

Automatische Klassifikation

Theoretische und praktische Methoden
zur Gruppierung und Strukturierung von Daten
(Cluster-Analyse)

Von

Dr. rer. nat. Hans Hermann Bock

Technische Universität Hannover

Mit 54 Abbildungen

Technische Hochschule Darmstadt FACHBEREICH INFORMATIK B I B L I O T H E K Inventar-Nr.: <u>2438</u> Sachgebiete: _____ Standort: _____
--



VANDENHOECK & RUPRECHT IN GÖTTINGEN

Inhalt

§ 1.	Problemstellung, Grundlagen	13
	a) Einführung	13
	b) Einige Anwendungsbeispiele	14
	c) Die Datenmatrix; quantitative und qualitative Merkmale	18
	d) Ausgangsdaten: Ähnlichkeiten, Relationen	20
	e) Vorläufige Präzisierung des Klassifikationsproblems; Bemerkungen	21
Teil I: Ähnlichkeits-, Distanz- und Homogenitätsmaße		
Kap. 1.	Ähnlichkeit und Distanz von Objekten	24
§ 2.	Ähnlichkeits- und Distanzmaße; Präordnungen	24
	a) Definition und Eigenschaften	24
	b) Kriterien zur Auswahl solcher Maße	26
	c) Die induzierte Präordnung	29
	d) Vergleich von Distanz- und Ähnlichkeitsmaßen	32
§ 3.	Ähnlichkeit und Distanz bei quantitativen Merkmalen	35
	a) Der euklidische Abstand	35
	b) L_p -Distanzen	39
	c) MAHALANOBIS-Distanz	40
	d) Der Korrelationskoeffizient als Ähnlichkeitsmaß	44
	e) Ein Distanzmaß von KENDALL	46
	f) Einige empirische Distanzmaße	47
§ 4.	Ähnlichkeit bei binären Merkmalen	48
	a) Binäre Merkmale und zugehörige Ähnlichkeitsmaße	48
	b) Symmetrische Merkmale und invariante Ähnlichkeitsmaße	50
	c) Einige vertauschungs-invariante Ähnlichkeitsmaße	51
	d) Einige nicht invariante Ähnlichkeitsmaße	53
	*e) Vergleich einiger Ähnlichkeitsmaße	55
	f) Assoziationsmaße	59
	g) Ein probabilistisches Ähnlichkeitsmaß	63
	h) Berücksichtigung von Merkmalshäufigkeit und Abhängigkeiten	65
§ 5.	Ähnlichkeit bei mehrstufigen Merkmalen	66
	a) Geordnete und ungeordnete Alternativen	67
	b) Verallgemeinerter M-Koeffizient und Modifikationen	68
	c) Verwendung von Merkmalshäufigkeit, Alternativenanzahl u. ä.	69
	*d) Ähnlichkeit bei geordneten Alternativen	71
§ 6.	Einzelfragen; verwandte Probleme	74
	a) Gemischte Merkmale	74
	b) Fehlende Daten	75
	c) Gewichtung von Merkmalen	76
	d) Transformation von Ähnlichkeiten in Distanzen und umgekehrt	77
	e) Mehrdimensionale Skalierung	79

Kap. 2:	Ähnlichkeit, Distanz und Homogenität bei Objektmengen	81
§ 7.	Ähnlichkeit und Distanz von Objektmengen	81
	a) Definition anhand einer Ähnlichkeits- oder Distanzmatrix	81
	b) Quantitative Merkmale	83
	c) Allgemeine Methoden zur Definition von Distanzmaßen	88
	d) Qualitative Merkmale	90
§ 8.	Homogenität und Inhomogenität einer Objektmenge	91
	a) Definition durch Ähnlichkeiten oder Distanzen	91
	b) Quantitative Daten	92
	c) Informationstheoretische Heterogenitätsmaße bei qualitativen Daten	93
	d) Weitere Heterogenitätsmaße	99
§ 9.*	Typische Objekte einer Menge	100
	a) Zentrale Punkte	100
	b) Kernpunkte	101
	c) Definition mit Hilfe von Heterogenitätsmaßen	102
	d) Verwendung von Hauptkomponenten	103

