Psychologische Beiträge, Band 25, 1983, S. 349–362

Probleme und Möglichkeiten bei der Bewertung von Clusteranalyse-Verfahren

III. Appendix: Kurzbeschreibung der verbreitetsten Clusteranalyse-Algorithmen

W. Schneider[1] und D. Scheibler[2]

Zusammenfassung, Summary, Résumé

Es wird eine relativ einfach gehaltene Kurzcharakteristik derjenigen Clusteranalyse-Algorithmen gegeben, die aufgrund eines Literaturüberblicks (SCHNEIDER & SCHEIBLER 1983a) als die in der Forschung hauptsächlich benutzten Verfahren einzustufen sind. Die Kurzbeschreibung verzichtet im wesentlichen auf statistische Details und verfolgt speziell das Ziel, dem Leser eine Vorstellung von Gemeinsamkeiten und Unterschieden in der Funktionsweise von hierarchischen Clusteranalysen, Optimierungs- bzw. Partitionierungstechniken, Dichteverfahren, ‚Clumping Techniques' und anderen Prozeduren zu geben.

On the evaluation of clustering algorithms, III. Appendix: A description of the most popular algorithms

This paper presents a summary of 18 clustering algorithms most frequently applied in research (cf. SCHNEIDER & SCHEIBLER 1983a). Only a short description of each procedure is provided which aims at highlighting the basic differences and commonalities of hierarchical clustering algorithms, iterative partitioning methods, mode seeking techniques, clumping techniques, and other procedures.

Problèmes et cossibilités d'évaluation de procédés des analyses Cluster, III. Appendix: Courte description des Algorithmes analyse Cluster les plus répendues

Les Algorithmes analyse Cluster qui sont décritent par (Schneider & Scheibler 1983) comme étant les procédés les plus répendus dans la recherche sont relatés ici de façon courte. Le description chématique exclue l'énnumération des détails statistiques et à pour but essentiel de transmettre au lecteur, entre autre, une représentation des rapports et des différences dans le mode de fonction des analyses Cluster hiérarchiques, des techniques d'optimation, et des procédés de population «Clumping techniques» etc. (Dr. W. Lohr)

Wenn im folgenden der Versuch gemacht wird, die Funktionsweise von einigen in der Forschung relativ häufig benutzten Clusteranalyse-Algorithmen summarisch zu beschreiben, so geschieht dies vor allem aus der Absicht heraus, einen vorläufigen Orientierungsrahmen bereitzustellen, der

1 Dr. Wolfgang Schneider, Max-Planck-Institut für psychologische Forschung, Leopoldstr. 24, 8000 München 40.
2 Dipl.-Psych. Dieter Scheibler, Neugasse 7, 6900 Heidelberg.

statt technischer Details überwiegend inhaltlich gehaltene Erläuterungen bietet. Es geht also insbesondere darum, dem mit Clusteranalysen nicht allzu sehr vertrauten Leser die Möglichkeit zu geben, die bei SCHNEIDER & SCHEIBLER (1983a, b) aufgeführten Algorithmen im Hinblick auf ihre Unterschiede und Gemeinsamkeiten besser beurteilen zu können. Angesichts der weit verbreiteten terminologischen Konfusionen mag es aber auch für den Clusteranalyse-Experten hilfreich sein, sich die unterschiedlichen Etikettierungen einzelner Algorithmen zu vergegenwärtigen.

Es darf allerdings kein Zweifel daran bestehen, daß der hier vorgelegte Überblick die z. T. ausgezeichneten Detaildarstellungen der einzelnen Algorithmen (s. ANDERBERG 1973; ECKES & ROSSBACH 1980; SPÄTH 1975; STEINHAUSEN & LANGER 1977; VOGEL 1975 u. a.) nicht ersetzen kann und soll. Anhand der hier vorgenommenen Vorstrukturierung ergibt sich für den potentiellen Anwender von Clusteranalysen die Möglichkeit, eine erste Selektion der theoretisch interessanten Algorithmen vorzunehmen. Es empfiehlt sich jedoch in jedem Falle, detailliertere Informationen zu diesen Prozeduren einer der erwähnten Monografien zur Clusteranalyse zu entnehmen.

