .....	255
4.5.1 Critères de sélection des variables.....	255
4.5.2 Procédures de sélection.....	256
4.5.3 Le choix du nombre optimal de variables.....	258
<b>4.6 Mesure de la qualité de la discrimination</b> .....	259
<b>4.7 Analyse discriminante sur variables         qualitatives</b> .....	261

<b>4.8 La méthode des K plus proches voisins .....</b>	<b>262</b>
<b>4.9 Une méthode de segmentation non paramétrique .....</b>	<b>264</b>
4.9.1 Introduction .....	264
4.9.2 La méthode de base (cas de 2 classes).....	265
4.9.3 Une méthode dans le cas multi-classe.....	269
4.9.4 Arbres de décision ternaires .....	271
<b>4.10 La discrimination dans SICLA.....</b>	<b>274</b>
4.10.1 La commande DISC.....	274
4.10.2 La commande FACDSC .....	274
4.10.3 La commande SELDSC.....	275
4.10.4 La commande VOISIN.....	275
4.10.5 La commande DNP .....	275