

Clusteranalyse mit Mikrocomputern

von Hans-Joachim Mucha



Akademie Verlag

INHALT

EINLEITUNG	11
1. MULTIVARIATE STATISTIK UND GEOMETRISCHE KONZEPTE	13
1.1. Ein Beispiel	13
1.2. Univariate Statistik	22
1.3. Überblick zur Clusteranalyse	22
1.4. Statistische Grundlagen	26
1.5. Maximum-Likelihood-Methode zur Klassifikation mehrdimensionaler Normalverteilungen	28
1.6. Der gewichtete euklidische Raum und weitere geometrische Begriffe	32
1.7. Deskriptive Statistik (Datenanalyse)	35
1.8. Zentroid, Varianz, Kovarianz, Korrelation	36
1.9. Multivariate Analysemethoden	39
2. DIE DATENMATRIX	41
2.1. Die Datenmatrix X mit I Zeilen und J Spalten	41
2.2. Inhalt und Kodierung, Graphiken	43
2.3. Gewichte der Spaltenpunkte, adaptive Gewichte	47
2.4. Massen der Zeilenpunkte	49
2.5. Aktive und supplementäre Zeilen- und Spaltenpunkte	50
2.6. Fehlende und ungültige Werte	50
2.7. Unglaubliche und ungesicherte Werte	51
2.8. Transformationen	53
2.9. Rangbildung, Rekodierung	55
2.10. Transposition	56
2.11. Aggregation von Zeilen und Spalten der Datenmatrix ("Collapsing")	57
2.12. Zusatzinformationen (Raum- und Zeitkoordinaten)	58
3. FAKTORIELLE ANALYSEN ZUR DATENREDUKTION	59
3.1. Überblick zu den faktoriellen Methoden	59
3.2. Allgemeines numerisches Approximierungsproblem (Singularwertzerlegung)	59
3.3. Hauptkomponentenanalyse	61
3.4. Hauptkomponentenwerte, Graphiken	63
3.5. Ein kleines Beispiel	64

3.6. Ranganalyse	66
3.7. Korrespondenzanalyse	68
3.8. Faktorenanzahl	70
3.9. Rekonstruierte Datenmatrix, Restwertmatrix	71
3.10. Simultane graphische Darstellungen der Zeilen- und der Spaltenpunkte	72
3.11. Permutation der Zeilen und Spalten	72
4. DISTANZ UND ÄHNLICHKEIT	73
4.1. Allgemeine Definition	73
4.2. Transformation von Ähnlichkeiten in Distanzen und umgekehrt	74
4.3. Distanzkoeffizienten für metrische (quantitative) Daten, adaptive Distanzen	75
4.4. Chiquadrat-Abstand, Entropie	79
4.5. Distanzkoeffizienten für nominale (qualitative) und 0-1-Daten	80
4.6. Distanzkoeffizienten für gemischte Daten	84
4.7. Korrelationskoeffizienten	85
4.8. Behandlung von fehlenden Werten	86
4.9. Beispiele	88
5. KLASSENBEGRIFF, DISTANZ ZWISCHEN KLASSEN	91
5.1. Allgemeiner Klassenbegriff	91
5.2. Abstände zwischen Klassen	94
6. HIERARCHISCHE CLUSTERANALYSE	95
6.1. Agglomerative Methoden	96
6.1.1. Single Linkage (Methode "Nächster Nachbar")	98
6.1.2. Complete Linkage (Methode "Entferntester Nachbar")	101
6.1.3. Average Linkage (gewichtete oder ungewichtete Mittelung)	102
6.1.4. Zentroid-Methode (Schwerpunkt-methode)	103
6.1.5. Das Verfahren von Ward (Minimalvarianzmethode)	104
6.1.6. Flexible Strategie nach Lance und Williams	106
6.2. Dendrogramm, Ultrametrik	106
6.3. Spezielle Dendrogramme	109
6.4. Klassenanzahl, Partition	109
6.5. Permutation der Punkte, Ergebnisdarstellung	113
6.6. Klassifikation supplementärer Punkte	113
6.7. Beispiel	114
6.8. Divisive Methoden	117