I. Grobeinteilung der verfügbaren Clusteranalyse-Algorithmen

Nach EVERITT (1974, S. 7 ff.) lassen sich Clusteranalysen ganz bestimmten Kategorien zuordnen, die im folgenden näher charakterisiert werden sollen:

1. Hierarchische Clusteranalysen

Es lassen sich bei diesen Techniken agglomerative und divisive Formen unterscheiden. Ausgehend von N Elementen (Personen) fügen agglomerative Verfahren stufenweise einzelne Elemente oder bereits geschaffene Cluster zu immer größeren Clustern (Klassen) zusammen, bis zum Schluß zwei Cluster resultieren, die zusammengefaßt die Gesamtmenge N der Elemente enthalten. In den Sozialwissenschaften sind diese Verfahren am gebräuchlichsten. Divisive Verfahren gehen genau umgekehrt vor, indem sie die Gesamtmenge schrittweise in immer kleinere Cluster aufspalten, z. B. in der Weise, daß bei jedem Schritt jeweils ein Cluster in zwei homogenere geteilt wird, bis zum Schluß jedes Element ein separates Cluster bildet. Die meisten divisiven Verfahren eignen sich nur für binäre oder dichotomisierte Daten und sind deswegen von geringerer praktischer Bedeutung als die agglomerativen hierarchischen Verfahren.

2. Optimierungs-Partitionierungs-Techniken

Es werden disjunkte (nicht überlappende) Cluster dadurch gebildet, daß ein festgelegtes Kriterium optimiert wird. Die wesentlichen Un-

terschiede zu den hierarchischen Verfahren bestehen darin, daß zum einen schlechte Anfangspartitionierungen korrigiert, die Elemente demnach re-allokiert werden können, und zum anderen die gewünschte Klassenzahl vom Untersucher a priori festzulegen ist.

3. Dichteverfahren (Mode Seeking Techniques)

Die Elemente werden als Punkte im metrischen Raum aufgefaßt und solche Regionen als Cluster interpretiert, in denen sich die Elemente besonders dicht konzentrieren.

4. ‚Clumping Techniques'

Im Unterschied zu den vorher erwähnten Verfahren, die lediglich disjunkte Klassen zulassen, können mit dieser Methode überlappende Cluster (Clumpen) gebildet werden.

5. Andere Verfahren

Hier sind Methoden gemeint, die keiner der vier anderen Kategorien eindeutig zugeordnet werden können.

II. Kurzbeschreibung der bei Schneider & Scheibler (1983a,b) erwähnten Clusteranalyse-Techniken

1. Hierarchisch agglomerative Verfahren

Alle nachfolgend aufgeführten hierarchischen Methoden verfahren nach dem selben Prinzip und unterscheiden sich allein in der Definition der Distanz zwischen zwei Clustern. Alle Verfahren beginnen damit, daß sie zunächst jene beiden Elemente zu einem Cluster verbinden, die am nächsten zueinander liegen. Im folgenden werden schrittweise immer jene zwei Cluster zu einem neuen Cluster zusammengefaßt (agglomeriert), die die geringste Distanz zueinander aufweisen (wobei auch Einzelelemente als Cluster definiert sind). Mit jedem Schritt reduziert sich also die Zahl der Cluster um 1, und es entstehen immer größere aber auch heterogenere Cluster, bis alle Elemente in einem großen Cluster verbunden sind.

Die Definition der Distanzen zwischen Clustern läßt sich für alle relevanten agglomerativen Verfahren in einer Rekursionsformel ausdrükken:

Wenn die Cluster P und Q zu einem neuen Cluster (P+Q) verbunden werden, dann ist die Distanz $d_{r(p+q)}$ zwischen dem Cluster R und (P+Q):

$$d_{r(p+q)} = \alpha_p d_{rp} + \alpha_q d_{rq} + \beta d_{pq} + \gamma |d_{rp} - d_{rq}|.$$

Die verschiedenen agglomerativen Verfahren unterscheiden sich allein in der Ausprägung der Parameter α, β und γ. Tabelle 1 gibt die Parametergröße für die nachfolgend genauer beschriebenen Verfahren wieder.

Tabelle 1:

Ausprägungen der Rekursionsformel-Parameter α, β und γ für die einzelnen hierarchisch-agglomerativen Cluster-Algorithmen

Clusteranalyse-Verfahren	α_p	α_q	β	γ
Single Linkage	1/2	1/2	0	− 1/2
Complete Linkage			0	1/2 (bei Distanzkoeffizient) − 1/2 (bei Ähnlichkeitskoeff.)
Average Linkage	$\frac{N_p}{N_p + N_q}$	$\frac{N_q}{N_p + N_q}$	0	0
Median	1/2	1/2	− 1/4	0
Centroid	$\frac{N_p}{N_p + N_q}$	$\frac{N_q}{N_p + N_q}$	$\frac{-N_p N_q}{(N_p + N_q)^2}$	0
WARD	$\frac{N_p + N_r}{N_p + N_q + N_r}$	$\frac{N_q + N_r}{N_p + N_q + N_r}$	$\frac{-N_r}{N_p + N_q + N_r}$	0
Flexible beta	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	β	0
McQUITTY	1/2	1/2	0	0

a) Single-Linkage-Methode (Nearest Neighbour)

