Psychologische Beiträge, Band 25, 1983, S. 238–254

Probleme und Möglichkeiten bei der Bewertung von Clusteranalyse-Verfahren II. Ergebnisse einer Monte-Carlo-Studie

W. SCHNEIDER[1] und D. SCHEIBLER[2]

Zusammenfassung, Summary, Resumé

Ziel der vorliegenden Untersuchung war es, Aufschluß über die unterschiedliche Qualität hierarchischer und nicht-hierarchischer (partionierender) Clusteranalyse-Verfahren zu gewinnen. Die Reproduktionsgüte beider Clusteranalyse-Varianten wurde anhand von 200 Monte-Carlo-Datensätzen (multivariat normalverteilte Mixturen) zu überprüfen versucht, wobei jeweils unterschiedliche Proportionen der Daten-Elemente klassifiziert werden mußten. Es zeigte sich, daß insgesamt gesehen die hierarchischen Algorithmen nach WARD und LANCE-WILLIAMS am besten dazu in der Lage waren, die vorgegebenen Datenstrukturen zu reproduzieren, andererseits aber die herangezogenen partitionierenden KMEANS-Verfahren nicht schlechter abschnitten, wenn die Lösung der WARD-Technik als Start-Partition vorgegeben wurde.

On the evaluation of clustering algorithms: A monte carlo approach

In this study, a number of hierarchical clustering algorithms and nonhierarchical (i.e. iterative-partitioning) methods were compared with regard to accuracy on the basis of 200 monte carlo data sets. As main results, the two hierarchical procedures by WARD and LANCE-WILLIAMS as well as two nonhierarchical k-means algorithm using WARDs solution as starting seeds proved to be most robust. Although some of the remaining algorithms showed acceptabel recovery values when only a certain proportion of the elements had to be classified, it is recommended to choose the few methods mentioned above for particular applications.

Problème et possibilité pour l'évaluation des procédés d'analyse de cluster

Le but de cette étude est d'obtenir des renseignements sur les différentes qualitées hiérarchiques et non-hierarchiques (partionaires) procèdés d'analyse de Clusters. Le qualité de reproduction des deux variantes d'analyse de Cluster a été relevée et controlée à l'aide de 200 groupes de Monte-Carlo (multivariation, mélange de distribution normale).

Pour chacune des proportions differentes, les éléments de données ont du être classés. On observe, dans l'ensemble, que l'algorithme hierarchique selon Ward et

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

Lance-Williams, est en mesure de reproduire, le mieux, les structures de données impliquées. D'autre part, les procédés appliqués de KMEANS-partionaires ne se détachent pas pour le moins de ces résultats lorsque la solution de la technique de Ward à été, au préalablement, donnée comme situation de départ. (Dr. Lohr)

1. Einleitung

Der vorangegangene Überblick über die vielfältigen Versuche zur Evaluation von Clusteranalyse-Verfahren (SCHNEIDER & SCHEIBLER, 1983 a) hatte ergeben, daß sich bestimmte hierarchisch-agglomerative Cluster-Algorithmen wie z. B. die Verfahren nach WARD bzw. nach LANCE-WILLIAMS wie auch iterativ-partitionierende Verfahren (KMEANS) in einer Reihe unterschiedlicher Monte-Carlo-Studien als relativ robust und akkurat herausgestellt hatten. Einschränkungen der Generalisierbarkeit dieser Befunde sind aber vor allem darin zu sehen, daß a) in der Mehrzahl der Untersuchungen auf identische oder ähnlich konstruierte Simulations-Datensätze zurückgegriffen wurde; b) überwiegend vollständige Lösungen, d. h. Lösungen mit hundertprozentiger Objekterfassung analysiert wurden und c) bislang lediglich drei Studien bekannt sind, in denen hierarchisch-agglomerative und iterativ-partitionierende Clusteranalyse-Verfahren systematisch verglichen wurden. Unglücklicherweise wird in zwei dieser Arbeiten (BAYNE, BEAUCHAMP, BEGOVICH & KRANE 1980; MEZZICH 1978) auf nur zwei Monte-Carlo-Datensätze zurückgegriffen, was die Befunde als wenig generalisierungswürdig erscheinen läßt. Demgegenüber ist der in der dritten Studie (MILLIGAN 1980) verwendete restringierte ‚Mixtur-Ansatz (überlappende Clusterstrukturen waren ausgeschlossen) deshalb zu kritisieren, weil die resultierenden Datensätze für alle CA-Verfahren relativ leicht in ihre konstitutionierenden Teile (Populationen) zu zerlegen waren und keine nennenswerte Varianz erzeugt wurdee. Die vorliegende Monte-Carlo-Studie sollte deshalb einmal dazu dienen, die Generalisierbarkeit der meist mit unrestringiert multivariat normalverteilten Datensätzen erzielten Befunde dadurch zu testen, daß restringierte multivariate normalverteilte Mixturen kreiert wurden, in denen ähnlich wie bei MILLIGAN (1980) überlappende Clusterstrukturen ausgeschlossen waren, die im Hinblick auf den Schwierigkeitsgrad jedoch wesentlich breiter streuten. Für diese Mixturen sollte weiterhin geprüft werden, ob die von EDELBROCK (1979; EDELBROCK & McLAUGHLIN 1980) berichteten Befunde bestätigt werden können, wonach die bei hundertprozentiger Objekterfassung schlecht abschneidenden Verfahren (z. B. average-linkage-, single-linkage-Methode) dann in der Reproduktionsgüte nahe an die Leistungen der WARD-Technik herankom-

men, wenn nicht alle Objekte klassifiziert werden müssen. Schließlich interessierte der Vergleich von hierarchischen und nichthierarchischen Clusteranalyse-Verfahren anhand einer ausreichend großen Stichprobe von Monte-Carlo-Datensätzen, um Hinweise über die Validität der bislang vorliegenden Ergebnisse zu erhalten.

