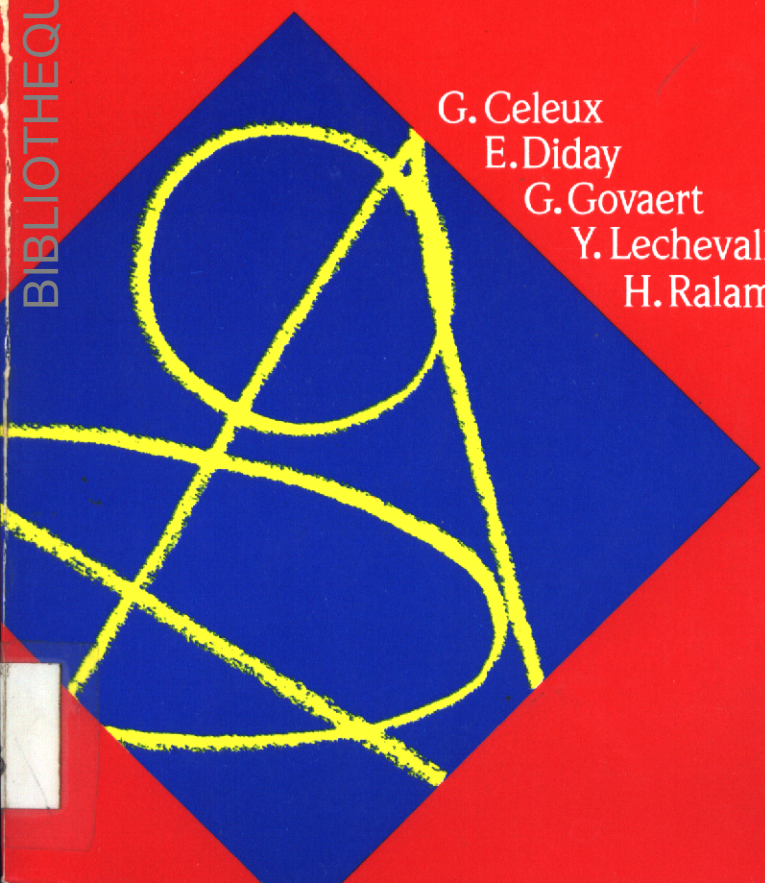
Classification automatique des données

Environnement statistique et informatique



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

DUNOD
informatique

Classification automatique des données

Classification automatique des données

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

ISBN : 2-04-018798-7

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.

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