Bei dieser Technik ist die Distanz zwischen zwei Clustern durch den Abstand jener zwei Elemente aus den beiden Clustern definiert, die die

geringste Distanz zueinander aufweisen. ‚Single Linkage' wurde ursprünglich von SNEATH (1957) als nichthierarchisches Verfahren eingeführt, bei dem die Cluster durch die Vorgabe eines Grenzwertes für die Distanz (bzw. Ähnlichkeit) zwischen jenen Elementen definiert sind, die im selben Cluster integriert sind. Durch systematische Verkleinerung (bzw. Vergrößerung) des Grenzwertes werden unterschiedliche Clusterlösungen produziert.

Das „Nearest Neighbour"-Verfahren stellt eine Verallgemeinerung der ursprünglichen Single-Linkage-Methode dar. Heute wird kaum noch zwischen beiden Verfahren unterschieden und beide Bezeichnungen synonym für die allgemeine Version verwendet (so bei allen gebräuchlichen Computerprogrammen, die dieses Verfahren enthalten).

Da das Single-Linkage-Verfahren dazu neigt, jeweils benachbarte Elemente aneinander zu hängen oder zu verketten (chaining effect), liefert es selten kompakte Cluster und ist insbesondere dann ungeeignet, wenn man Cluster mit möglichst geringen Varianzen anstrebt.

Zudem ist der Chaining-Effekt in der Regel unerwünscht, weil schlecht getrennte Cluster (z. B. bei sich überlappenden Klassen) nicht „erkannt" werden.

Das Verfahren ist jedoch dann zu empfehlen, wenn es mehr auf den Zusammenhang der Elemente als auf deren Ähnlichkeit ankommt, d. h. auch langgestreckte, bananen-, schlangen- oder ringförmige Cluster sinnvoll erscheinen (z. B. wenn Simplex- oder Circumplexstrukturen gesucht werden).

Es ist typisch für ‚Single Linkage', daß bei den unteren und mittleren Hierarchiestufen viele Elemente eigene Cluster bilden; selbst auf höheren Hierarchiestufen finden sich häufig einzelne Elemente. Dies ist bei der Suche kompakter Cluster unerwünscht, macht das Verfahren jedoch noch interessant für die Suche nach Elementen mit ungewöhnlichen Meßwertkombinationen (outlier), also solchen Objekten, die sich von den anderen in irgendeiner Weise abheben.

b) Complete-Linkage-Methode (Furthest Neighbour)

Bei diesem Ansatz ist die Distanz zwischen zwei Clustern als der Abstand zwischen den am weitesten auseinanderliegenden Elementen aus den jeweiligen Clustern definiert. Im Vergleich zu ‚Single-Linkage' liefert dieses Verfahren recht homogene Klassen. Es bildet vor allem auf den unteren Hierarchieebenen kleine, kompakte Cluster, kann darum aber auf den höheren Hierarchiestufen dem Kriterium der Varianzminimierung innerhalb der Cluster nicht immer gerecht werden. Mit Sicherheit vermeidet es den ‚chaining'-Effekt und outlier. Ursprünglich war auch ‚Complete Linkage' (wie Single Linkage) nicht als hierarchisches Verfahren konzi-

piert, sondern arbeitete mit systematisch variierten Grenzwerten für die zulässigen Distanzen (bzw. Ähnlichkeiten) innerhalb der Cluster. Auch hier gilt, daß heute die Bezeichnung ‚Complete Linkage' und ‚Furthest Neighbour' synonym für die verallgemeinerte Version als hierarchisches Verfahren verwendet werden.

c) Average-Linkage-Methode

Diese Methode kann zwischen den beiden Extremen ‚Single-Linkage' und ‚Complete-Linkage' eingeordnet werden. Bei diesem Verfahren ist die Distanz zwischen zwei Clustern definiert als arithmetisches Mittel der Distanzen zwischen allen Paaren von Elementen, die man aus den Elementen der beiden bilden kann.

Das Verfahren hat die Eigenschaft, großen Clustern mehr Gewicht zu geben als kleinen, so daß schwach repräsentierte Klassen leicht übersehen werden können. Dies liegt daran, daß der Mittelwert eines durch die Fusion zweier Cluster gebildeten neuen Clusters zum Zentrum des jeweils größeren Ausgangsclusters hin verschoben ist; dies umso mehr, je unterschiedlicher die Anzahl der Elemente in den beiden Clustern ist.

