Klassifikationsverfahren

Thorsten Meiser und Stefanie Humburg

Unter *Klassifikation* versteht man allgemein die Einteilung einer Menge von Objekten in Gruppen. In diesem Beitrag geht es um Klassifikationen, die auf explizit ausgewählten Merkmalen der Objekte und auf zuvor definierten Kriterien der Zuordnung zu Gruppen beruhen. In diesem Sinne sind die dargestellten Klassifikationsverfahren im Gegensatz zu „naiven" Alltagsklassifikationen transparent und objektiv. Die Zuordnung soll dazu dienen, durch die Gruppenzugehörigkeit eines Objektes Rückschlüsse auf weitere Merkmale ziehen zu können. Die Verfahren ermöglichen eine Informationsverdichtung sowie die strukturelle Analyse von Objektmengen, deren Gliederung bislang unbekannt ist.

1 Ziele und Grundlagen der Klassifikation und Diskrimination

Es ist zu unterscheiden zwischen Verfahren der *Diskrimination* und der *Klassifikation*. Bei der Diskrimination geht es darum, Zuordnungsregeln aus Stichproben von Objekten abzuleiten, deren Gruppierung bereits vorliegt (zur *Diskriminanzanalyse* vgl. Andres, in diesem Band). Klassifikation hingegen bezieht sich auf die Anwendung von Zuordnungsregeln auf Objekte, deren Gruppenzugehörigkeit noch nicht bekannt ist. Klassifikation kann sich demnach an eine Diskriminationsphase anschließen, wenn auf neue Objekte Zuordnungsregeln angewendet werden, die mit Hilfe einer Stichprobe bereits klassifizierter Objekte gewonnen wurden. Wenn es eine solche Stichprobe nicht gibt, bedeutet Klassifikation, eine Menge von Objekten nach ausgewählten Merkmalen und Kriterien in Gruppen einzuteilen, wobei nur die innere Struktur dieser einen Menge benutzt werden kann. Damit ist nicht nur die Zuordnung, sondern auch die Gruppenstruktur selbst Ergebnis der Klassifikation.

Clusteranalytische Methoden sind die wohl bekanntesten Verfahren für den zuletzt genannten Fall. Grundlage der Gruppierung stellen dabei zuvor definierte Merkmale der Objekte dar, die es ermöglichen, die Distanz oder (Un-)Ähnlichkeit zwischen den Objekten zu bestimmen. Hauptanliegen von Clusteranalysen ist es dabei, dem Betrachter eine möglichst einfache Darstellung einer komplexen multivariaten Datensituation zu geben. Neben der Reduktion großer Objektmengen auf eine überschaubare Anzahl homogener Gruppen bieten sie im Sinne der exploratorischen Datenanalyse auch die Möglichkeit zur Generierung von Hypothesen.

Die multivariate Datensituation, auf die sich Klassifikationsverfahren beziehen, stellt sich formal wie folgt dar (vgl. auch Andres, in diesem Band): n Objekte (Personen, Lebewesen, Gegenstände, ...) werden hinsichtlich p Merkmale untersucht. Die resultierende $n \times p$ Rohdatenmatrix $\mathbf{X} := [x_{iq}]$ enthält die Realisation jedes Objektes i ($i = 1, \ldots, n$) in jeder Merkmalsvariablen X_q ($q = 1, \ldots, p$). Für jedes

https://doi.org/10.25521/HQM22

Objekt liegt somit ein p-dimensionaler Merkmalsvektor vor. Geometrisch betrachtet besteht das Ziel einer Clusteranalyse nun darin, die Objekte, die in dem durch die Merkmale aufgespannten p-dimensionalen Raum eine geringe Distanz zueinander aufweisen, zu einem gemeinsamen *Cluster* zusammenzufassen. Damit ergibt sich das allen clusteranalytischen Verfahren gemeinsame Kriterium: Objekte eines Clusters sollen möglichst nahe beieinander liegen, d.h. ein hohes Maß an Ähnlichkeit aufweisen (*interne Homogenität*), wohingegen Objekte unterschiedlicher Cluster weiter voneinander entfernt liegen, d.h. einander unähnlicher sein sollen (*externe Heterogenität* oder *Isolation*). Die Vielfalt der Verfahren ergibt sich nun einerseits durch verschiedene Herangehensweisen an das Aufspüren intern homogener und extern voneinander abgegrenzter Subpopulationen. Andererseits sind die Verfahren durch unterschiedliche Konzeptualisierungen des Distanz- oder (Un-)Ähnlichkeitsbegriffes zwischen Objekten und Clustern definiert. Wie unten beschrieben wird, unterscheiden sich clusteranalytische Verfahren wesentlich darin, welche *Kriteriumsfunktion* bei der Vereinigung von Objekten oder Clustern zu neuen Clustern optimiert, d.h. welches Kriterium der (Un-)Ähnlichkeit oder Distanz herangezogen wird.

