

Detlef Steinhausen · Klaus Langer

# Clusteranalyse

Einführung in Methoden und Verfahren  
der automatischen Klassifikation

Mit zahlreichen Algorithmen, FORTRAN-Programmen,  
Anwendungsbeispielen und einer Kurzdarstellung  
der multivariaten statistischen Verfahren



Walter de Gruyter · Berlin · New York 1977

# Inhalt

1. Einleitung . . . . .	11
1.1 Problemstellung . . . . .	11
1.2 Zum Begriff „Clusteranalyse“ . . . . .	13
1.3 Ziel und Funktion . . . . .	14
1.4 Das Clusteranalyseproblem . . . . .	16
1.5 Ablaufschema . . . . .	19
2. Grundzüge multivariater Verfahren . . . . .	25
2.1 Vorbemerkung . . . . .	25
2.2 Allgemeine Voraussetzungen . . . . .	26
2.2.1 Grundbegriffe und Bezeichnungen . . . . .	26
2.2.2 Skalierung einer Variablen . . . . .	28
2.3 Regressionsanalyse . . . . .	30
2.4 Varianz- und Kovarianzanalyse . . . . .	32
2.5 Kanonische Analyse . . . . .	37
2.6 Diskriminanzanalyse . . . . .	39
2.7 Faktoren- und Hauptkomponentenanalyse . . . . .	42
2.8 Multidimensionale Skalierung . . . . .	46
2.9 Zusammenfassung . . . . .	47
2.10 Übungen und Ergänzungen . . . . .	49
3. Ähnlichkeits- und Distanzfunktionen . . . . .	51
3.1 Definition einer Ähnlichkeits- und Distanzfunktion . . . . .	51
3.2 Ähnlichkeits- und Distanzfunktionen bei qualitativen Variablen . . . . .	53
3.2.1 Nominale Variablen . . . . .	53
3.2.2 Ordinale Variablen . . . . .	56
3.3 Ähnlichkeits- und Distanzfunktionen bei quantitativen Variablen . . . . .	58
3.3.1 Euklidische Distanz . . . . .	58
3.3.2 Mahalanobis-Distanz . . . . .	59
3.3.3 $L_r$ -Distanzen . . . . .	61
3.3.4 Q-Korrelationskoeffizient . . . . .	62
3.4 Ähnlichkeits- und Distanzfunktionen bei gemischten Variablen . . . . .	63
3.5 Ähnlichkeits- und Distanzfunktionen bei Elementgruppen . . . . .	64
3.6 Übungen und Ergänzungen . . . . .	66
4. Clusteranalysealgorithmen . . . . .	69
4.1 Vorbemerkung . . . . .	69
4.1.1 Kriterien zur Systematisierung . . . . .	69
4.1.2 Datenstruktur und Gruppierung . . . . .	70

4.1.3 Programmstandards . . . . .	71
4.2 Hierarchische Verfahren . . . . .	73
4.2.1 Agglomerative Verfahren . . . . .	75
4.2.2 Ein graphentheoretisches Verfahren . . . . .	94
4.2.3 Divisive Verfahren . . . . .	98
4.3 Verfahren zur Verbesserung einer Anfangspartition . . . . .	100
4.3.1 Zielfunktionen . . . . .	100
4.3.1.1 Varianzkriterium . . . . .	101
4.3.1.2 Determinantenkriterium . . . . .	103
4.3.1.3 Spur( $W^{-1}B$ )-Kriterium . . . . .	104
4.3.1.4 Varianzkriterium bei transformierten Daten . . . . .	105
4.3.1.5 Zielfunktion für die $L_r$ -Clustering . . . . .	106
4.3.2 Sift-and-Shift Verfahren . . . . .	106
4.3.2.1 Iteriertes Minimaldistanzverfahren . . . . .	107
4.3.2.2 Austauschverfahren . . . . .	118
4.3.2.3 Minimaldistanzverfahren und Austauschverfahren für andere Zielfunktionen . . . . .	127
4.3.2.4 Austauschverfahren für beliebige Distanzmatrizen . . . . .	135
4.3.2.5 Anfangspartitionen . . . . .	137
4.3.2.6 Überwindung lokaler Extrema . . . . .	138
4.4 Andere Verfahren . . . . .	138
4.4.1 Q-Analyse . . . . .	138
4.4.2 Konfigurationsfrequenzanalyse . . . . .	148
4.4.3 Clustering unter Verwendung der Punktdichte . . . . .	156
4.5 Übungen und Ergänzungen . . . . .	158
5. Spezielle Probleme . . . . .	161
5.1 Clusteranalyse bei Variablen . . . . .	161
5.2 Probleme der Beurteilung von Cluster-Lösungen . . . . .	169
5.2.1 Beurteilungskriterien . . . . .	169
5.2.2 Bestimmung der Clusteranzahl . . . . .	170
5.2.3 Vergleich mehrerer Lösungen . . . . .	172
5.3 Probleme der praktischen Durchführung . . . . .	175
5.3.1 Große Elementanzahl . . . . .	175
5.3.2 Große Variablenanzahl . . . . .	176
5.3.3 Fehlende Daten . . . . .	176
6. Zusammenfassender Überblick . . . . .	179
7. Anhang: Grundbegriffe aus der Mengenlehre und Linearen Algebra . . . . .	185
7.1 Grundbegriffe aus der Mengenlehre . . . . .	185
7.2 Grundbegriffe aus der Linearen Algebra . . . . .	187
Literatur . . . . .	197
Autoren- und Sachregister . . . . .	201