Um diesen Effekt zu unterbinden, wurde eine Abwandlung empfohlen, bei der die verfahrensimmanente Gewichtung durch die Clustergröße aufgehoben wird. Diese Version entspricht mathematisch exakt dem Verfahren nach McQUITTY, doch läuft es in der Literatur manchmal als ‚Weighted Average Linkage', manchmal auch als ‚unweighted', um es von der einfacheren Average-Linkage-Methode unterscheiden zu können. Dies ist letztlich nicht gelungen, weil nun beide Verfahren einmal so und einmal andersherum bezeichnet werden. Gerade bei diesem Verfahren herrscht hinsichtlich der Terminologie besondere Konfusion. Dies rührt daher, daß SOKAL & MICHENER (1958) bei der Einführung dieser Unterscheidung jenes Verfahren unter taxonometrischen Gesichtspunkten als ‚unweighted' bezeichneten, bei dem explizit (durch Abwandlung des Algorithmus) eine Gewichtung der Mittelwerte durchgeführt wird (um die implizite Gewichtung aufzuheben). Im Zweifelsfall sei empfohlen, den Verfahrenstyp anhand der Rekursionsformel zu identifizieren, die häufig bei Programmbeschreibungen zu finden ist.

d) Centroid-Methode

Das Prinzip dieser Methode läßt sich am leichtesten geometrisch erklären, wenn man sich die Cluster als Punkte-Haufen oder -Wolken in einem euklidischen Raum vorstellt. Die Distanz zwischen zwei Clustern wäre dabei als die Distanz zwischen den Centroiden dieser Cluster definiert.

Das Centroid-Verfahren hat eine gewisse Ähnlichkeit mit ‚Average

Linkage'. Im Gegensatz zum Centroid-Verfahren, bei dem ein Cluster allein durch seinen Centroid (Mittelwert im mehrdimensionalen Raum) repräsentiert ist, berücksichtigt die Average-Linkage-Methode bei der Bildung eines neuen Clusters auch die Varianz der beiden fusionierten Cluster. Das Centroid-Verfahren bildet demnach weniger kompakte Cluster.

Die Centroid-Methode zeigt ähnlich wie ‚Average Linkage' die Eigenschaft, daß bei der Fusion zweier (extrem) unterschiedlich großer Cluster das Centroid des neu gebildeten Clusters (sehr viel) näher bei dem größeren Ausgangscluster liegt. Dies führt dazu, daß kleine Klassen von großen mehr oder weniger stark „aufgesogen" werden.

e) Median-Methode (nach GOWER)

Die Distanz d (R,P+Q) zwischen irgendeinem Cluster R und dem Cluster, das durch die Fusion von P und Q gebildet wird, ist bei diesem Verfahren als die Distanz vom Centroid von R zum Mittelpunkt der Strecke definiert, die die Centroide von P und Q verbindet. Dieses Verfahren stellt eine Modifikation der Centroidmethode dar, bei der kleinen Clustern das gleiche Gewicht gegeben wird wie großen. Trotzdem neigt auch diese Methode dazu, heterogene Cluster zu bilden und weiterhin stärker als das Centroid-Verfahren dazu, ‚outlier' zu produzieren.

f) Clusteranalyse nach WARD

Bei diesem Verfahren werden schrittweise jeweils diejenigen Cluster fusioniert, die den geringsten Zuwachs zur Gesamtvarianz innerhalb der Cluster (= Summe der Varianzen der gebildeten Cluster) erzeugen. Dieser Ansatz zielt also nicht darauf ab, immer die ähnlichsten Gruppen zu fusionieren, sondern möglichst homogene Cluster zu bilden, was für viele Anwendungen in der psychologischen Forschung evident erscheint. Das Verfahren läßt große wie auch kleine Cluster zu und vermeidet die Bildung von ‚outliern'. Sehr stabile Clusterlösungen finden sich vor allem bei (annähernd) symmetrisch und eingipflig verteilten Gruppen.

g) LANCE & WILLIAMS' Flexible beta-Methode (Flexible Strategy)

LANCE & WILLIAMS (1966, 1967) haben auf der Grundlage der allgemeinen Rekursionsformel für agglomerative Verfahren eine Methode entwickelt, deren Eigenschaften durch die freie Wahl des β-Parameters (im Bereich zwischen 1 und -1; siehe Tab. 1) variiert werden kann. Interessant sind dabei nur β-Werte zwischen 0 und -1, da das Verfahren umso stärker zur Kettenbildung (Aneinanderreihung einzelner Elemente) neigt, je mehr sich β dem Wert 1 nähert. Bei einem β um -0.5 ($\pm 0{,}2$) zeigt das Verfahren ähnliche Eigenschaften wie die Methode von WARD. Je näher β bei -1 liegt, desto stärker neigt das Verfahren dazu, auf unteren bis mittleren