Bezeichne M die Menge der Objekte, so kann ein Maß d für die Distanz zwischen zwei Objekten aus M als Funktion von $M \times M$ in die Menge der reellen Zahlen angesehen werden. Die Funktion d heißt *Metrik auf* M, und der Funktionswert $d(i,j)$ heißt *Distanz zwischen den Objekten* i *und* j, wenn die folgenden Axiome gelten:

$$d(i,j) = 0 \Leftrightarrow i = j \quad (Identität), \tag{1}$$

$$d(i,j) = d(j,i) \quad (Symmetrie), \tag{2}$$

$$d(i,j) \leq d(i,k) + d(k,j) \quad (Dreiecksungleichung). \tag{3}$$

Anschaulich bedeutet die Dreiecksungleichung (3), daß der Umweg von i nach j über ein drittes Objekt k nicht kleiner sein darf als der direkte Weg von i zu j. Aus den Axiomen (1), (2) und (3) folgt zudem

$$d(i,j) \geq 0 \quad (Positivität). \tag{4}$$

Im Kontext clusteranalytischer Verfahren spielt zudem das Konzept der *Ultrametrik* (Johnson, 1967) eine zentrale Rolle, das durch die obigen Axiome unter Verschärfung der Dreicksungleichung definiert ist. Statt (3) wird hier gefordert

$$d(i,j) \leq max(d(i,k), d(k,j)). \tag{5}$$

Während Ultrametriken als Ergebnis spezieller clusteranalytischer Verfahren resultieren (siehe Abschnitt 2.5), stellt die $n \times n$ Matrix der Distanzwerte $d(i,j)$ einer ausgewählten Metrik die Ausgangsbasis für zahlreiche clusteranalytische Methoden dar. Häufig werden die Distanzen zwischen den Objekten im p-dimensionalen Raum durch einen Spezialfall der Klasse der *Minkowski-r-Metriken*

$$d(i,j) = \left(\sum_{q=1}^{p} |x_{iq} - x_{jq}|^r \right)^{\frac{1}{r}} \quad mit \quad r \geq 1 \tag{6}$$

definiert. Die bekanntesten Spezialfälle ergeben sich durch die Spezifikationen $r = 1$ (*City-Block*-Metrik), $r = 2$ (Euklidische Metrik) und $r = \infty$ (Dominanz- oder

Supremum-Metrik). Bei der Festlegung von r ist zu beachten, daß hiermit theoretische Interpretationen verbunden sind. So entspricht die Distanz zwischen zwei Objekten bei Verwendung der *City-Block*-Metrik der ungewichteten Summe der Differenzbeträge auf allen Merkmalsdimensionen. Mit wachsendem r steigt der Einfluß von Dimensionen, auf denen für ein gegebenes Objektpaar größere Differenzbeträge vorliegen. Im Extremfall der Dominanz-Metrik schließlich ist die Distanz zweier Objekte identisch mit ihrem Differenzbetrag auf derjenigen Dimension, auf der sie die größte Differenz aufweisen. Es ist in jedem Anwendungsfall zu klären, ob große Differenzen auf einzelnen Dimensionen durch kleine Differenzen auf anderen Dimensionen kompensiert werden oder ob große Differenzen auf einigen oder auch nur einer einzigen Dimension bereits zu hohen Distanzwerten zwischen Objekten führen.

Die Verwendung einer Metrik nach (6) setzt voraus, daß die p Merkmale mindestens auf Intervallskalenniveau erhoben wurden. Weitere Maße für die Ähnlichkeit von Objekten, einschließlich solcher für kategoriale und ordinale Merkmale, werden von Bock (1974), Eckes und Roßbach (1980), Everitt (1993), Hand (1981) und Moosbrugger und Frank (1992) dargestellt. Es muß erwähnt werden, daß nicht alle derartigen Maße für die Ähnlichkeit von Objekten Distanzmaße im Sinne einer Metrik (Axiome (1)-(3)) sind. Neben den bisher besprochenen *indirekten* Verfahren, d.h. Distanz- oder (Un-)Ähnlichkeitseinschätzungen durch Verwendung von p Objektmerkmalen rechnerisch zu ermitteln, können ebenso *direkte* Verfahren eingesetzt werden (vgl. Eckes & Roßbach, 1980; Moosbrugger & Frank, 1992; Oldenbürger, 1983). Direkte Einschätzungen können beispielsweise durch Befragungen gewonnen werden, in denen Personen angeben, wie ähnlich Objekte ihrer Meinung nach sind. Es gibt eine Vielzahl von Möglichkeiten, auf direktem Wege Ähnlichkeitseinschätzungen zu erhalten, und auch hier sollte die Wahl des Verfahrens dem untersuchten Gegenstandsbereich und dem theoretischen Hintergrund gerecht werden.

2 Hierarchische clusteranalytische Verfahren

Hierarchische Clusterverfahren erstellen *Sequenzen von Partitionen* auf der Menge der Objekte. Partitionen auf M sind Aufteilungen der Menge M in disjunkte und gemeinsam erschöpfende nichtleere Teilmengen. Bei hierarchisch *agglomerativen* Verfahren besteht die Anfangspartition aus n ein-elementigen Teilmengen, d.h. jedes Objekt bildet eine Teilmenge oder – anders ausgedrückt – ein eigenes Cluster. In $n-1$ Stufen werden nun Cluster aufgrund ihrer Distanz oder (Un-)Ähnlichkeit zusammengefaßt, bis schließlich die Endpartition erreicht ist, die aus nur einer Menge mit allen n Objekten besteht. Welche beiden Cluster auf einer Stufe des Agglomerationsprozesses zu einem neuen Cluster vereinigt werden, ist abhängig von der gewählten Kriteriumsfunktion, die durch die Wahl der zu vereinigenden Cluster minimiert werden soll. Die diversen Verfahren unterscheiden sich hinsichtlich dieser Kriteriumsfunktion, und somit ist die Sequenz der Partitionen *methodenabhängig*. Zu beachten ist außerdem, daß die Zuordnung von Objekten zu einem gemeinsamen Cluster irreversibel ist, d.h. daß eine Zusammenfassung von Objekten auf einer späteren Stufe des Prozesses nicht rückgängig gemacht werden kann. Hierarchisch *divisive* Clusterverfahren hingegen starten mit der vollständigen Menge M als Anfangspartition und zerlegen diese in $n-1$ Schritten bis zur Endpartition, die hier aus

n ein-elementigen Teilmengen besteht. Die Richtung des Prozesses ist also entgegengesetzt zu dem der hierarchisch agglomerativen Methoden. Da die agglomerativen Verfahren in der Literatur sowie in der Anwendungspraxis eine bedeutend größere Rolle spielen, beschränken wir uns hier auf die Darstellung ausgewählter Verfahren dieser Methodenklasse, nämlich des *single linkage*, des *complete linkage*, des *group average linkage* und der *Methode nach Ward*. Darstellungen divisiver Clusterverfahren finden sich in der einschlägigen Literatur (vgl. Bock, 1974; Eckes & Roßbach, 1980; Everitt, 1993; Hand, 1981; Moosbrugger & Frank, 1992).

