Hans Theo Forst

Zur hierarchischen Klassifizierung von Beobachtungseinheiten nach komparativen Merkmalen

(On the hierarchical classification of observation units according to comparative characteristics)

Forst, H. T.: **Zur hierarchischen Klassifizierung von Beobachtungseinheiten nach komparativen Merkmalen.** (On the hierarchical classification of observation units according to comparative characteristics.)
In: Intern. Classificat. 5 (1978) No. 2, p. 81–85

In the group of hierarchical classification procedures those ones have been predominant so far which classify observation units according to metrical or classificatory characteristics. Classification procedures based on comparative characteristics differing only in their degree of intensity have so far been neglected. The present paper looks into the problems of similarity measurement in the use of comparative characteristics and investigates – by means of data simulation – the efficiency of a pertinent procedure proposed by *Lance* and *Williams*. (Author)

1. Aufgabenstellung

Das Interesse an statistischen Verfahren, die eine zweckmäßige Zusammenfassung der Vielfalt realer Erscheinungen gestatten, ist in den letzten Jahren zunehmend größer geworden. Ziel der Klassifizierungsmethode, nach dem englisch-amerikanischen Sprachgebrauch auch Cluster-Analyse genannt, ist eine Untergliederung der Gesamtheit aller Beobachtungseinheiten in eine überschaubare Anzahl von Klassen, die bezüglich mehrerer interessierender Merkmale in sich möglichst homogen, untereinander dagegen möglichst heterogen sein sollen.

Hierarchische Klassifizierungsverfahren beginnen auf der ersten Stufe mit einer Messung der paarweisen Ähnlichkeit der zu betrachtenden Einheiten und verschmelzen dann schrittweise diejenigen Einheiten und Gruppen von Einheiten, die untereinander am ähnlichsten sind. Das Ergebnis einer solchen schrittweise vorgenommenen Agglomeration von Einheiten und Gruppen wird in einem Dendrogramm dargestellt (*Anderberg* 1973, *Vogel* 1975, *Sodeur* 1975, *Bock* 1974).

In der Literatur wird eine Fülle solcher hierarchischen Prozeduren vorgeschlagen, die sich in der Messung der Ähnlichkeit zwischen Einheiten und/oder Gruppen unterscheiden. Die anzuwendenden Maße sind ganz entscheidend abhängig von der Art der Merkmale, die zur Klassifizierung herangezogen werden. In der Statistik werden im wesentlichen drei Merkmalstypen mit unterschiedlichem Meß- oder Skalenniveau unterschieden: klassifikatorische Merkmale, deren Ausprägungen sich nur der Art nach unterscheiden, komparative Merkmale, deren Ausprägungen nur eine Anordnung aufweisen und metrische Merkmale, zwischen deren Ausprägungen ein Abstand definiert werden kann. Insbesondere für den letzteren Merkmalstyp, der bisher in der multivariablen Statistik bevorzugt untersucht worden ist, existiert eine Reihe von Unähnlichkeits- bzw. Distanz- und Streuungsmaßen, die sich in vielen praktischen Anwendungen bewährt haben. Für klassifikatorische Merkmale haben vor allem die australischen Forscher *Lance* und *Williams* (1967A) spezielle Informationsmaße angegeben, die den Maßen für metrische Merkmale analoge Eigenschaften aufweisen. Auch diese Maße wurden in empirischen Untersuchungen mit Erfolg eingesetzt.

Die Gruppe der komparativen Merkmale, die vom Meß- und Informationsniveau eine mittlere Stellung zwischen klassifikatorischen und metrischen Merkmalen einnimmt, wurde dagegen bisher vernachlässigt. Dies läßt sich vermutlich darauf zurückführen, daß für diesen Merkmalstyp kein geschlossenes System von Streuungs- bzw. Heterogenitätsmaßen angegeben werden kann, wie es für die anderen Merkmalstypen möglich ist. Dennoch ist es zumindest erstaunlich, daß die Versuche zur Definition eines diesem Merkmalstyp adäquaten Klassifizierungsverfahrens nur sehr spärlich sind, da komparative Merkmale in vielen Anwendungsbereichen, etwa der Soziologie, Psychologie oder Wirtschaftswissenschaften, recht häufig auftreten.