2. Beschreibung des Untersuchungsansatzes

Aufbauend auf früheren Erfahrungen mit Simulationen (SCHEIBLER & SCHNEIDER 1978) und der ausgiebigen Analyse einer Reihe von Monte-Carlo-Studien wurde versucht, nach einer Prozedur zu verfahren, bei der einerseits frühere Fehler und unnötiges Beiwerk vermieden und andererseits viele Aspekte berücksichtigt wurden, die wertvolle Informationen über Eigenschaften eines breiten Spektrums von Clusteranalyseverfahren liefern. Dabei schien es unerläßlich, ein vollkommen neues Untersuchungsdesign zu entwerfen. Wir verzichteten auf die Übernahme eines schon existierenden Ansatzes, versuchten aber die meisten Pluspunkte früherer Untersuchungen zu übernehmen. Trotzdem muß betont werden, daß auch dieser Ansatz nur einen Kompromiß zwischen Idealvorstellungen und der Notwendigkeit darstellt, Abstriche aus technischen, ökonomischen und formalen Gründen vorzunehmen.

2.1. Design

In Anlehnung an die gängigsten Modellvorstellungen der psychologischen Methodenlehre (lineare Statistik) wählten wir als Datenmodell die Multinormalverteilung. Zur Evaluation werden dabei Datensätze generiert, von denen jeder als Stichprobe aus einer Mixtur (Gesamtpopulation) von multinormalverteilten (Teil-)Populationen definiert ist. Die Teilpopulationen unterscheiden sich hinsichtlich der Mittelwerte und Varianzen der einzelnen normalverteilten Variablen sowie im Hinblick auf die Korrelationen zwischen den Variablen und somit auch hinsichtlich der Faktorenstruktur.

Es wurde davon abgesehen, die Variablen mit einem Meßfehler (random noise) zu versehen, wie das einzelne Autoren (z. B. BLASHFIELD 1976) tun, da man die Clusteralgorithmen dadurch vor die prinzipiell unlösbare Aufgabe stellt, die „wahren Werte" zu erraten. Wir meinen, daß eine Evaluation nur mit einer eindeutigen Datenstruktur möglich ist.

Bei der Generierung der Daten durch automatische Ziehung aus den Mischpopulationen ergab sich das Problem, daß bei manchen Datensätzen die Teilstichproben von fast allen Clusteranalyseverfahren problemlos getrennt werden konnten, auf der anderen Seite aber auch solche Daten-

sätze existierten, bei denen sich die Teilstichproben so stark überlappten, daß eine Trennung faktisch überhaupt nicht möglich war. Diese extremen Datensätze wurden im neuen Ansatz dadurch ausgeschlossen, daß zunächst solche Datensätze eliminiert wurden, die durch das Wardsche Verfahren fehlerfrei getrennt werden konnten (Ausschluß von zu leichten Datensätzen). In einem zweiten Schritt unterzogen wir die restlichen Datensätze einer Diskriminanzanalyse und eliminierten all diejenigen, die von diesem Verfahren nicht vollständig separiert werden konnten. Aus dem verbliebenen Pool von Datensätzen wurde nach einem Quotenplan eine Stichprobe von 200 Datensätzen gezogen, bei der darauf geachtet wurde, daß eine möglichst breite Streuung von Datensätzen unterschiedlichen Schwierigkeitsgrades resultierte (s. Tab. 1).

Tabelle 1:

Verteilung der Zahl von Subpopulationen und Variablen bei den analysierten 200 Datensätzen

Anzahl von Subpopulationen	Anzahl von Datensätzen
2	40
3	40
4	40
5	40
6	40
Anzahl von Variablen	**N = 200**
3–10	29
11–15	29
16–20	58
21–25	64
	N = 200

Die Populationsparameter der einzelnen Datensätze bestimmten wir in Anlehnung an BLASHFIELD durch Zufallsstichprobenziehung aus gleichverteilten Populationen, wobei allerdings ein größeres Spektrum an unterschiedlichen Datensätzen angestrebt wurde (Tab. 2).

Tab. 2:
Verteilung der Populationsmuster in der Monte-Carlo-Studie

	kleinster	(theoretisch) größter Wert
Anzahl der Subpopulationen (k) (gleichverteilt)	2	6
Stichprobengröße pro Subpopulation (gleichverteilt)	5	50
Stichprobengröße pro Datensatz (ungefähr gleichverteilt)	10	300
Anzahl der Variablen (p) (verteilt nach Tab. 3)	3	25
Anzahl der Faktoren (Hauptkomponenten)	2	10
Mittelwerte (Erwartungswerte) (gleichverteilt)	45.	60.
Varianzen (gleichverteilt)	5.	30.
Korrelationskoeffizienten r	−1.	+1.
arccos (r) (gleichverteilt)	0	2

2.2. Auswahl der untersuchten Clusteranalyse-Verfahren

Unser Ziel bestand darin, eine möglichst große Zahl der gebräuchlichsten und bewährtesten Clusterverfahren zu evaluieren. Dieses Vorhaben ließ sich aus verschiedenen Gründen nicht voll realisieren. Die große Zahl von generierten Datensätzen, die für eine sinnvolle Monte-Carlo-Untersuchung notwendig ist, legt es nahe, die gesamte Analyse weitgehend zu automatisieren (anders wäre die Arbeit auch nicht zu leisten gewesen). Diese Automatisierung setzt nun aber wieder Grenzen in bezug auf die Auswahl der Clusteranalyse-Verfahren. Wir mußten uns schon aus rein technischen Gründen auf solche Verfahren beschränken, die disjunkte (sich nicht überlappende) Cluster generieren und außerdem keinen allzugroßen Rechenzeitbedarf aufweisen.