2.1 Single-Linkage- oder Minimum-Methode

Wie bereits erwähnt, werden auf jeder Stufe des Agglomerationsprozesses die Objekte oder Cluster zu neuen Clustern zusammengefaßt, für die eine reellwertige Kriteriumsfunktion C minimiert wird. In den meisten Verfahren kann C als Maß für die Entfernung oder Unähnlichkeit von Clustern zueinander interpretiert werden. In der *Single-Linkage-* oder Minimum-Methode ist die Kriteriumsfunktion C wie folgt definiert (vgl. Hubert, 1974; Johnson, 1967): Seien M_s und M_t zwei Elemente einer Partition auf M, d.h. Teilmengen, die auf einer bestimmten Stufe des Prozesses nicht in einem Cluster vereinigt sind. Dann ist für dieses Clusterpaar

$$C(M_s, M_t) := min\{d(i,j)|i \in M_s, j \in M_t\}. \tag{7}$$

Damit ist die Distanz zwischen zwei Clustern definiert als die Distanz der beiden Objekte $i \in M_s$ und $j \in M_t$, die sich am nächsten sind. Es werden also jeweils die beiden Cluster zu einem neuen Cluster fusioniert, für die die Distanz *eines* Objektpaares minimal ist, wobei die anderen Objekte der Cluster irrelevant sind.

Anders ausgedrückt bildet beim *single linkage* ein Cluster eine Gruppe von Objekten, in der jedes Mitglied mindestens einem anderen Mitglied desselben Clusters ähnlicher ist als den Mitgliedern eines anderen Clusters. Diese Definition und die dazugehörige Kriteriumsfunktion (7) führen dazu, daß durch das *single linkage* leicht sogenannte Ketten als Cluster entstehen (*chaining effect*), da nur die Distanzen benachbarter Objekte berücksichtigt werden. Demnach können in einem Cluster sehr weit voneinander entfernte Objekte enthalten sein, wenn sie nur durch eine Kette hinreichend naher Objekte miteinander verbunden sind. Diese Tatsache führt dazu, daß häufig viele Objekte in einer geringen Anzahl von Clustern oder in nur einem Cluster mit mehr als einem Objekt zusammengefaßt werden (vgl. Hubert & Schultz, 1975). Dieser Effekt hängt mit der *Kontraktion des Raumes* um ein anwachsendes Cluster zusammen: Die Distanz eines neuen Clusters zu den übrigen Objekten oder Clustern ist kleiner oder gleich den vorherigen Distanzen der zusammengelegten Cluster zu den übrigen Objekten oder Clustern. Daher zieht sich durch Vergrößerung eines Clusters der Raum um dieses Cluster gleichsam zusammen, was tendenziell zum weiteren Anwachsen desselben Clusters führt („*space-contracting strategy*", Hubert & Schultz, 1975; vgl. auch Blashfield & Aldenderfer, 1988).

2.2 Complete-Linkage- oder Maximum-Methode

Hinsichtlich der Kriteriumsfunktion sowie der Eigenschaften der gebildeten Cluster ist die *Complete-Linkage-* oder Maximum-Methode gegensätzlich zum *Single-*

Linkage-Verfahren. Die zu minimierende Kriteriumsfunktion ist durch

$$C(M_s, M_t) := max\{d(i,j)|i \in M_s, j \in M_t\} \tag{8}$$

definiert (vgl. Hubert, 1974; Johnson, 1967), wobei M_s und M_t wie oben zwei beliebige Elemente einer gegebenen Partition auf M sind. Es werden also jeweils die beiden Cluster miteinander vereinigt, für die die Distanz der am weitesten entfernten Objekte $i \in M_s$ und $j \in M_t$ minimal ist. Hier spielt also nicht die Distanz benachbarter Objekte zweier Cluster eine Rolle, sondern der maximale Durchmesser des resultierenden Clusters in bezug auf alle enthaltenen Objekte.

Aufgrund dieses Kriteriums führt das *Complete-Linkage*-Verfahren im Gegensatz zum *Single-Linkage*-Algorithmus eher zum Anwachsen mehrerer kleinerer Cluster, in denen sich alle Objekte relativ nahe sind. Dies hängt mit der *Streckung des Raumes* durch Anwachsen der Cluster zusammen: Die Distanz eines neuen Clusters zu den umgebenden Objekten oder Clustern ist größer oder gleich den Distanzen der fusionierten Einzelcluster zu den umgebenden Objekten oder Clustern. Dies führt zu einem Auseinanderrücken der Cluster durch Anwachsen und zur Bildung einer größeren Anzahl von Clustern mit mehreren Objekten, die etwa gleich stark besetzt sind („*space-dilating technique*", Hubert & Schultz, 1975; vgl. auch Blashfield & Aldenderfer, 1988).