Hierarchiestufen die Elemente zu einzelnen sehr kleinen kompakten Gruppen zusammenzufassen (ähnlich Complete Linkage). LANCE & WILLIAMS empfehlen zwar einen β-Wert von $-1/4$, bei unserer eigenen Untersuchung liefert das Verfahren bei $\beta = -1/2$ allerdings bessere Ergebnisse.

h) Die Methode nach McQUITTY

Dieses Verfahren ist identisch mit der Flexible beta-Methode für $\beta = 0$ und mit dem von SOKAL & MICHENER (1958) als ‚Unweighted Average Linkage' benannten Verfahren. Es hat ähnliche Eigenschaften wie die Average-Linkage-Methode, mißt jedoch kleinen wie großen Clustern gleiches Gewicht bei, was in z. T. deutlich besseren Clusterlösungen resultiert.

2. Optimierungs-Partitionierungs-Techniken

a) Optimierungstechnik nach MacQUEEN

Bei dieser Partitionierungs-Prozedur wird versucht, die Spur einer Matrix W (Summe der Elemente in der Hauptdiagonalen) zu minimieren, wobei W die gepoolte Kovarianz-Matrix der Abweichungsquadrate und Kreuzprodukte innerhalb der Gruppen darstellt. Die Logik des Vorgehens ist evident, da im Prinzip die Varianz innerhalb der Cluster minimiert wird, was zu möglichst homogenen Gruppen führt.

Die Anzahl der Cluster, die gebildet werden sollen, muß ad hoc vorgegeben werden. Außerdem benötigt das Verfahren eine vorläufige Clusterlösung als Startkonfiguration. Danach wird versucht, diese Anfangspartitionierung nach folgendem heuristischen Prinzip zu verbessern: Für alle Dateneinheiten (Elemente, Fälle, Personen) werden nacheinander die Distanzen zu allen Cluster-Centroiden berechnet. Wenn der am nächsten gelegene Centroid nicht zu dem Cluster gehört, in dem das betreffende Element liegt, wird es dem anderen (näher gelegenen) Cluster zugeordnet (relociert). Danach werden die Centroide des Clusters, das ein Element verloren hat sowie dessen, das eines dazu bekommen hat neu berechnet, und der Vorgang wird für die weiteren Elemente fortgesetzt. Danach beginnt das Verfahren wieder mit dem ersten Element und setzt die Prozedur so lange fort, bis keine Elemente mehr ausgetauscht werden müssen, d. h. die Summe der Abweichungsquadrate nicht weiter abnimmt. Dieses Verfahren ist u. a. in dem Programm MIKCA von McRAE (1971) realisiert. Es gibt eine Reihe ähnlicher Verfahren, die entweder im Optimierungsalgorithmus oder in der Definition der Distanz zwischen Elementen und/oder Clustern abweichen. Eine gute Übersicht findet man bei ANDERBERG (1973).

b) Das Programm RELOCATE von WISHART (1975)

RELOCATE enthält als Kernstück den Algorithmus von McQUEEN. Es bietet die Möglichkeit, eine Anfangspartitionierung vom Programm generieren zu lassen (Zufallseinteilung) oder die Lösung von einer vorgeschalteten anderen Clusterprozedur zu übernehmen. Wenn RELOCATE eine verbesserte Lösung gefunden hat, besteht die Möglichkeit, die Zahl der Cluster durch Agglomeration nach dem WARDschen Prinzip schrittweise zu reduzieren, wobei nach jeder Fusion zweier Cluster die Optimierungsstrategie nach McQUEEN abläuft.

c) Optimierungstechnik nach SPAETH (1975)
Das Programm KMEANS

Dieses Programm baut ebenfalls auf dem Ansatz von McQUEEN auf, berechnet aber nicht die Distanzen der Elemente zu den Centroiden der Cluster, sondern zu deren Medianen. (Dadurch soll das Verfahren auch für Daten auf Rang- oder Ordinalskalenniveau anwendbar sein). Es wird sodann die Summe der quadrierten Distanzen zwischen den Medianen und den dazugehörigen Elementen minimiert (was streng genommen doch Intervallskalenniveau voraussetzt). SPÄTH empfiehlt, der Prozedur eine Standardisierung der Daten vorzuschalten, und zwar der Gestalt, daß alle Variablen so (linear) transformiert werden, daß sie das gleiche range bzw. gleiche Minima und Maxima aufweisen. Die Methode verhält sich weitestgehend wie RELOCATE, wenn es mit denselben Startkonfigurationen beginnt. Es hat den Nachteil, daß es nicht wie RELOCATE eine automatische Agglomeration benachbarter Cluster erlaubt.

d) Die Prozedur EUCLID von WISHART (1975)

Das hier angesprochene Verfahren verwendet die CAUCHY-Methode, um die Quadratsumme S der euclidischen Distanz durch schrittweise Verbesserung einer kontinuierlichen Klassifikationsmatrix y zu minimieren. Das Element Y_{ik} kann dabei als Wahrscheinlichkeit interpretiert werden, daß der i-te Punkt zum k-ten Cluster gehört.