Weitere Grenzen ergaben sich aus der Verfügbarkeit einzelner Algorithmen. Bestimmte Verfahren blieben deshalb unberücksichtigt, weil sie zu große Restriktionen im Hinblick auf die Anzahl von Variablen und Objekten setzen, wie z. B. Euclid von WISHART (1975). Das wegen seines

im deutschsprachigen Raum größeren Bekanntheitsgrades grundsätzlich interessante Verfahren der Automatischen Klassifikation von FABER & NOLLAU (eine Variante der partitionierenden Verfahren) sollte ursprünglich auch in den Kreis der zu evaluierenden Verfahren aufgenommen werden. Da mit diesem Verfahren aufgrund seines enorm hohen Rechenzeit- und Speicherplatz-Bedarfs lediglich relativ kleine Datenmatrizen (selten mehr als 100 Elemente bzw. 2 bis 5 Variablen) verarbeitet werden können, war ein Vergleich mit den bei SCHNEIDER & SCHEIBLER (1983 a) aufgeführten, wesentlich schnelleren Algorithmen praktisch nicht möglich. Die wohl umfassendste Sammlung von Clusterverfahren ist in dem schon erwähnten Programmsystem CLUSTAN 1C (WISHART 1975, 1977) enthalten, aus dem alle hierarchisch-agglomerativen Verfahren sowie das partitionierende Verfahren RELOCATE ausgewählt wurden. Ergänzend dazu untersuchten wir noch das von SPÄTH (1975) hoch bewertete Verfahren KMEANS, das vom Algorithmus her sehr viel Ähnlichkeit mit RELOCATE aufweist.

Es wurden demnach folgende zehn Clusteranalyse-Verfahren berücksichtigt:

1) single-linkage-Methode
2) complete-linkage-Methode
3) average-linkage-Methode
4) Centroid-Methode
5) Median-Methode
6) WARD-Methode
7) LANCE-WILLIAMS-flexible-beta-Methode
8) McQUITTYs Ähnlichkeits-Analyse
9) Prozedur RELOCATE
10) Prozedur KMEANS (nach SPÄTH 1975).

Das Verfahren von LANCE-WILLIAMS bietet die Möglichkeit, über einen Parameter – den Beta-Koeffizienten[3] – die Definition der Distanz zwischen zwei Clustern zu variieren. LANCE & WILLIAMS empfehlen einen Beta-Wert von – 0.25. Wir testeten das Verfahren zusätzlich mit einem Parameter-Wert von –0.5.

Die nicht-hierarchischen Verfahren RELOCATE und KMEANS erfordern eine Anfangspartitionierung (vorläufige Clusterlösung). Dabei bieten beide Verfahren die Möglichkeit, entweder eine „Zufallspartitionierung" vom Programm generieren zu lassen oder „von außen" eine vorläufige Clusterlösung einzugeben, die dann von dem Verfahren optimiert wird. Beide Alternativen wurden erprobt, wobei wir bei der zweiten jene

3 Zur Bedeutung dieses Parameters vgl. die Kurzbeschreibung der Clusteranalyse-Verfahren bei Schneider & Scheibler (1983 b).

Clusterlösung vorgaben, die durch das WARDsche Verfahren gewonnen wurde. Da diese Technik im allgemeinen schon sehr gute Lösungen liefert, interessierte die Frage, wieweit diese Lösung überhaupt noch verbessert werden kann.

Zu RELOCATE und KMEANS sei noch erwähnt, daß sich beide Verfahren auf den von MacQUEEN (1967) entwickelten Algorithmus „K-means" stützen (denselben Algorithmus verwendet McRAE (1971) in seinem Programm MIKCA). Der entscheidende Unterschied zwischen beiden Techniken besteht darin, daß als Repräsentant für ein Cluster bei RELOCATE der Centroid gewählt wird, bei KMEANS dagegen der Median des betreffenden Clusters (vgl. im Anhang die Beschreibung beider Verfahren).

Ein weiterer Unterschied ist dadurch gegeben, daß bei KMEANS eine andere Form der Standardisierung der Daten durchgeführt wird. Dieses Verfahren fällt damit etwas aus dem Rahmen, der durch die einheitliche z-Transformation bei den anderen Clusteralgorithmen gesetzt wurde. Eine Angleichung von KMEANS wäre zwar leicht möglich gewesen, doch stellt der Programmautor (SPÄTH 1975) diese Form der Standardisierung als anderen überlegen dar, weshalb wir es dabei beließen und das Verfahren in der Form testeten, wie es von Späth propagiert wird. Durch den Vergleich mit RELOCATE bietet sich damit die Möglichkeit zu prüfen, inwieweit sich die genannten Unterschiede auswirken.