Da im *Single-Linkage*-Verfahren weit auseinander liegende, durch Ketten verbundene Objekte in Clustern zusammengefaßt werden können, kann diese Methode als eher „liberal" bezeichnet werden, wohingegen der *Complete-Linkage*-Algorithmus als eher „konservativ" angesehen werden kann (Blashfield & Aldenderfer, 1988).

2.3 Group-Average-Linkage-Methode

Die Verfahren des *single linkage* und *complete linkage* stehen einander konträr gegenüber und betonen unterschiedliche Aspekte der Daten. Dies hängt damit zusammen, daß bei der Definition der Clusterdistanz in beiden Verfahren nur die Distanz eines einzigen Objektpaares herangezogen wird, nämlich des Paares mit der kleinsten beziehungsweise größten Distanz. Es liegt daher nahe, Verfahren heranzuziehen, die einen Kompromiß zwischen diesen beiden Extrema darstellen und mehr Dateninformation der Cluster einbeziehen. Eine derartige Methode ist das *Group-Average-Linkage*-Verfahren. Hierin ist die Kriteriumsfunktion C definiert als arithmetisches Mittel der Distanzwerte $d(i,j)$ aller Objektpaare mit $i \in M_s$ und $j \in M_t$, d.h.

$$C(M_s, M_t) := \frac{1}{n_s n_t} \sum_{i \in M_s} \sum_{j \in M_t} d(i,j). \tag{9}$$

In (9) bezeichnet n_s die Anzahl der Objekte in M_s, n_t entsprechend die Anzahl der Objekte in M_t. Bei Verwendung dieser Kriteriumsfunktion müssen die Distanzwerte allerdings höheres Skalenniveau haben als bei den Verfahren des *single linkage* und *complete linkage*, bei denen die gebildete Clusterstruktur invariant unter monotonen Transformationen von $d(i,j)$ ist. Demgegenüber schlagen Hubert und Schultz (1975) mit der *r-Diameter-Methode* eine Klasse von Verfahren vor, die die Methoden des *single linkage* sowie des *complete linkage* als Spezialfälle enthält und ebenfalls ausschließlich die ordinale Information der Distanzwerte nutzt (man beachte die unterschiedliche Bedeutung von r an dieser Stelle gegenüber Gleichung (6)).

2.4 Die Methode nach Ward

Ward (1963) stellt einen allgemeinen Algorithmus zur hierarchischen Gruppierung von Objekten vor, der auf der Optimierung einer „objektiven Funktion" im Sinne der oben eingeführten Kriteriumsfunktion beruht. Der Autor betont, daß inhaltliche Gründe für die Wahl dieser Funktion herangezogen werden müssen. Lediglich als Beispiel einer solchen objektiven Funktion benutzt er das unten vorgestellte Varianzkriterium (für den Spezialfall $p = 1$). Es erscheint daher fraglich, ob die übliche Bezeichnung „Wardsche Methode" für die Anwendung dieser speziellen Kriteriumsfunktion dem Anspruch des Autors gerecht wird, da er eigentlich den Grundstein einer weitaus größeren Klasse von Verfahren legte (vgl. auch Eckes & Roßbach, 1980). So bemerkt auch Johnson (1967, S. 242), daß die Methode von Ward (1963) so generell ist, daß sich die von ihm vorgestellten, unter monotonen Transformationen von Ähnlichkeitsdaten invarianten Verfahren darin einordnen lassen. Dennoch übernehmen wir hier den üblichen Sprachgebrauch und bezeichnen das hierarchisch agglomerative Clusterverfahren, in dem das folgende Varianzkriterium minimiert wird, als „Methode nach Ward" oder „Wardsches Verfahren".

Kehren wir zurück zur geometrischen Vorstellung, daß die n Objekte in einem p-dimensionalen Merkmalsraum liegen. Für ein Cluster M_s mit n_s Objekten kann dann der p-dimensionale Mittelwertsvektor (*Zentroid*) bestimmt werden. Im Wardschen Verfahren wird nun die Summe der quadrierten Abweichungen (Quadratsumme, QS), die die Objekte in M_s von ihrem Zentroiden auf den einzelnen Dimensionen aufweisen, betrachtet:

$$QS_{M_s} = \sum_{i \in M_s} \sum_{q=1}^{p} \left(x_{iq} - \sum_{i \in M_s} \frac{x_{ig}}{n_s} \right)^2 . \tag{10}$$

Die rechte Seite von Gleichung (10) kann anschaulich als Summe der quadrierten euklidischen Distanzen der Clusterelemente zu ihrem Zentroiden interpretiert werden. Die zu minimierende Kriteriumsfunktion in der Methode nach Ward ist nun die Summe der Intra-Cluster-Quadratsummen (10) über die Cluster einer Partition

$$C := \sum_{s} QS_{M_s} = \sum_{s} \sum_{i \in M_s} \sum_{q=1}^{p} \left(x_{iq} - \sum_{i \in M_s} \frac{x_{iq}}{n_s} \right)^2 . \tag{11}$$

Folglich werden auf einer Stufe des Agglomerationsprozesses die beiden Cluster vereinigt, die zu einer minimalen Steigerung von (11) führen. Der Zuwachs Δ in C, der durch die Fusionierung zweier Cluster M_s und M_t auf einer Stufe des Prozesses erfolgt, kann ausgedrückt werden durch

$$\begin{aligned} \Delta(M_s, M_t) &= QS_{M_s \cup M_t} - QS_{M_s} - QS_{M_t} = \\ & n_s \textstyle\sum_{q=1}^{p} \left(\sum_{i \in M_s} \frac{x_{iq}}{n_s} - \sum_{i \in M_s \cup M_t} \frac{x_{iq}}{n_s + n_t} \right)^2 + \\ & n_t \textstyle\sum_{q=1}^{p} \left(\sum_{i \in M_t} \frac{x_{iq}}{n_t} - \sum_{i \in M_s \cup M_t} \frac{x_{iq}}{n_s + n_t} \right)^2 . \end{aligned} \tag{12}$$