7. PARTITIONIERENDE CLUSTERANALYSE	119
7.1. Methode der Radiusrestriktion	120
7.2. Minimaldistanzmethode und Beispiel	121
7.3. Austauschverfahren	125
7.4. Klassenanzahl	126
7.5. Klassifikation supplementärer Punkte	127
7.6. Beispiel	131
7.7. Neue, stabilisierte Klassifikationsalgorithmen für adaptive Distanzen	141
8. MULTIVARIATE GRAPHISCHE DARSTELLUNGEN KLASSIFIZierter DATEN	143
8.1. Eine Beispieldatei	143
8.2. Klassendarstellung im faktoriellen Raum, Dendrogramm mit Variablenbewertung	144
9. STABILITÄT DER KLASSEN UND FAKTOREN	151
9.1. Kriterien der Klassifikationsleistung	151
9.2. Variation der Menge aktiver und supplementärer Punkte	158
9.3. Simulationen zur internen Stabilität	159
9.4. Simulationen zur externen Stabilität	165
9.5. Variablenselektion	169
10. EXPERTENSYSTEME ZUR CLUSTERANALYSE	171
11. ANWENDUNGSBEISPIELE ZUR AUTOMATISCHEN KLASSIFIKATION UND FAKTORIELLEN ANALYSE	173
11.1. Vogelatlas Berlin: Klassifikation der Beobachtungsgebiete und Vogelarten	173
11.2. Aufdeckung von Kausalbeziehungen in der medizinischen Forschung und Prädiktion für neue Patienten	182
11.3. Regionalisierung der Beobachtungsgebiete in der geologischen Forschung und Erkundung	183
11.4. Produktionsoptimierung in der Chipfertigung	186
11.5. Leistungsvergleich von Wirtschaftseinheiten in Industrie, Landwirtschaft sowie Handel und Versorgung	187
11.6. Weitere Anwendungen im Überblick	188

12. FORTRAN-PROGRAMME	189
12.1. Kommandosprache und Systemfile (Datenbank)	189
12.2. <i>Create</i> : Kommandointerpretation, Datenmanagement und Simulationstechniken	190
12.3. Faktorielle Methoden zur Datenreduktion und multivariaten graphischen Darstellung	190
12.3.1. <i>CorrAn</i> : Hauptkomponenten- und Korrespondenzanalyse	190
12.3.2. <i>XYPlot</i> : universelle Pseudographik	191
12.4. Hierarchische und partitionierende Clusteranalyse	191
12.4.1. <i>DistAn</i> : Berechnung von Distanzmatrizen	191
12.4.2. <i>Agglom</i> : hierarchische Clusteranalyse	191
12.4.3. <i>KMeans</i> : partitionierende Clusteranalyse	192
12.4.4. <i>CluDia</i> : deskriptive Statistik von Klassen und Klassenhierarchien	192
12.4.5. <i>Neighbor</i> : M-Nächste-Nachbarn-Klassifikation	193
12.4.6. <i>Cross</i> : Kontingenztafelerstellung und Beschreibung von Datentabellen	193
12.5. Softwarekompatibilität und Varianten der Arbeit mit der Software	193
ANHANG A: GRUNDBEGRIFFE DER MATRIZEN- UND VEKTORRECHNUNG	195
ANHANG B: DIE MEHRDIMENSIONALE NORMALVERTEILUNG	197
ANHANG C: χ^2 -VERTEILUNG (TABELLE DER KRITISCHEN WERTE)	199
ANHANG D: F-VERTEILUNG (TABELLE DER KRITISCHEN WERTE)	200
ANHANG E: IRIS-DATEN (KLASSIFIKATIONS DATEN NACH FISHER)	201
LITERATUR	203
SACHVERZEICHNIS	205