2.3. Zur Wahl des Distanzkoeffizienten

Monte-Carlo-Studien von EDELBROCK (1979) sowie EDELBROCK & MCLAUGHLIN zeigten, daß die Güte von Clusterlösungen sehr stark von der Wahl des Distanz- oder Ähnlichkeitskoeffizienten abhängen kann. Einzelne Verfahren, die mit der Euklidischen Distanz schlechte Ergebnisse lieferten, schnitten mit dem Korrelationskoeffizienten als Ähnlichkeitsmaß bedeutend besser ab. Dies mag zu der Auffassung verleiten, daß das Ziel adäquater Evaluationsstudien die Suche nach der optimalen Kombination zwischen einem bestimmten Cluster-Algorithmus und einem bestimmten Distanz- oder Ähnlichkeitskoeffizienten sein sollte, bzw. daß man über solche Studien auch das optimale Distanz-/Ähnlichkeitsmaß ermitteln kann. Dem glauben wir entgegenhalten zu müssen, daß sich die Wahl eines geeigneten Maßes in erster Linie aus der Fragestellung ergibt, in deren Rahmen die Clusteranalyse eingesetzt wird, bzw. aus theoretischen Vorüberlegungen.

Um diese Auffassung zu verdeutlichen, wollen wir zunächst eine Erklärung dafür liefern, daß in den Studien von EDELBROCK einige Clusterverfahren deutlich bessere Resultate erbrachten, wenn als Ähnlich-

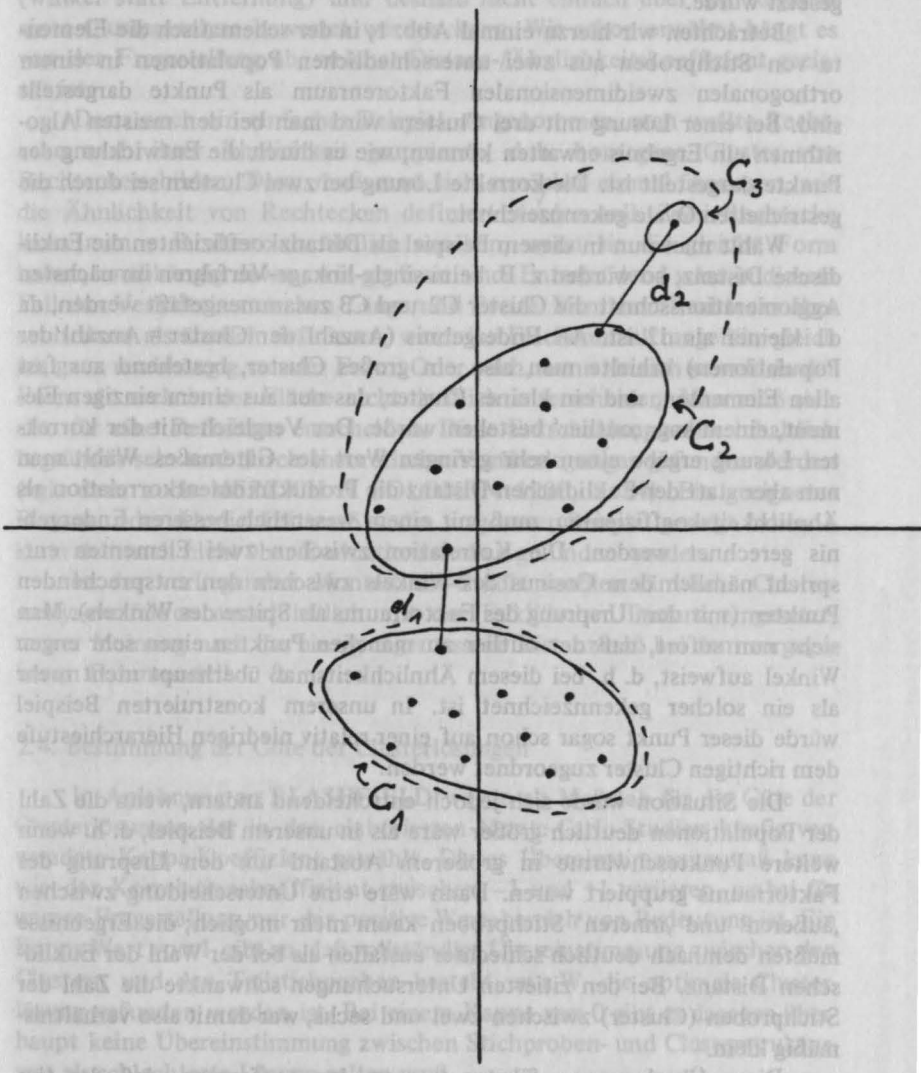

Abb. 1: Veranschaulichung des Effekts von „outliern" auf das Resultat von Cluster-Algorithmen (z. B. von linkage-Methoden)

keitsmaß statt der Euklidischen Distanz der Korrelationskoeffizient eingesetzt wurde.

Betrachten wir hierzu einmal Abb. 1, in der schematisch die Elemente von Stichproben aus zwei unterschiedlichen Populationen in einem orthogonalen zweidimensionalen Faktorenraum als Punkte dargestellt sind. Bei einer Lösung mit drei Clustern wird man bei den meisten Algorithmen ein Ergebnis erwarten können, wie es durch die Entwicklung der Punkte dargestellt ist. Die korrekte Lösung bei zwei Clustern sei durch die gestrichelten Ovale gekennzeichnet.

Wählt man nun in diesem Beispiel als Distanzkoeffizienten die Euklidische Distanz, so würden z. B. beim single-linkage-Verfahren im nächsten Agglomerationsschritt die Cluster C2 und C3 zusammengefaßt werden, da d1 kleiner als d2 ist. Als Endergebnis (Anzahl der Cluster = Anzahl der Populationen) erhielte man also ein großes Cluster, bestehend aus fast allen Elementen, und ein kleines Cluster, das nur aus einem einzigen Element, einem sog. ‚outlier' bestehen würde. Der Vergleich mit der korrekten Lösung ergäbe einen sehr geringen Wert des Gütemaßes. Wählt man nun aber statt der Euklidischen Distanz die Produktmomentkorrelation als Ähnlichkeitskoeffizienten, muß mit einem wesentlich besseren Endergebnis gerechnet werden. Die Korrelation zwischen zwei Elementen entspricht nämlich dem Cosinus des Winkels zwischen den entsprechenden Punkten (mit dem Ursprung des Faktorraums als Spitze des Winkels). Man sieht nun sofort, daß der outlier zu manchen Punkten einen sehr engen Winkel aufweist, d. h. bei diesem Ähnlichkeitsmaß überhaupt nicht mehr als ein solcher gekennzeichnet ist. In unserem konstruierten Beispiel würde dieser Punkt sogar schon auf einer relativ niedrigen Hierarchiestufe dem richtigen Cluster zugeordnet werden.