Es zeigt sich, daß $\Delta(M_s, M_t)$ die gewichtete Summe der quadrierten euklidischen Distanzen der Zentroiden von M_s und M_t zu dem gemeinsamen Zentroiden des entstehenden Clusters ist (zu verwandten Interpretationen vgl. auch Eckes & Roßbach,

1980, S. 74f.). Δ kann auch als Maß für die Distanz zweier Cluster auf einer Stufe des Agglomerationsprozesses betrachtet werden, so daß wiederum die beiden Cluster vereinigt werden, die verglichen mit allen anderen Clusterpaaren die geringste Distanz zueinander aufweisen. Zu beachten ist, daß die Merkmale bei Verwendung dieses Kriteriums mindestens Intervallskalenniveau haben müssen.

Die Kriteriumsfunktion (11) heißt auch *Spurkriterium*, da C hier die *Spur*, d.h. die Summe der Diagonalelemente, der multivariaten *Within-Cluster-Dispersionsmatrix* darstellt. Aus der varianzanalytischem Denken nahestehenden Zerlegung der Dispersionsmatrix in eine *Within-Cluster-* und eine *Between-Cluster-Dispersionsmatrix* (vgl. z.B. Eckes & Roßbach, 1980; Everitt, 1993; Hand, 1981) folgt, daß mit der Minimierung von (11) gleichzeitig die Summe der Inter-Cluster-Quadratsummen (Spur der *Between-Cluster-Dispersionsmatrix*) maximiert wird. Das bedeutet, daß simultan die interne Homogenität im Sinne einer kleinen Summe von quadrierten Intra-Cluster-Distanzen der Objekte zu ihren Zentroiden sowie die externe Heterogenität im Sinne einer großen Summe gewichteter quadrierter Inter-Cluster-Distanzen der Zentroide zum Gesamt-Zentroiden aller Objekte optimiert werden.

2.5 Hierarchische Clusterverfahren und das Konzept der Ultrametrik

Johnson (1967) argumentiert, daß unterschiedliche hierarchisch agglomerative Verfahren notwendigerweise zu identischen Lösungen kommen, falls die Distanzwerte $d(i,j)$ die Axiome einer Ultrametrik (s.o.) erfüllen (vgl. auch Hubert, 1974). Andererseits zeigt er, daß das von ihm definierte Konzept eines „*hierarchical clustering scheme*“ (S. 243) eine Ultrametrik auf der Menge der Objekte induziert. Sei $h(i,j)$ der Wert der Kriteriumsfunktion C auf der Stufe des Agglomerationsprozesses, auf der die Objekte i und j zum ersten Mal einem gemeinsamen Cluster angehören. $h(i,j)$ wird dann auch als *Wert der Partition*, die durch Vereinigung der beiden Cluster M_s und M_t mit $i \in M_s$ und $j \in M_t$ zustande kommt, bezeichnet. Unter der Voraussetzung, daß die Werte der Partitionen im Laufe der Sequenz monoton ansteigen, ist h eine Ultrametrik auf M. Johnson (1967) setzt diese Monotonie in seiner Definition eines hierarchischen Clusterschemas voraus. Milligan (1979) leitet hinreichende Bedingungen für die Gültigkeit dieser Voraussetzung bezüglich aller hierarchisch agglomerativen Verfahren im hier verwendeten allgemeineren Sinne her.

Daß die Funktion h die Axiome (1) und (2) erfüllt, ist leicht einzusehen, da ausschließlich die Anfangspartition den Wert Null hat und sich die Symmetrie unmittelbar ergibt. Interessanter ist die Gültigkeit der Ungleichung (5). Hier lassen sich zwei Fälle unterscheiden: Erstens, das Objekt k ist bereits mit Objekt i in einem gemeinsamen Cluster M_s enthalten, wenn die Cluster M_s und M_t fusioniert werden. Dann ist $h(i,k) < h(i,j) = h(k,j)$, und folglich ist $h(i,j) \leq max(h(i,k), h(k,j))$ erfüllt. Zweitens, das Objekt k ist weder in M_s noch in M_t enthalten, wenn diese beiden Cluster vereinigt werden. Dann werden i und j gemeinsam an einer späteren Stelle der Sequenz mit k in einem Cluster vereint, und es ergibt sich $h(i,j) < h(i,k) = h(j,k)$. Folglich ist (5) auch in diesem Fall erfüllt. Daher resultiert notwendigerweise eine Ultrametrik aus einer hierarchischen Clusteranalyse, sofern die Werte der Partitionen monoton ansteigen. Dies trifft auf die unter 2.1 bis 2.4 erläuterten Verfahren zu.

Die hier nicht beschriebenen hierarchisch agglomerativen Verfahren der *Zentroid-* und *Medianmethode* hingegen gewährleisten diese Eigenschaft nicht (vgl. etwa Bock, 1974; Eckes & Roßbach, 1980; Milligan, 1979).

