



Inhaltsverzeichnis

8.1	Problemstellung	496
8.2	Vorgehensweise	499
8.2.1	Auswahl der Clustervariablen	500
8.2.2	Bestimmung der Ähnlichkeiten	503
8.2.2.1	Ausgangsbeispiel und Überblick zu Proximitätsmaßen	504
8.2.2.2	Proximitätsmaße bei metrisch skalierten Variablen	505
8.2.3	Auswahl des Fusionierungsalgorithmus	512
8.2.3.1	Ablaufschritte der hierarchisch-agglomerativen Verfahren	514
8.2.3.2	Single Linkage, Complete Linkage und Ward-Verfahren	518
8.2.3.3	Fusionierungseigenschaften ausgewählter Clusterverfahren	526
8.2.3.4	Verdeutlichung der Fusionierungseigenschaften an einem erweiterten Beispiel	528
8.2.4	Bestimmung der Clusterzahl	533
8.2.4.1	Analyse von Scree-Plot und Elbow-Kriterium	534
8.2.4.2	Regeln zur Bestimmung der Clusterzahl	535
8.2.4.3	Beurteilung von Robustheit und Güte einer Cluster-Lösung	537
8.2.5	Interpretation einer Cluster-Lösung	538
8.2.6	Empfehlungen zum Ablauf einer hierarchisch, agglomerativen Clusteranalyse ...	540
8.3	Fallbeispiel	540
8.3.1	Problemstellung	540
8.3.2	Durchführung einer Cluster Analyse mit SPSS	542
8.3.3	Ergebnisse	545
8.3.3.1	Ausreißer-Analyse mittels Single Linkage-Verfahren	545
8.3.3.2	Clusterung mithilfe des Ward-Verfahrens	546
8.3.3.3	Interpretation der Zwei-Cluster-Lösung im Fallbeispiel	552
8.3.4	SPSS-Kommandos	555
8.4	Modifikationen und Erweiterungen	558

8.4.1	Proximitätsmaße bei nicht metrischen Daten	558
8.4.1.1	Proximitätsmaße bei nominalem Skalenniveau	558
8.4.1.2	Proximitätsmaße bei binären Variablen	562
8.4.1.3	Proximitätsmaße bei gemischt skaliertem Variablenstruktur	568
8.4.2	Zentrale partitionierende Clusterverfahren	571
8.4.2.1	K-Means Clusteranalyse	571
8.4.2.2	Two-Step Clusteranalyse	574
8.4.2.3	Vergleich von KM-CA und TS-CA	578
8.5	Anwendungsempfehlungen	578
	Literatur	580

8.1 Problemstellung

Empirische Untersuchungen erheben oft die Einschätzungen einer Vielzahl von Personen gegenüber bestimmten Sachverhalten. Häufig ergeben sich dabei große Unterschiede in den Einschätzungen der einzelnen Personen, sodass die Daten in einer Erhebung stark voneinander abweichen. Es besteht eine hohe *Heterogenität*. Werden stark heterogene Daten, z. B. durch einen Mittelwert beschrieben, so ist dieser mit einer entsprechend großen Varianz bzw. Standardabweichung verbunden. Diese deutet statistisch bereits darauf hin, dass der Mittelwert als charakteristisches Maß für die Gesamterhebung eine nur geringe Aussagekraft besitzt. Je geringer hingegen die Heterogenität, desto verlässlicher ist die Aussage des Mittelwertes und desto kleiner die zugehörige Standardabweichung.

Eine Möglichkeit, das Problem der Heterogenität von Daten zu lösen, liegt in der Zusammenfassung von Personen (oder Objekten) zu Gruppen, die in ihren Einschätzungen relativ vergleichbar sind. Die ursprüngliche Erhebungsgesamtheit wird so in Gruppen zerlegt, bei denen die Personen (oder Objekte) in einer Gruppe eine hohe Homogenität aufweisen (*Intragruppen-Homogenität*). Zwischen den Gruppen hingegen besteht eine hohe Heterogenität (*Intergruppen-Heterogenität*). Auf diese Weise können statistische Analysen pro Gruppe vorgenommen werden, die dann pro Gruppe deutlich verlässlichere Ergebnisse liefern.

Die Clusteranalyse ist das methodische Instrument, um heterogene Erhebungsergebnisse in homogene Gruppen zu zerlegen. Sie ist in vielen Disziplinen wie Medizin, Soziologie, Biologie oder Ökonomie anwendbar und wird hier verwendet, um Ähnlichkeiten z. B. von Patienten, Käufern, Pflanzenarten, Unternehmen oder Produkten zu bestimmen. Tab. 8.1 zeigt ausgewählte Fragestellungen in unterschiedlichen Anwendungsfeldern, die alle auf die Bildung von Gruppen von Personen bzw. Objekten abzielen.

Die Fragen machen deutlich, dass die Clusteranalyse zu den explorativen Datenanalyseverfahren gehört, da sie im *Ergebnis* zu Vorschlägen für eine Gruppierung erhobener Untersuchungsobjekte führt und damit „neue Erkenntnisse“ generiert bzw. Strukturen in Datensätzen entdeckt.