Die Situation würde sich jedoch entscheidend ändern, wenn die Zahl der Populationen deutlich größer wäre als in unserem Beispiel, d. h. wenn weitere Punkteschwärme in größerem Abstand um den Ursprung des Faktorraums gruppiert wären. Dann wäre eine Unterscheidung zwischen ‚äußeren' und ‚inneren' Stichproben kaum mehr möglich; die Ergebnisse müßten demnach deutlich schlechter ausfallen als bei der Wahl der Euklidschen Distanz. Bei den zitierten Untersuchungen schwankte die Zahl der Stichproben (Cluster) zwischen zwei und sechs, war damit also verhältnismäßig klein.

Diese Überlegungen führten uns zu der Auffassung, daß das (im Schnitt) bessere Abschneiden einzelner Clusteralgorithmen bei der Anwendung der Korrelation als Ähnlichkeitsmaß vor allem darauf zurückzuführen ist, daß outlier das Resultat nicht mehr in dem Maße beeinträchtigen können, wie das bei der Euklidischen Distanz der Fall ist.

Das Beispiel macht aber deutlich, daß der Unterschied bzw. die

Ähnlichkeit von Elementen bei der Korrelation ganz anders definiert ist (Winkel statt Entfernung) und deshalb nicht einfach über das Resultat einer Clusteranalyse bewertet werden kann. Wie schon erwähnt hängt es von der Fragestellung ab, welcher Distanz-/Ähnlichkeitskoeffizient geeignet ist.

Dazu noch ein einfaches Beispiel. Angenommen, man wollte Rechtecke nach ihrer Ähnlichkeit gruppieren, d. h. homogene Cluster von Rechtecken bilden. Dazu muß man sich zunächst einmal überlegen, wie die Ähnlichkeit von Rechtecken definiert werden soll. Zwei Rechtecke kann man z. B. dann als ähnlich ansehen, wenn sie die gleiche Form haben, unabhängig davon, wie groß sie sind. Entscheidend wäre in diesem Fall das Verhältnis zwischen Länge und Breite. Man kann Rechtecke aber auch dann als ähnlich definieren, wenn ihr Flächeninhalt ungefähr gleich ist, ganz unabhängig von der Form. Oder auch, wenn sie sich sowohl in der Form als auch in der Fläche nicht deutlich unterscheiden. Man muß sich also für eine Definition entscheiden. Diese Entscheidung sollte inhaltlich begründet sein und durch theoretische Vorüberlegungen gefunden werden (vgl. dazu auch MEZZICH & SOLOMON 1980, S. 14). Ein geeignetes Distanz- oder Ähnlichkeitsmaß – das mag unser Beispiel gezeigt haben – kann demnach nicht über Evaluationsstudien gefunden werden.

In der vorliegenden Monte-Carlo-Studie wurden sämtliche Clusteranalyseverfahren ausschließlich mit der Euklidschen Distanz eingesetzt, unsrer Meinung nach das einzig angemessene Distanzmaß bei dem vorgegebenen Datenmodell.

2.4. Bestimmung der Güte der Clusterlösungen

In Anlehnung an BLASHFIELD wurde als Maßstab für die Güte der Clusterlösungen der in den elaborierten Monte-Carlo-Studien häufig verwendete Kappa-Koeffizient gewählt. Dieses Übereinstimmungsmaß kann wie der Korrelationskoeffizient zwischen −1 und +1 variieren, wobei für unsere Fragestellung nur der positive Wertebereich von Bedeutung ist. Ein Kappa-Wert von 1 gibt an, daß vollständige Übereinstimmung zwischen den Clustern und den Teilstichproben besteht, m.a.W. die optimale Clusterlösung gefunden worden ist. Bei einem Kappa von 0 gibt es dagegen überhaupt keine Übereinstimmung zwischen Stichproben- und Clusterstruktur, was als schlechteste Lösung gelten muß.

Wie schon oben angedeutet, liegt ein kritischer Punkt bei der Evaluation von Clusteranalyse-Verfahren in der Existenz von ‚outliern', also Einzelelementen, die bei bestimmten Verfahren Mini-Cluster bilden. Diese Verfahren schneiden demnach besonders schlecht ab, wenn sie genausoviele Cluster bilden sollen, wie Teilstichproben vorhanden sind. Es scheint

daher angebracht, solche Mini-Cluster bei der Analyse gesondert zu berücksichtigen.

Um das Verhalten der einzelnen Verfahren gegenüber solchen Mini-Clustern genauer zu untersuchen, wurde eine Vorgehensweise geewählt, die schrittweise immer mehr Mini-Cluster in die Berechnung der Kappa–Koeffizienten mit einbezieht. Dabei ließen wir von jedem Clusteranalyseverfahren zunächst K+9 Cluster bestimmen (K = Anzahl der Subpopulatioen bzw. Teilstichproben) und analysierten nur jene K Cluster, die die größere Übereinstimmung mit dem vorgegebenen Teilstichproben aufwiesen. Auf diese Weise blieben zunächst outlier und Mini-Cluster unberücksichtigt. Daraufhin wurde die Zahl der Restcluster schrittweise um eines reduziert, bis die Anzahl der Cluster der Anzahl der Teilstichproben exakt entsprach. Auf jeder der zehn Stufen wurde der Kappa-Koeffizient als Übereinstimmungsmaß zwischen den K ausgesuchten Clustern und den vorgegebennen Teilstichproben berechnet.