EUCLID generiert für eine vorgegebene Zahl von Clustern gleichverteilte Wahrscheinlichkeit für Y, die so skaliert sind, daß die Summe der „Wahrscheinlichkeiten" (fractions) gleich 1 ist. Durch eine Exponential-Transformation von Y ergeben sich neue Klassifikationsvariablen (kontinuierlich-differenzierbar und monoton-ansteigend).

Eine Folge von Klassifikationsmatrizen wird nun durch sukzessive Applikation der CAUCHY-Korrekturen so generiert, daß bei jedem Schritt ein neues Minimum von S gefunden wird. Die Sequenz wird bei derjenigen Interaktion beendet, die für jeden Punkt i einen Wert k so kombiniert, daß Y_{ik} wenigstens .999 ergibt und die partielle Ableitung von S im Hinblick auf Y_{ik} negativ wird.

Danach wird wie bei RELOCATE versucht, durch Re-Allocation eine weitere Reduktion von S zu erreichen.

3. Dichte-Verfahren (Mode-Seeking Techniques)

(a) Mode Analysis nach WISHART

Diese Technik läßt sich der Gruppe der Dichteverfahren zuordnen. Der Algorithmus sieht vor, daß zunächst um jedes Element je ein Raum mit dem Radius R aufgespannt und dann untersucht wird, wie viele benachbarte Elemente ebenfalls in dieser Sphäre liegen. Wenn der Raum bei bestimmten Individuen K und mehr Elemente enthält, werden diese zu Dichtezentren erklärt. Der Parameter R wird sukzessive vergrößert, wobei mit neuen Dichtezentren verschiedenartig verfahren werden kann: wird das neue Zentrum von allen andernen Dichtepunkten durch eine Distanz getrennt, die größer als R ist, muß ein neues Cluster initiiert werden; ist die Distanz dagegen kleiner als R, ergeben sich verschiedene Möglichkeiten:

1) wenn sich der neue Punkt im Bereich von Dichtezentren befindet, die einen gemeinsamen Clusterkern besitzen, wird er diesem angeschlossen;
2) liegt er im Bereich von Dichtezentren, die unterschiedlichen Clusterkernen angehören, werden die betreffenden Cluster zusammengeschlossen.

Weiterhin wird bei jedem Zyklus die geringste Distanz D zwischen Dichtezentren berechnet, die zu bestimmten Clustern gehören, und letztere mit einem bestimmten Grenzwert verglichen: falls D diesen Grenzwert nicht übersteigt, werden die betreffenden Cluster kombiniert.

(b) Das Verfahren der Automatischen Klassifikation (AUKLA) nach FABER & NOLLAU

AUKLA versucht im vollständigen oder reduzierten Datenraum (aufgespannt durch eine begrenzte Zahl von Faktoren) „geometrisch vorgegebene Punktehäufungen innerhalb dieses Datenraums festzustellen und die jeweils zusammengehörigen Elemente zu ermitteln" (FABER & NOLLAU 1969, S. 8).

Die Grundidee des Verfahrens geht auf SCHNELL (1964) zurück. Dieser betrachtete die p Meßwerte einer Person (p = Anzahl der Variablen) als Erwartungswerte einer p-dimensionalen Normalverteilung (zum besseren Verständnis sei auf die analoge Betrachtungsweise in der Klassischen Testtheorie hingewiesen, bei der man annimmt, daß aufgrund eines Meßfehlers der wahre Wert mit einer bestimmten Wahrscheinlichkeit – definiert durch eine Normalverteilung mit dem Meßwert als Mittelwert – in

einem gewissen Intervall um den gemessenen Wert liegen kann).

Die Verteilungen der Meßwerte mehrerer Personen überlagern sich mehr oder weniger stark, wobei das jeweilige Ausmaß davon bestimmt wird, wie dicht die Werte im Datenraum beieinanderliegen und wie groß die Varianzen dieser Verteilungen sind. Ist die Varianz bei allen Personen gleich Null, können sich die Verteilungen verschiedener Individuen nicht überlagern (es sei denn, sie hätten identische Meßwerte in allen Variablen).