In vielen empirischen Untersuchungen werden komparative Merkmale entweder wie klassifikatorische oder wie metrische Merkmale behandelt. In Grenzfällen, wenn etwa die Anzahl der Ausprägungen sehr gering oder sehr groß ist, mag diese Vorgehensweise gerechtfertigt sein, in den vorherrschenden Fällen wird jedoch bei der Behandlung als klassifikatorisches Merkmal ein Teil der Information, nämlich die in den Ausprägungen enthaltene Anordnungsinformation, nicht ausgenutzt, oder aber es wird bei der Behandlung als metrisches Merkmal unterstellt, daß die Unterschiede zwischen den einzelnen Ausprägungen des komparativen Merkmals annähernd gleich groß sind.

Die vorliegende Arbeit untersucht die Probleme bei der Ähnlichkeitsmessung für komparative Merkmale und stellt ein hierarchisches Klassifizierungsverfahren für komparative Merkmale vor, das ursprünglich von *Lance* und *Williams* (1967B) vorgeschlagen wurde. Die Leistungsfähigkeit dieses merkmalsadäquaten Verfahrens wird anschließend anhand von Datensimulationen überprüft.

2. Messung der paarweisen Ähnlichkeit von Einheiten bei komparativen Merkmalen

Zu Beginn der hierarchischen Klassifizierungsprozedur wird die Ähnlichkeitsmatrix berechnet, die die Maße für die paarweise Ähnlichkeit der Einheiten als Komponenten enthält:

	1	2	...	n
1	s_{11}	s_{12}	...	s_{1n}
2	s_{21}	s_{22}	...	s_{2n}
...	...	...	...	...
n	s_{n1}	s_{n2}	...	s_{nn}

Diese Matrix ist symmetrisch, d.h. es gilt:

(1) $s_{ij} = s_{ji}$ für $i < j = 1, 2, ..., n$,

wobei n = Anzahl der Beobachtungseinheiten,
und enthält in der Hauptdiagonale den Maximalwert des Ähnlichkeitsmaßes bzw. den Minimalwert des Unähnlichkeits- oder Heterogenitätsmaßes. Die Betrachtung kann daher auf die untere oder obere Dreiecksmatrix beschränkt werden.

Das zu berechnende Ähnlichkeits- bzw. Unähnlichkeitsmaß sollte bei komparativen Merkmalen speziell die Forderung erfüllen, daß es die in den Ausprägungen enthaltene Ordnungsinformation ausnutzt.

Betrachtet man eine komparatives Merkmal mit den Ausprägungen $A_1, A_2, ... A_k$, so bestehen zwischen diesen Ausprägungen folgende Relationen:

(2) $A_1 < A_2 < \ldots < A_k$

Aufgrund dieser zwischen den Ausprägungen geltenden Ordnungsbeziehungen lassen sich entsprechende Ordnungsrelationen für Ähnlichkeitsmaße fordern, die die Ähnlichkeit der Ausprägungen untereinander messen. Eine Einheit, die die Ausprägung A_1 trägt, ist z.B. der Einheit mit der Ausprägung A_2 ähnlicher als einer Einheit mit der dritten Ausprägung des komparativen Merkmals, d.h. für das Ähnlichkeitsmaß muß gefordert werden, daß

(3) $S(A_1, A_2) > S(A_1, A_3)$

Allgemein müssen folgende Ordnungsrelationen erfüllt sein, die anhand einer Ähnlichkeitsmatrix für Ausprägungen abgeleitet werden können ($s_{ij} = S(A_i, A_j)$):

	A_1	A_2	A_3	...	A_k
A_1		s_{12}	s_{13}	...	s_{1k}
A_2			s_{23}	...	s_{2k}
....	.....	.....	.....	.....	.....
A_{k-1}					$s_{k-1,k}$
A_k					

Relationen zwischen Zeilenelementen:

(4) $s_{12} > s_{13} > \ldots > s_{1k}$
$s_{23} > \ldots > s_{2k}$
.....