2.5. Zusammenfassende Darstellung der Methode

Vereinfacht dargestellt, setzte sich die Analyse für jeden Cluster-Algorithmus aus folgenden Schritten zusammen: zunächst wurde ein Datensatz über einen Zufallsgenerator erzeugt. Dieser Datensatz war als eine Ziehung von k Stichproben aus einer Mischpopulation von k multinormalverteilten Subpopulationen definiert. Der generierte Datensatz wurde von den Algorithmen in einzelne Cluster zerlegt. Dabei lieferte jedes CA-Verfahren 10 Lösungen mit maximal k+9 und minimal k Clustern. Anschließend wurden alle Clusterlösungen mit den Teilstichproben verglichen und Kappa-Werte zur Bestimmung der Reproduktionsgüte berechnet. Diese Prozedur wurde pro Algorithmus 200-mal wiederholt, d. h. es wurden 200 Datensätze generiert und clusteranalysiert. Die Ergebnisse wurden gespeichert und nach Abschluß der 200 Durchgänge ausgewertet.

Diese Prozedur unterscheidet sich nur geringfügig von der bei SCHEIBLER & SCHNEIDER (1978) benutzten Vorgehensweise. Es entfällt lediglich die Diskriminanzanalyse, da sie sich aufgrund der Vorselektion der Datensätze erübrigt (die vollständige Trennbarkeit der Teilstichproben ist gewährleistet).

3. Ergebnisse

Das Resultat der Analyse ist in Tab. 3 zusammengefaßt. Es werden die Mediane von jeweils 200 Kappa-Werten wiedergegeben (je ein Median pro Verfahren und Analyse-Stufe).

Tab. 3:

Ergebnisse der Monte-Carlo-Studie mit restringiert multivariat normalverteilten Mixturen:
Mediane der Kappa-Werte auf verschiedenen Hierarchieebenen auf der Basis von 200 Datensätzen

	single link.	compl. link	aver. link	Median-Meth.	Centroid-Meth.	WARD's-Meth.	beta = –.5	LANCE-WILLIAMS'-Meth. beta = –.25	beta = 0	MCQUITTY's Meth.	RELOCATE*	KMEANS
K+9	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.98
K+8	.89	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
K+7	.86	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
K+6	.82	.98	1.	1.	1.	1.	1.	1.	1.	1.	1.	.92
K+5	.79	.94	.98	1.	1.	1.	1.	1.	1.	1.	1.	.91
K+4	.75	.89	.97	.87	.90	1.	1.	1.	.97	.97	.86	
K+3	.70	.82	.95	.82	.86	1.	.98	.98	.95	.96	.85	
K+2	.59	.69	.90	.75	.80	.96	.93	.94	.90	.91	.77	
K+1	.50	.57	.77	.60	.65	.90	.87	.86	.75	.80	.69	
K	.04	.38	.16	.04	.05	.791	.77	.72	.43	.58	.55	

K = Anzahl der Teilstichproben

RELOCATE*	.775
KMEANS*	.773

* Anfangspartitionierung durch Zufallsaufteilung der Stichprobe

** Anfangspartitionierung entspricht der WARD-Lösung bei K Clustern

Aus der Tabelle geht hervor, daß die Ergebnisse der Clusteranalyse umso schlechter ausfallen, je mehr Mini-Cluster aus der Analyse ausgeschlossen werden. Der Verlauf der Kappa-Koeffizienten macht aber auch deutlich, daß es nicht genügt, nur die ‚letzte' Analysestufe zu interpretieren. Ähnlich wie in den meisten der oben zitierten Untersuchungen schneidet auch in dieser Simulationsstudie das single-linkage-Verfahren am schlechtesten ab. Auf der letzten Analysestufe unterscheidet es sich zwar praktisch kaum von der Median- und Centroid-Methode, doch zeigen jene Verfahren auf den früheren Stufen durchweg bessere Ergebnisse (wobei ein gewisser Vorsprung der Centroid-Methode erkennbar ist).

Die complete-linkage-Methode liefert bei k Clustern zwar das beste Ergebnis der 5 erstgenannten Verfahren, doch schneidet sie bei steigender Cluster-Anzahl schlechter als die average-linkage-Methode sowie die Median- bzw. Centroid-Technik ab. Insgesamt gesehen wird damit die ‚outlier'-Anfälligkeit der 5 erstgenannten Verfahren bestätigt. Aus Tab. 3 geht aber auch hervor, daß diese Algorithmen nicht annähernd an das überlegene Resultat der WARD-Methode herankommen. Lediglich der Algorithmus von LANCE & WILLIAMS scheint der WARD-Technik ebenbürtig zu sein (insbesondere für beta = −.5). Es fällt auf, daß sich beide Verfahren als äußerst robust gegenüber ‚outliern' erweisen.