3 Nicht-hierarchische Clusterverfahren

Die Anwendung hierarchischer Clusterverfahren wird bei großen Objektmengen unmöglich, da die Berechnung der Distanzen für alle Objektpaare dann nicht praktikabel ist. Eine Alternative bieten die *iterativ partitionierenden* Clusterverfahren. Auch hier werden Partitionen auf der Objektmenge erstellt, es gibt allerdings zwei wesentliche Unterschiede zu den hierarchischen Methoden: Erstens ist die Zuordnung von Objekten zu Clustern nicht irreversibel, d.h. Objekte können im Verlauf der Berechnung verschiedenen Clustern zugeordnet werden. Zweitens bleibt die Anzahl der Teilmengen in den Partitionen konstant und wird a priori spezifiziert. Während bei den hierarchisch agglomerativen Verfahren alle Stadien von n Clustern bis zu einem Cluster durchlaufen werden, ändert sich bei den partitionierenden Verfahren also lediglich die Zuordnung von Objekten zu den Clustern.

Auch bei den partitionierenden Verfahren wird die Zuordnung von Objekten im Sinne der Optimierung einer Kriteriumsfunktion durchgeführt. Ausgehend von einer Startpartition der Objektmenge in eine vorgegebene Anzahl k von Clustern werden Objekte zwischen den Clustern „verschoben", bis keine Verschiebung mehr zu einer weiteren Optimierung der Funktion führt. Das am weitesten verbreitete partitionierende Verfahren ist der *k-Means-Algorithmus*, der ebenfalls auf der Kriteriumsfunktion (11) beruht. In diesem Verfahren werden die Cluster-Zentroide zu Beginn des Prozesses für die Startpartition festgelegt und im weiteren Verlauf für die gebildeten Partitionen berechnet. Danach wird für jedes Objekt geprüft, ob es zu dem Zentroiden eines anderen Clusters eine geringere Distanz aufweist als zu dem des eigenen Clusters. Ist dies der Fall, wird das Objekt in das Cluster mit dem näheren Zentroiden verschoben. Die Zentroide der neuen Cluster werden dabei entweder nach jeder Verschiebung eines einzelnen Objektes oder aber, nachdem alle Objekte hinsichtlich ihrer Clusterzugehörigkeit überprüft wurden, neu berechnet. Durch diese Neuberechnung können wieder Verschiebungen von Objekten notwendig werden. Bildlich gesprochen kann sich der Schwerpunkt eines Clusters im Raum durch Aufnahme neuer oder Verlust bisheriger Elemente verändern und damit von eigenen Elementen wegrücken oder sich fremden annähern. Daraus ergibt sich ein iterativer Prozeß, der zum Ende kommt, wenn kein Punkt mehr näher an einem fremden als an dem eigenen Zentroiden liegt. Da die Verschiebung von Objekten deren Distanz zum jeweiligen Cluster-Zentroiden minimiert und die neu berechneten Zentroide ihrerseits die internen Quadratsummen minimieren, optimiert dieser Algorithmus die Kriteriumsfunktion (11).

Ein Problem stellt bei diesem Verfahren die Möglichkeit eines *lokalen Optimums* dar: Der Optimierungsprozeß wird beendet, wenn keine Verschiebung eines Objektes *innerhalb einer gegebenen Partition* mehr zu einem geringeren Wert von C führt. Dies schließt nicht aus, daß es andere Partitionen gibt, die im Sinne der Kriteriumsfunktion besser sind. Um die Möglichkeit auszuschließen, bei einer suboptimalen Lösung zu landen, müßten alle möglichen Partitionen von M in k Cluster

miteinander verglichen werden, was bereits bei moderaten Datenmengen praktisch unmöglich ist. Um die Gefahr eines lokalen Optimums zu reduzieren, empfiehlt es sich, die Startpartition für den Algorithmus sorgfältig auszuwählen. Simulationsstudien von Milligan (1980) deuten darauf hin, daß die *k-Means*-Methode hinsichtlich der Auffindung einer vorgegebenen Clusterstruktur mindestens ebenso gut ist wie die bekannten hierarchischen Verfahren, sofern in die Startpartition Informationen über die tatsächliche Clusterstruktur oder über das Ergebnis einer zuvor durchgeführten hierarchischen Clusteranalyse (der Autor verwendet hierbei die Zentroide zur Bildung der Startpartition, die mit Hilfe der *Group-Average-Linkage*-Methode gewonnen wurden) eingehen.

4 Weitere Verfahren der Klassifikation

Neben den clusteranalytischen Methoden gibt es weitere Verfahrensklassen, die dem Ziel dienen, Gruppierungen von Objekten aus einer multivariaten Datenmatrix $\mathbf{X}$ zu bilden, ohne daß Zuordnungsregeln aus einer Stichprobe bereits klassifizierter Objekte zur Verfügung stehen. Wolfe (1970) hat *finite Mischverteilungsmodelle* (vgl. Rost & Erdfelder, in diesem Band) als Alternative vorgeschlagen und Anwendungen gemischter multivariater Normalverteilungen vorgestellt. Für p diskrete Indikatoren stellt die latente Klassenanalyse (Langeheine & Rost, in diesem Band) den wohl bekanntesten Spezialfall finiter Mischverteilungsmodelle dar. Giegler und Rost (1993) vergleichen clusteranalytische Verfahren und Modelle der latenten Klassenanalyse sowohl methodologisch als auch empirisch.

Ein bedeutsamer Unterschied der finiten Mischverteilungsmodelle zu den clusteranalytischen Verfahren besteht darin, daß bei den letzteren jedes Objekt genau einem Cluster zugeordnet wird, wohingegen die Objekte in finiten Mischverteilungsmodellen für jede Subpopulation eine Wahrscheinlichkeit der Zugehörigkeit haben, die (von Grenzfällen abgesehen) größer als Null ist.