Teil II: Disjunkte Gruppierung

§ 10.	Disjunkte Gruppierung	104
	a) Problemstellung und Überblick	104
	b) Möglichkeiten zur Beschreibung einer Partition	107
	*c) Die Anzahl der Partitionen einer Menge	109
Kap. 3:	Entscheidungstheoretische Modelle bei Normalverteilungen	113
§ 11.*	Modell I: Kovarianzmatrizen bis auf einen Faktor bekannt	113
	a) Allgemeine Voraussetzungen	113
	b) Fall A: Gleiche Kovarianzmatrizen; ML-Verfahren	114
	c) Bayesverfahren für Fall A: Einfache Verlustfunktion	116
	d) Bayesverfahren für Fall A: Quadratische Verlustfunktion	122
	e) Fall B: Kovarianzmatrix $\sigma_i^2 \cdot \Sigma$ in Klasse A_i ; ML-Verfahren	128
	f) Bayesverfahren für Fall B: Einfache Verlustfunktion	130
§ 12.*	Modell II: Unbekannte Kovarianzmatrizen	137
	a) Fall C: Gleiche Kovarianzmatrizen; ML-Verfahren	137
	b) Bayesverfahren für Fall C: Einfache Verlustfunktion	138
	c) Fall D: Verschiedene Kovarianzmatrizen; ML-Verfahren	142
	d) Bayesverfahren für Fall D: Einfache Verlustfunktion	143
§ 13.*	Der Fall unbekannter Klassenanzahl	146
	a) Ein simultanes Test- und Klassifikationsproblem	147
	b) Signifikanztests bei allgemeineren Voraussetzungen	153
Kap. 4:	Optimale, disjunkte Gruppierungen	159
§ 14.	Optimale, disjunkte Gruppierungen: Problemstellung	159

§ 15.	Varianzkriterium und zugehörige Verfahren	162
	a) Definition und allgemeine Eigenschaften	162
	b) Die Äquivalenz zweier Extremalprobleme.	164
	c) Notwendige Bedingungen für die optimale Gruppierung der Objekte	169
	d) Iterierte Minimal-Distanz-Partitionen.	171
	e) Gruppierung in $m = 2$ Klassen; hierarchische Verfahren.	174
	f) Der eindimensionale Fall	177
	*g) Lineare und nichtlineare Optimierung.	180
§ 16.	Verwandte Optimalitätskriterien.	184
	a) Determinantenkriterium.	185
	b) Notwendige Bedingungen für die optimale Partition	187
	c) Das Spur-Kriterium	191
	d) Verwandte Kriterien	192
	e) Der Fall unbekannter-Klassenanzahl.	193
	f) Gleichzeitige Gruppierung von Objekten und Merkmalen	194
§ 17.	Charakterisierung der Klassen durch Hyperebenen.	195
	a) Ein kontinuierliches Extremalproblem.	195
	b) Ein äquivalentes Problem	198
	c) Ein entsprechendes Gruppierungsverfahren	200
§ 18.	Optimalitätskriterien: Beliebige Ähnlichkeits- oder Distanzmaße	202
	a) Mittlere Heterogenität der Klassen.	202
	b) Separation der Klassen.	203
	c) Entsprechende Kriterien für Ähnlichkeitsmaße.	204
	d) Kriterien, die einen variablen Parameter enthalten	205
§ 19.	Ein Kriterium für qualitative, ungeordnete Merkmale	206
§ 20.	Ein Kriterium für dichotome Ähnlichkeitsmaße	210
	a) Dichotome Ähnlichkeitsmaße	210
	b) Ein zugehöriges Optimalitätskriterium.	211
	c) Ein asymptotisches Ergebnis	211
§ 21.	Kriterien, die durch eine Präordnung definiert werden	212
	a) Die Kriterien von BENZÉCRI und de la VEGA	212
	b) Eigenschaften und Vergleich dieser Kriterien.	214
	*c) Asymptotische Verteilung der Kriterien.	216
Kap. 5:	Numerische Verfahren.	218
§ 22.	Iterative Verbesserung einer Anfangsklassifikation	219
	a) Austauschverfahren	220
	b) Iterierte Minimal-Distanz-Partitionen.	222
	c) Wahl der Anfangsklassifikation	223
§ 23.	Rekursiver Aufbau von Gruppen um Kerne	224
	a) Konstruktionsprinzip	225
	b) Präzisierung der einzelnen Schritte.	226
	c) Wahl der Parameter.	228
	d) Einige spezielle Verfahren.	229