Die iterativ-partitionierenden Verfahren RELOCATE und KMEANS erbrachten nur mittelmäßige Ergebnisse, wenn sie mit einer Zufallspartitionierung starten mußten. Hier zeigt sich der Nachteil ihres heuristischen Algorithmus', der darin zu sehen ist, daß sie sich bei schlechten Startkonfigurationen im lokalen (nicht absoluten) Minimum verfangen. Diese Algorithmen sind demnach zu empfehlen, wenn die Anfangspartitionierung hinreichend gut ist, wenn also z. B. die Ergebnisse des WARD-Verfahrens dazu herangezogen werden. Überraschend ist allerdings, daß beide Verfahren unter dieser Bedingung nicht besser als die WARD-Technik abschneiden, ja im Schnitt sogar geringfügig schlechtere Ergebnisse liefern. Dieser Befund sollte aber mit Vorsicht interpretiert werden. Eine genauere Analyse der Einzelergebnisse zeigt nämlich, daß etwa die Hälfte der Kappa-Koeffizienten kleinere und die andere Hälfte gleiche oder größere Werte enthielt. In allen Fällen, in denen RELOCATE das Ergebnis von WARD (i.d.R. nur geringfügig) modifizierte, wurde die Summe der Abstandsquadrate innerhalb der Cluster vermindert. Es ist nun allerdings schwer zu sagen, ob ein anderes Minimierungskriterium bessere Ergebnisse liefern würde. In diesem Zusammenhang scheint noch erwähnenswert, daß RELOCATE und KMEANS bei sehr schwer separierbaren Stichproben tendenziell besser abschnitten.

Schließlich sei noch am Rande bemerkt, daß die in unseren Studien auf der ‚letzten' Analysestufe aufgefundenen Mediane der Kappa-Werte für

das single-linkage-, complete-linkage-, average-linkage-Verfahren sowie den WARD-Algorithmus fast exakt mit den Ergebnissen von BLASHFIELD (1976) übereinstimmen. Dies überrascht vor allem deshalb, weil eine Rekonstruktion der BLASHFIELD-Ergebnisse überhaupt nicht angestrebt wurde und auch Unterschiede zwischen den beiden Untersuchungsansätzen bestehen. Die Ähnlichkeit der Ergebnisse im Hinblick auf die absoluten Kappa-Werte muß daher als Zufall gewertet werden. Allerdings wird damit erneut die von BLASHFIELD gefundene Rangordnung der Verfahren, insbesondere die Überlegenheit der WARD-Technik bestätigt.

4. Zusammenfassung und Diskussion

Für die einleitend formulierten Hauptfragestellungen der vorliegenden Monte-Carlo-Studie ließen sich relativ eindeutige Antworten finden. Zunächst einmal konnte gezeigt werden, daß sich das im Rahmen von mit unrestringiert multivariat normalverteilten Datensätzen operierenden Monte-Carlo-Studien gewonnene Ergebnismuster auch auf den vorliegenden Fall von restringiert multivariat normalverteilten Mixturen übertragen ließ. Bei hundertprozentiger Objekterfassung erwiesen sich bei den hierarchisch-agglomerativen Verfahren die Methoden nach WARD bzw. LANCE-WILLIAMS erneut als am leistungsstärksten, wie die relativ hohen Mediane der Kappa-Werte demonstrieren können. Interessant ist in diesem Zusammenhang, daß die von LANCE-WILLIAMS empfohlenen Variante des ‚flexible-beta'-Verfahrens (beta = −.25) zu niedrigeren Reproduktionswerten als die von uns zusätzliche erprobte Variante (beta = −.5) führt. Weiterhin ließ sich bestätigen, daß die bei vollständiger Objektklassifikation äußerst ungenau rekonstruierenden hierarchisch-agglomerativen Verfahren (single-linkage-, average-linkage-, Median- und Centroid-Methode) dann bedeutend besser abschneiden, wenn lediglich der überwiegende Teil der Elemente klassifiziert werden muß. Wenn dies auch im Trend die Befunde von EDELBROCK (1979) bestätigt, sind die Verbesserungen dennoch nicht so deutlich ausgeprägt, daß damit die Leistungen der Verfahren nach WARD bzw. LANCE-WILLIAMS erreicht werden könnten.

Dies ändert allerdings nichts daran, daß auch bei einer solchen Betrachtungsweise die oben als überlegen eingestuften Algorithmen ihren Vorsprung (wenn auch nicht mehr so ausgeprägt) behalten. Es kann also aufgrund der vorliegenden Monte-Carlo-Studie kaum ein Zweifel daran bestehen, daß die WARD-Methode und das LANCE-WILLIAMS-Verfahren bei Verwendung Euklidscher Distanzen mit Abstand die robustesten hierarchischen Algorithmen darstellen und damit sicherlich zur Anwendung zu empfehlen sind. Es muß allerdings hier die Einschränkung gemacht wer-

den, daß auch bei den Monte-Carlo-Studien nicht von genereller Validität ausgegangen werden kann.

Schließlich wird im Hinblick auf den Vergleich von hierarchischen und nichthierarchischen Clusteranalyse-Verfahren deutlich, daß letztere sicherlich gleichwertig robuste Prozeduren darstellen, wenn geeignete Anfangspartitionierungen vorgenommen werden. Erfolgen die Anfangspartitionierungen dagegen nach dem Zufallsprinzip (wie etwa von SPÄTH (1975) empfohlen), so pendeln sich die Reproduktionskennwerte auf einem deutlich niedrigen Niveau ein. Damit werden die von BLASHFIELD (1977) gefundenen Ergebnisse voll bestätigt.