Dimensionale Methoden wie *Faktorenanalysen* (vgl. Schönemann & Borg, in diesem Band) der $n \times n$-Korrelationsmatrix der Objekte statt der $p \times p$-Korrelationsmatrix der Merkmale und *multidimensionale Skalierungsverfahren* (vgl. Borg, in diesem Band; Borg & Schönemann, in diesem Band) für $n \times n$-Nähematrizen können ebenfalls zur Klassifikation herangezogen werden. Als Kriterien der Gruppierung dienen hierbei die Ladungsmuster der Objekte auf den Faktoren respektive die Werte, die die Objekte auf den einzelnen Dimensionen der resultierenden Konfiguration aufweisen, sowie geometrische Gruppenbildungen in der Ergebnisdarstellung.

5 Möglichkeiten und Probleme objektiver Klassifikationsverfahren

Einen wesentlichen Vorteil im Sinne der exploratorischen Zielsetzung, eine einfache und illustrative Annäherung an die komplexe multivariate Datensituation zu ermöglichen, stellen *Dendrogramme* als Visualisierung hierarchischer Clusteranalysen dar. In Dendrogrammen oder Baumdarstellungen wird die Sequenz der Zusammenfassung von Objekten zu Clustern abgebildet. Dieses und weitere graphische Hilfsmittel dienen dem Ziel, die vorgefundene Struktur in den Daten zu veranschaulichen und die Interpretation der Clusterlösung zu vereinfachen (vgl. Moosbrugger & Frank, 1992).

Wie bereits erwähnt, liefern jedoch unterschiedliche clusteranalytische Verfahren unterschiedliche Ergebnisse. Das Resultat einer Klassifikation ist daher keine unmittelbare Abbildung einer datenimmanenten Struktur. Vielmehr handelt es sich um eine hochgradig methodenabhängige Beschreibung der Daten, die jeweils besondere Eigenschaften in den Daten hervorhebt. Ein entscheidendes Kriterium für das Ergebnis einer Klassifikation ist die Nützlichkeit für die weitere Forschung (vgl. etwa Eckes & Roßbach, 1980; Everitt, 1993; Marradi, 1990). Es ist vom betrachteten Gegenstandsbereich und den jeweiligen theoretischen Vorannahmen abhängig zu machen, welche Methode herangezogen wird: „This is a general point as far as cluster analysis goes, and one that cannot be over-emphasized: the different methods have different properties and what is a disadvantage for one problem could be an advantage for another." (Hand, 1981, S. 169). Dies schließt allerdings nicht aus, mit Hilfe simulierter Daten die Fähigkeit der einzelnen Verfahren zu untersuchen, bekannte Strukturen aufzudecken. Eine Serie derartiger Monte-Carlo-Studien stellt beispielsweise Milligan (1980) vor. Der Autor evaluiert fünfzehn clusteranalytische Verfahren hinsichtlich der Rekonstruktion vorgegebener Cluster und der Anfälligkeit gegenüber ausgewählten Fehlerquellen in den Daten. Es zeigt sich, daß auch die Auswirkungen von Fehlerquellen verfahrensspezifisch sind. Von den oben dargestellten clusteranalytischen Verfahren weist hierbei der *k-Means*-Algorithmus eine besondere Robustheit auf, sofern die Startpartition nicht zufällig gewählt wird.

Ein Problem clusteranalytischer Verfahren besteht weiterhin darin, daß sie eine Klassifikation unabhängig davon liefern, ob den Daten eine Clusterstruktur innewohnt oder nicht. Es gilt daher, die Grundannahme zu prüfen, daß den Daten überhaupt eine Struktur zugrunde liegt, die sich auf verfahrensspezifische Art und Weise im Ergebnis ausdrückt. Hierfür wurden *interne Kriterien* vorgeschlagen, die den Zusammenhang zwischen den Ausgangsdaten und der Zuordnung der Objekte in der ermittelten Partition erfassen, also ausschließlich die Information innerhalb des Clusterprozesses berücksichtigen (vgl. etwa Hubert, 1974; Milligan, 1980, 1981).

Oldenbürger und Schwibbe (1980) schlagen vor, die Clusterbarkeit unabhängig vom Ergebnis einer Clusteranalyse zu testen. Ihr Verfahren beruht auf einem Maß für die Abweichung der empirischen Distanzwerte von einer Ultrametrik.

Externe Kriterien dienen dagegen dem Ziel, eine ermittelte Partition anhand gegebener Informationen über die Struktur der Daten, einer theoretisch angenommenen Gruppierung oder aber des Ergebnisses einer anderen Analyse zu evaluieren (vgl. etwa Hubert & Baker, 1977; Milligan, 1980, 1981; Rand, 1971, und Morey & Agresti, 1984). Diese Kriterien verwenden also Informationen zur Beurteilung eines Clusterergebnisses, die außerhalb des jeweiligen Clusterprozesses liegen.

Zur Evaluation einer Clusterlösung bieten sich auch Vorgehensweisen an, die die Reproduzierbarkeit der ermittelten Struktur zum Gegenstand haben. So können die gleichen Daten mit unterschiedlichen Verfahren analysiert werden, es können Ergebnisse mit denen von Teilstichproben von Objekten verglichen werden, oder disjunkte Objektmengen können im Sinne einer Kreuzvalidierung eingesetzt werden.

Bei der Interpretation von Klassifikationsergebnissen ist die Festlegung der Anzahl von Gruppen von besonderer Bedeutung. Im Zusammenhang clusteranalytischer Verfahren sind sogenannte *stopping rules* vorgeschlagen worden, um die optimale Anzahl von Clustern, z.B. in einer hierarchischen Sequenz, zu ermitteln. Milligan

und Cooper (1985) evaluieren in einer Monte-Carlo-Studie 30 ausgewählte *stopping rules* (vgl. auch Atlas & Overall, 1994). Eher intuitive Verfahren, die Clusterzahl festzulegen, basieren etwa auf der Betrachtung von „Sprüngen" der Werte in der Kriteriumsfunktion. Obwohl im Gegensatz zu clusteranalytischen Verfahren bei der Anwendung finiter Mischverteilungsmodelle eine Betonung auf dem Aspekt statistischer Modellbildung und -testung liegt, ist auch hier die Ermittlung der Anzahl von Subpopulationen nicht unproblematisch.