Relationen zwischen Spaltenelementen:

(5) $s_{23} > s_{13}$
.....
$s_{k-1,k} > \ldots > s_{2k} > s_{1k}$

Für Unähnlichkeitsmaße gelten entsprechende entgegengesetzte Ordnungsrelationen.

Die für metrische Merkmale entwickelten Ähnlichkeits- bzw. Abstandsmaße erfüllen sämtliche der genannten Ordnungsrelationen, wenn man sie auf die (numerisch codierten) Ausprägungen der komparativen Merkmale anwendet. Als Beispiel sei der einfache euklidische Abstand genannt, der die Unähnlichkeit der Ausprägungen durch die absolut gesetzte Differenz zwischen den Indices i und j mißt:

(6) $d_{ij} = |i - j|$

Solche Maße können jedoch deswegen nicht auf komparative Merkmale übertragen werden, weil ihre Anwendung die Annahme gleich großer Unterschiede bzw. Abstände zwischen den Ausprägungen der komparativen Merkmale impliziert, d.h. es wird angenommen, daß die Maße in den Nebendiagonalen der Ähnlichkeits- bzw. Unähnlichkeitsmatrix übereinstimmende Werte annehmen. Für die oben dargestellte Ähnlichkeitsmatrix ergäbe sich folgende Annahme:

(7) $s_{12} = s_{23} = \ldots = s_{k-1,k}$
$s_{13} = s_{24} = \ldots = s_{k-2,k}$
.....

Eine solche Annahme entspricht jedoch nicht dem Informationsniveau komparativer Merkmale.

Als Ausweg aus diesem Dilemma bietet sich lediglich die Möglichkeit an, auf ein niedrigeres Informationsniveau hinabzusteigen und solche Ähnlichkeits- bzw. Unähnlichkeitsmaße zu verwenden, die für klassifikatorische Merkmale entwickelt worden sind. Als Heterogenitätsmaß bietet sich speziell das Entropiemaß nach *Shannon* (1948) an, das sehr günstige Eigenschaften besitzt. Für Paare von Einheiten ist es wie folgt definiert (*Lance/Williams* 1967):

(8) $H_{ij} = a_{ij} \cdot 2 \log 2$,

wobei a_{ij} = Anzahl der Merkmalsausprägungen, in denen die Einheiten i und j nicht übereinstimmen.

Das Maß nimmt den Wert 0 an, wenn die Einheiten bezüglich aller Merkmale übereinstimmende Ausprägungen aufweisen. Jede Nichtübereinstimmung führt dagegen zu einer Vergrößerung des Entropiemaßes um 2 log 2. Da die Merkmalsausprägungen nur auf Übereinstimmung geprüft werden, die „Entfernung" zwischen ihnen jedoch außer acht gelassen wird, geht die Anordnungsinformation auf dieser Stufe verloren.

3. Das Klassifizierungsverfahren nach *Lance* und *Williams*

Lance und *Williams* (1967B) haben versucht, ein Verfahren für komparative Merkmale zu entwickeln, das zumindest auf den weiteren Stufen der hierarchischen Klassifizierung die Anordnung der Merkmalsausprägungen berücksichtigt. Das Verfahren ist informationsanalytisch und basiert auf dem Shannonschen Entropiemaß (*Shannon* 1948). Methodisch stellt es eine Modifikation des informationsanalytischen Verfahrens für klassifikatorische Merkmale dar, das zunächst kurz beschrieben werden soll (*Dale, Lance, Albrecht* 1971).