Scheinen diese Befunde die Validität der früheren Evaluationsstudien zu dokumentieren, so muß doch folgende Einschränkung gemacht werden: Wenn auch in der eigenen Untersuchung ein sehr breites Spektrum an Datensätzen untersucht wurde, deren Zahl hinreichend groß schien, um die Güte der Verfahren zuverlässig schätzen zu können, muß man dennoch mit der Generalisierung der Befunde vorsichtig sein. Streng genommen erstreckt sich ihr Gültigkeitsbereich nur auf solche Daten vom selben Typ, d. h. auf Stichproben aus Mischpopulationen, die sich aus multinormalverteilten Subpopulationen zusammensetzen. Es bleibt unklar, wie sich die CA-Verfahren z. B. bei extrem schiefen (asymmetrischen) Verteilungen verhalten. Das hier verwendete Datenmodell erzeugt sphärische (elliptoide) Teil-Stichproben. Es bleibt damit auch offen, wie die zur sphärischen Clusterbildung tendierenden Algorithmen (WARD-Methode, LANCE-WILLIAMS-Verfahren und RELOCATE bzw. KMEANS) bei nicht-sphärischen Stichproben abgeschnitten hätten. Man kann allerdings davon ausgehen, daß die von uns ermittelten ‚Spitzen-Verfahren' immer dann gute Ergebnisse liefern, wenn die Verteilungen (Dichten) der Subpopulationen eingipflig und nicht extrem schief sind.

Wie schon eingangs erwähnt, wurde die hier vorgelegte Beurteilung von CA-Verfahren nicht zuletzt auch unter dem Aspekt vorgenommen, Aussagen zur praktischen Handhabbarkeit bzw. zur Benutzerfreundlichkeit machen zu können. Die hier analysierten Verfahren wurden alle dem Software-Paket CLUSTAN 1C von WISHART (1975) bzw. im Fall von KMEANS aus der Sammlung von ‚stand-alone'-Programmen bei SPÄTH (1975, 1980) entnommen. Sind schon die erforderlichen Vorkenntnisse der Benutzer bei CLUSTAN 1C erheblich zu nennen, so können die bei SPÄTH (wie auch etwa die bei STEINHAUSEN & LANGER 1977) publizierten Programme nur mit einigermaßen guten FORTRAN-Kenntnissen sachgemäß eingesetzt werden (bei KMEANS ist darüberhinaus nicht auszuschließen, daß bei bestimmten Datensätzen Endlosschleifen produziert werden (s. MÖBUS 1982)).

Gerade für den mit EDV-Problemen weniger vertrauten Benutzer

stellte es sich nun bisher als geradezu fatal heraus, daß die großen und komfortablen Programmsysteme in dieser Hinsicht unzureichend ausgestattet waren. Während im SPSS bis dato kein Cluster-Algorithmus eingebaut ist, sind erst in den neuesten Editionen der beiden anderen relativ weit verbreiteten Systemen (SAS und BMDP) deutliche Fortschritte erkennbar geworden. Dies gilt insbesondere für die neueste SAS-Version, in der nicht nur das hierarchische Verfahren von WARD sowie die Centroid- und average-linkage-Methode, sondern auch eine KMEANS-Prozedur verfügbar ist, die sich speziell für den Fall großer Stichproben (bis zu 100 000 Beobachtungen) anbietet.

Die Installation effizienter Clusteranalyse-Algorithmen in benutzerfreundlichen Software-Paketen wird dem zukünftigen Anwender Frustrationen ersparen helfen, wie sie den Verfassern dieser Arbeit geläufig waren und sich in dem (im Universitätsrechenzentrum Heidelberg aufgefundenen) Sprüchlein vielleicht am besten zusammenfassen lassen:

„Die Hälfte seines Lebens
wartet der Programmierer vergebens."

Literatur

Bayne, C. K., J. J. Beauchamp, C. L. Begovich & V. E. Kane: Monte-Carlo-comparisons of selected clustering procedures. Pattern Recognition, 12, 1980, 51–62.

Blashfield, R. K.: Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods. Psychological Bulletin, 83, 1976, 377–388.

Blashfield, R. K.: A consumer report on cluster analysis software: (3) Iterative partitioning methods. Report No. 3 from N.S.F. grant 74–20007. Gainesville, Florida, March 1977.

Edelbrock, C.: Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 14, 1979, 367–384.

Edelbrock, C. & B. McLaughlin: Hierarchical cluster analysis using intraclass correlations: A mixture model study. Multivariate Behavioral Research, 15, 1980, 299–318.

Faber, E. & W. Nollau: Über ein Verfahren zur automatischen Klassifikation. Schriftenreihe des DRZ, Heft S. 6, Darmstadt 1969.

McIntyre, J. B. & R. K. Blashfield: A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 15, 1980, 225–238.

Mezzich, J. E.: Evaluating clustering methods for psychiatric diagnosis. Biological Psychiatry, 13, 1978, 265–281.

Mezzich, J. E. & H. Solomon: Taxonomy and behavioral science-Comparative performance of grouping methods. New York: Academic Press 1980.

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

Möbus, C.: On the nontermination of the k-means clustering algorithm for certain data sets. EDV in Medizin und Biologie, 13, 1982.

Scheibler, D. & W. Schneider: Probleme und Ergebnisse bei der Evaluation von Clusteranalyse-Verfahren. Bericht aus dem Psychologischen Institut der Universität Heidelberg, Nr. 11, Juni 1978.

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalysen. I. Ein Überblick über einschlägige Evaluationsverfahren. Psychologische Beiträge, 24, 1983, H. 1 (im Druck) (a).

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalysen. II. Appendix: Kurzbeschreibung der verbreitetsten Clusteranalyse-Algorithmen. Psychologische Beiträge, 24, 1983, H. 3 (im Druck) (b).

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

Späth, H.: Cluster analysis algorithms. Chichester, England: Ellis Horwood 1980.

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

Wishart, D.: CLUSTAN 1C user manual. London: Computer Center 1975.