Erhöht man nun die Varianz allmählich (und für alle Personen gleichmäßig), dann werden zunächst nur solche Punkte im Datenraum (Personen) deutliche Überlagerungen ihrer Verteilungen zeigen, die dicht beieinander liegen.

Das Verfahren AUKLA berechnet eine sog. Belegfunktion, die neben einem Haupt- meist mehrere Nebenmaxima (Gipfel) aufweist. Diese Maxima deuten auf stärkere Überlagerungen von Verteilungen und damit auf Punktkonzentrationen im Datenraum hin. Sie können als Clusterzentren interpretiert werden. Je größer nun die Varianz der Verteilungen ist, desto eher kommt es auch zu ausgeprägten Überlagerungen von Verteilungen um Punkte, die weiter voneinander entfernt sind. Eine stetige Vergrößerung der Varianz führt also auch gleichzeitig zu einer Verringerung der Anzahl von Maxima in der Belegfunktion, die (in der Endstufe) auf N = 1 reduziert wird. Wenn nun trotz stärkerer Veränderungen der Varianz die Anzahl der Maxima lange Zeit konstant bleibt, deutet dies darauf hin, daß eine (relativ) optimale Anzahl von Clustern gefunden ist, von der man bei der Ergebnisinterpretation ausgehen sollte.

Das Verfahren besticht durch seine Eleganz, hat sich in der Praxis jedoch kaum durchsetzen können, da es selbst moderne Großrechenanlagen bei einer Variablen- bzw. Faktorenzahl größer als 3 und bei Stichproben größer als N = 100 rasch in der Speicherkapazität überfordert und zudem äußerst rechenintensiv und damit unwirtschaftlich ist.

4. ‚Clumping techniques'

(a) Das Taxonome-Programm (TAXO) von CATTELL & COULTER

Obwohl dieses Verfahren im Grunde genommen nicht exakt unter Kategorie 4 subsummiert werden kann, wird es dennoch nicht hier dargestellt, weil es als einzige der in diesem Kontext berücksichtigten Prozeduren überlappende Cluster zuläßt.

Die konzeptuelle Grundlage ist in dem Typus-Begriff von CATTELL zu sehen: Als ‚Typen' werden relative Maxima in einer mehrdimensionalen Häufigkeitsverteilung definiert. Im ersten Analyseschritt werden aus einer Korrelationsmatrix sich überlappende sog. ‚Phenomenal clusters' (PhCl) als Grundeinheiten extrahiert, in denen Elemente zusammengefaßt sind, die

sich oberhalb eines beliebig zu wählenden Cutoff-Werts (Grenzwerts) ähneln. Im zweiten Schritt lassen sich die Überlappungsbereiche zweier oder mehrerer PhCl als sog. ‚Nuclear cluster' bestimmen, die wiederum die Voraussetzung für die im dritten Schritt erzeugten ‚Segregates' (Verkettungen einzelner oder mehererer PhCl) bilden, in denen nicht unbedingt alle Elemente einander ähnlich sein müssen. Eine Optimierung der Gruppen-Extraktion kann durch systematische Manipulation des Cutoff-Wertes sowie des Parameters der Überlappungsgröße erreicht werden (genauere Vorschläge zur Optimierung finden sich bei BAUMANN 1971, S. 115 f. und S. 157 f.).

5. Sonstige Verfahren

(a) Die Konfigurationsfrequenzanalyse (KFA):

Das Verfahren analysiert eine endliche Zahl von Variablenmustern (Konfigurationen) für kategoriale Daten (bei n dichotomen Variablen sind beispielsweise 2^n Konfigurationen möglich). Ein ‚Typus' läßt sich als signifikant überfrequentierte Konfiguration charakterisieren, ein ‚Anti-Typus' dementsprechend als signifikant unterfrequentierte Konfiguration bezeichnen. Personen, die keinem Typus (bz. Antitypus) angehören, werden nicht erfaßt. Vor der genaueren Analyse der einzelnen Typen wird jedoch jeweils überprüft, inwieweit die beobachteten von den theoretisch zu erwartenden Konfigurations-Frequenzen abweichen. Falls die Nullhypothese (es besteht kein signifikanter Unterschied) nicht widerlegt werden kann, unterbleibt eine weitere Analyse.