§ 24.	Heuristische und kombinierte Verfahren	232
	a) Modifikation der Verfahren aus § 22	232
	b) Ein sequentielles, heuristisches Verfahren.	235
	c) Minimal-Distanz-Partitionen bei nicht exhaustiver Gruppierung	236
§ 25.	Projektionsmethoden	237
	a) Hauptkomponentenmethode.	237
	b) Projektion auf die Ebene der Klassenmittelpunkte (Modell B)	242
	c) Analoge Verfahren bei beliebiger Kovarianzmatrix	245
	*d) Faktoranalyse; nichtlineare Reduktionsmethoden	247
Kap. 6:	Analyse von Punkt- und Verteilungsdichte	249
§ 26.	Verteilungsmischungen	250
	a) Verteilungsmischungen	250
	b) Mischung von Normalverteilungen: Identifizierbarkeit	252
	c) Mischung von Normalverteilungen: Schätzung der Parameter	257
	d) Andere Schätzmethoden; Mischungen nicht normaler Verteilungen	262
§ 27.*	Nichtparametrische Schätzung einer Verteilungsdichte	263
	a) Schätzung mit Hilfe von Kernfunktionen	265
	b) Verwendung von Reihentwicklungen; das Histogramm.	268
	c) Verwendung von Ranggrößen	269
§ 28.	Gruppierung unter Verwendung der Punktdichte	270
	a) Unimodale und multimodale Verteilungsdichten.	270
	b) Gruppen relativ hoher Punktdichte; das Verfahren von SCHNELL	272
	c) Gruppen absolut hoher Punktdichte.	276
	d) Das Verfahren von WISHART	278
§ 29.	Sequentielle, selbst-adaptierende Verfahren	279
	a) Problemstellung	279
	b) Ein kontinuierliches Extremalproblem	281
	c) Stochastische Approximation	284
	d) Kontinuierliches Varianzkriterium; das Verfahren von BRAVERMAN	287
	e) Das Verfahren von MACQUEEN.	291
	*f) Die Methode der Potentialfunktion	293
	g) Das Verfahren von DOROFYUK; beliebige Merkmale	297
Kap. 7:	Graphentheoretische Methoden	298
§ 30.	Gruppierung durch Zusammenhangskomponenten	298
	a) Gruppen der Stufe d	298
	b) Graphentheoretische Interpretation	300
	c) Gruppierung mit Hilfe des Minimalbaumes.	303
	d) Bemerkungen und Modifikationen	305
§ 31.	Verwandte Methoden	307
	a) Gruppierung durch reziproke Paare	308
	b) K -Gruppen	312

Teil III: Nichtdisjunkte und hierarchische Gruppierung

Kap. 8:	Nichtdisjunkte Gruppierung	316
§ 32.	Maximale Cliquen	318
	a) Definition maximaler Cliquen	318
	b) Elimination unwesentlicher Gruppen.	320
	c) Maximale Cliquen und einklassige Objekte	322
	d) Das Konstruktionsverfahren von HARARY/ROSS.	324
	e) Bemerkungen und Modifikationen	329
§ 33.	Iterative, heuristische Verfahren	332
	a) Das Verfahren I von DATTOLA	332
	b) Gruppierung mit kontrollierbarer Überschneidung.	336
§ 34.	R-Gruppen und verwandte Gruppenarten	338
	a) R-Gruppen	339
	b) Konstruktion von R-Gruppen.	341
	*c) Einige Sätze über R-Gruppen	344
	d) S-Gruppen und S*-Gruppen	347
§ 35.	GR-Gruppen	349
	a) Definition und Eigenschaften von GR-Gruppen; Modifikationen	349
	b) Konstruktion von GR-Gruppen.	351
	*c) Beweis von Satz 35.3	353
Kap. 9:	Hierarchische Gruppierung	356
§ 36.	Einführung	356
§ 37.	Formale Beschreibung einer Hierarchie; Definitionen	359
	a) Hierarchien.	360
	b) Indizierte Hierarchien (Dendrogramme)	361
	c) Partitionenhierarchie	363
	d) Die zugeordnete Ultrametrik	365
	*e) Die Anzahl der Hierarchien einer Objektmenge	368
§ 38.*	Optimale Hierarchien	370
	a) Das Gruppierungsproblem; Überblick	370
	b) Die maximale, dominierte Ultrametrik δ^-	371
	c) Minimale, dominierende Ultrametrien δ^+	377
	d) Die Hierarchie von APRESJAN	379
	e) Optimalitätskriterien	382
§ 39.	Agglomerative Verfahren I.	383
	a) Konstruktionsprinzip	384
	b) Single-Linkage-Methode.	387
	c) Complete-Linkage-Methode	392
§ 40.	Agglomerative Verfahren II	399
	a) Zentroid-Methode	400
	b) Average-Linkage-Methode	402

	c) Verfahren mit rekursiver Distanzdefinition	404
	d) Verwendung von Heterogenitätsmaßen	407
	e) Disjunkte Gruppierung mit hierarchischen Verfahren	409
§ 41.	Divisive Verfahren	411
	a) Konstruktionsprinzip	411
	b) Polythetische Methoden	412
	c) Erstellung eines Dendrogramms.	414
	d) Monothetische Verfahren	416
	e) Bemerkungen.	418
§ 42.	Andere hierarchische Methoden	419
	a) Das Verfahren von SCHNELL	420
	b) Das hierarchische Verfahren von WISHART	420
	c) Nichtdisjunkte Hierarchien.	423
	d) k -Ultrametrien, k -Dendrogramme	427
	Anhang: Mathematische Bezeichnungen.	430
	Literaturverzeichnis	435
	Autorenverzeichnis.	468
	Stichwortverzeichnis.	473