Abschließend sollen noch zwei Probleme clusteranalytischer Verfahren angesprochen werden, die primär die Ausgangsdaten betreffen: die Auswahl der Merkmale und die Berücksichtigung möglicherweise unterschiedlicher Maßeinheiten der Merkmale. Für das Ergebnis der Klassifikation ist es entscheidend, welche Merkmale in die Berechnung einbezogen werden. Die Merkmale müssen daher vor dem theoretischen Hintergrund der Fragestellung ausgewählt und somit bedeutsam im Hinblick auf das Ziel der Klassifikation sein. Zudem spielt die Reliabilität, mit der die Merkmale erhoben werden, eine Rolle. Milligan (1980) zeigt, daß clusteranalytische Verfahren sensitiv sind für Zufallsdimensionen. Die Hinzunahme von stark fehlerbehafteten Merkmalen kann also schädlich sein für den Nutzen der ermittelten Clusterstruktur. Werden die Merkmalsvariablen in sehr unterschiedlichen Einheiten gemessen, so dominieren die Merkmale mit den größten Varianzen die Distanzmaße (6). Eine scheinbare Lösung besteht in der Standardisierung aller Merkmalsvariablen auf gleiche Varianz. Eine Umgewichtung der Variablen kann die Clusterlösung jedoch hochgradig verändern (vgl. Hand, 1981, S. 159) oder aber die in den Daten vorhandene Struktur aufheben (vgl. Everitt, 1993, S. 39). Auch hier sollte dem Gegenstandsbereich angemessen entschieden werden, ob eine implizite Gewichtung der Merkmale gewünscht oder zumindest zulässig ist oder aber ausgeschlossen werden sollte (zu diesem und anderen Problemen der Gewichtung, vgl. auch Eckes & Roßbach, 1980).

6 Weiterführende Literatur

Die hier erörterten und zahlreiche weitere Verfahren werden etwa von Blashfield und Aldenderfer (1988), Eckes und Roßbach (1980), Everitt (1993), Hand (1981) und Oldenbürger (1983) dargestellt. Moosbrugger und Frank (1992) geben neben einer primär praxisorientierten Übersicht über zahlreiche methodische Ansätze vier ausführliche Anwendungsbeispiele aus verschiedenen Bereichen der Psychologie. Bei Bock (1974) findet sich eine umfassende Behandlung eines großen Spektrums von Ansätzen mit ihren formalen Grundlagen. Blashfield (1980) bietet eine historische und wissenschaftssoziologische Beschreibung der Entstehung und Verbreitung hierarchischer clusteranalytischer Ansätze. Marradi (1990) diskutiert wissenschaftstheoretische Aspekte von Klassifikationen und Typologien allgemein. Computerprogramme für die praktische Anwendung clusteranalytischer Verfahren werden von Everitt (1993) und Moosbrugger und Frank (1992) aufgelistet und kurz kommentiert.

Literaturverzeichnis

Atlas, R. S. & Overall, J. E. (1994). Comparative evaluation of two superior stopping rules for hierarchical cluster analysis. *Psychometrika, 59*, 581–591.

Blashfield, R. K. (1980). The growth of cluster analysis: Tryon, Ward, and Johnson. *Multivariate Behavioral Research, 15*, 439–458.

Blashfield, R. K. & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 447–473). New York: Plenum Press.

Bock, H. H. (1974). *Automatische Klassifikation: Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten.* Göttingen: Vandenhoeck & Ruprecht.

Eckes, T. & Roßbach, H. (1980). *Clusteranalysen.* Stuttgart: Kohlhammer.

Everitt, B. S. (1993). *Cluster analysis.* London: Arnold.

Giegler, H. & Rost, J. (1993). Typenbildung und Responsesets beim Gießen-Test: Clusteranalyse versus Analyse latenter Klassen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 137–152.

Hand, D. J. (1981). *Discrimination and classification.* Chichester: Wiley.

Hubert, L. J. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association, 69*, 698–704.

Hubert, L. J. & Baker, F. B. (1977). The comparison and fitting of given classification schemes. *Journal of Mathematical Psychology, 16*, 233–253.

Hubert, L. J. & Schultz, J. (1975). Hierarchical clustering and the concept of space distortion. *British Journal of Mathematical and Statistical Psychology, 28*, 121–133.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Marradi, A. (1990). Classification, typology, taxonomy. *Quality & Quantity, 24*, 129–157.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika, 44*, 343–346.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325–342.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46*, 187–199.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Moosbrugger, H. & Frank, D. (1992). *Clusteranalytische Methoden in der Persönlichkeitsforschung.* Bern: Huber.

Morey, L. C. & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance aggreement. *Educational and Psychological Measurement, 44*, 33–37.

Oldenbürger, H. A. (1983). Clusteranalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (= Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 4, S. 390–439). Göttingen: Hogrefe.

Oldenbürger, H. A. & Schwibbe, M. (1980). Konstruktive Kritik des Einsatzes dimensionaler Dekompositionsverfahren für EEG-Frequenzkomponenten. In S. Kubicki, W. M. Herrmann & G. Laudahn (Hrsg.), *Faktorenanalyse und Variablenbildung aus dem Elektroenzephalogramm* (S. 47–60). Stuttgart: Fischer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 263–244.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329–350.