(b) Die Clusteranalyse nach LORR et al.:

Bei diesem Verfahren beginnt der Cluster-Suchprozeß mit einem Auflisten der Code-Nummern (Vpn-Nummern) aller Standardwert-Profile, die mit einem bestimmten Profil X oberhalb eines festgelegten Grenzwerts (Signifikanz-Schranke) korrelieren (giltlt für alle X_n). Das Profil mit der längsten Nummernliste eröffnet ein Cluster: es werden sukzessive diejenigen Profile zugeführt, deren durchschnittliche Korrelation mit den Profilen innerhalb des Clusters am größten ist. Ein weiterer Grenzwert gibt an, wann ein Profil (trotz möglicherweise immer noch beträchtlicher Korrelation) als nicht mehr ähnlich gewertet und eliminiert wird, wodurch die Möglichkeit besteht, deutlich separierbare Klassen zu konstruieren. Aus der Residualmatrix wird nun das zweite Cluster initiiert und der Klassifikationsprozeß solange weitergeführt, bis keine Klassen mit wenigstens vier Items mehr gebildet werden können.

(c) Mehrdimensionale automatische Clustersuchstrategie nach Von EYE:

Bei diesem Verfahren wird a priori die Homogenität der Klassen dadurch festgelegt, daß alle Cluster im Testraum als gleich große d-dimensionale Quader (d = Anzahl der Meßwertkategorien) darstellbar sind.

In einem ersten Analyseschritt wird das Konfidenzintervall um das zentrale Element (bei intervallskalierten Daten) bzw. um das Streuungsmaß (bei niedriger skalierten Daten) von allen d Meßwertkategorien errechnet. Danach wird dasjenige Element entweder vorgegeben (bei Hypothesen über die Datenstruktur) oder aber stochastisch aufgesucht, das sich am besten zur Eröffnung eines Clusters eignet, wobei sich dieser Prozeß der sukzessiven Zuordnung solange fortsetzt, bis entweder das Maximum der vorgegebenen Klassenzahl erreicht ist oder aber alle Elemente in einem Cluster zusammengefaßt sind. Um festzustellen, ob die entstandenen Gruppierungen stabil bleiben, wird die Prozedur schließlich mit variierenden Quadergrößen wiederholt. Die Kantenlängen bestimmen die Homogenität der Cluster und können in Analogie zu den Hierarchie-Ebenen der hierarchischen Clusteranalysen interpretiert werden.

Literatur

Anderberg, M. R.: Cluster analysis for applications. New York: Academic Press, 1973.

Baumann, U.: Psychologische Taxometrie – Eine Methodenstudie über Ähnlichkeitskoeffizienten, Q-Clusteranalyse, Q-Faktorenanalyse. Bern: Huber 1971.

Cattell, R. B. & Coulter, M. A.: Principles of behavioral taxonomy and mathematical basis of the taxonomic computer program. British Journal of Mathematical and Statistical Psychology, 1966, 19, 237–269.

Eckes, T. & Rossbach, H.: Clusteranalysen. Stuttgart: Kohlhammer, 1980.

Everitt, B.: Cluster analysis. London: Heinemann, 1974.

Lance, G. N. & Williams, W. T.: A Generalized Sorting Strategy for Computer Classification. Nature, 212, 1966, 218.

Lance, G. N. & Williams, W. T.: A General Theory of Classificatory Systems. Computer Journal, 9, 1967, 373–380.

Lorr, M., Klett, C. J. & D. M. McNair: Syndromes of psychosis. New York: MacMillan, 1963.

McRae, D. J.: MIKCA: A FORTRAN IV iteration k-means cluster analysis program. Behavioral Sciences, 16, 1971, 423–424.

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalyse-Verfahren. I. Ein Überblick über einschlägige Evaluationsstudien. Psychologische Beiträge, 24, 1983, 208–237 (a).

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalysen. II. Ergebnisse einer Monte-Carlo-Studie. Psychologische Beiträge, 24, 1983, 238–254 (b).

Schnell, P.: Eine Methode zur Auffindung von Gruppen. Biometrika 6, 1964, 47–48.

Sneath, P. H. A.: A comparison of different clustering methods as applied to randomly spaced points. Classification Society Bulletin, 1, 1966, 2–18.

Sokal, R. R. & Michener, C. D.: A statistical method for evaluating systematic relationships. Univ. Kansas Scientific Bulletin, 38, 1958, 1409–1438.

Spaeth, H.: Cluster-Analyse-Algorithmen zur Objektklassifikation und Datenreduktion. München, Oldenbourg, 1975.

Steinhausen, D. & Langer, K.: Clusteranalyse. Berlin: de Gruyter, 1977.

Vogel, F.: Probleme und Verfahren der automatischen Klassifikation. Göttingen: Vandenhoeck & Ruprecht, 1975.

Von Eye, A.: Zum Vergleich zwischen der hierarchischen Clusteranalyse nach WARD und MACS, einer mehrdimensionalen, automatischen Clustersuchstrategie. Psychologische Beiträge, 1977, 19, 201–217.

Wishart, D.: CLUSTAN IC user manual. London: Computer Center, 1975.