Betrachtet man zwei Gruppen von Einheiten, so wird der Homogenitätsverlust, der bei der Verschmelzung dieser Gruppen entstehen würde (die „Unähnlichkeit" der beiden Gruppen) durch den Zuwachs an Entropie innerhalb der Gruppen gemessen. Die Entropie einer Gruppe C_h ist wie folgt definiert:

(8) $H(C_h) = \sum_{i=1}^{m} H_i(C_h)$

mit

(9) $H_i(C_h) = n_h \log n_h - \sum_{j=1}^{k_i} n_{ij} \log n_{ij}$, wobei

$H_i(C_h)$ = Entropie des Merkmals i in C_h
n_{ij} = Anzahl der Einheiten mit der j-ten Ausprägung des Merkmals i in C_h

k_i = Anzahl der Ausprägungen des Merkmals i
n_h = Anzahl der Einheiten in C_h
m = Anzahl der Merkmale

Der Entropiezuwachs ΔH, der durch die Fusion zweier Gruppen C_h und C_l entsteht, ist:

(10) $\Delta H(C_h, C_l) = H(C_h \cup C_l) - H(C_h) - H(C_l)$

Bei der Modifikation des Verfahrens für komparative Merkmale werden die Ausprägungen des jeweiligen Merkmals vor Berechnung der Entropiemaße dichotomisiert, und zwar so, daß die Ordnungsrelation in der Dichotomie erhalten bleibt. Betrachtet man ein komparatives Merkmal mit den Ausprägungen 1, 2, ..., k, so sind folgende Dichotomisierungen unter Beibehaltung der Ordnungsrelation möglich:

1 ; 2, 3, . . . , k
1,2; 3, 4, . . . , k
. . . . . .
1, 2, . . . , k-1; k

Insgesamt sind k-1 Dichotomien konstruierbar.

Für jede dieser Dichotomien wird der durch die Fusion zweier Gruppen C_h und C_l entstehende Entropiezuwachs für das jeweilige Merkmal i berechnet:

(11) $\Delta H_{id} = H_{id}(C_h \cup C_l) - H_{id}(C_h) - H_{id}(C_l)$
für d = 1, ..., $k_i - 1$

Als Beitrag des Merkmals i zum Gesamtentropiezuwachs wird der maximale Wert von ΔH_{id} ausgewählt:

(12) $\Delta H_i = \text{Max}\,(\Delta H_{id})$

Diese Berechnungen werden für alle m Merkmale durchgeführt. Der Gesamtentropiezuwachs ist dann gleich der Summe der abgeleiteten maximalen Entropiezuwächse:

(13) $\Delta H = \sum_{i=1}^{m} \Delta H_i$

Die einzelnen Schritte des Algorithmus möge folgendes einfache Beispiel verdeutlichen, wobei die entsprechenden Ergebnisse für das Verfahren mit klassifikatorischen Merkmalen gegenübergestellt werden:

Datenmatrix

Einheit	Merkmal 1	2	3
1	1	1	1
2	2	1	1
3	1	3	2
4	3	2	2
5	2	2	3

Matrix mit Informationsmaßen für Paare von Einheiten (für beide Verfahren identisch!):

	1	2	3	4	5
1	0	1,39	2,77	4,16	4,16
2	1,39	0	4,16	4,16	2,77
3	2,77	4,16	0	2,77	4,16
4	4,16	4,16	2,77	0	2,77
5	4,16	2,77	4,16	2,77	0

Min ΔH = 1,39 (für beide Verfahren)
Die Einheiten 1 und 2 werden fusioniert.

Korrigierte Matrix mit Entropiezuwächsen (für beide Verfahren identisch):

	1,2	3	4	5
1,2	0	4,34	5,72	4,34
3	4,34	0	2,77	4,16
4	5,72	2,77	0	2,77
5	4,34	4,16	2,77	0

Min ΔH = 2,77 (für beide Verfahren)
Die Einheiten 3 und 4 werden fusioniert (keine eindeutige Lösung!)

Korrigierte Matrix mit Entropiezuwächsen:

	1,2	3,4	5
1,2	0	6,41	4,34
3,4	6,41	0	2,96
5	4,34	2,96	0

Min ΔH = 2,96
Die Gruppen 3,4 und 5 werden fusioniert.

	1,2	3,4	5
1,2	0	6,93	4,34
3,4	6,93	0	4,34
5	4,34	4,34	0

Min ΔH = 4,34 (keine eindeutige Lösung!)
Die Gruppen 1,2 und 5 werden fusioniert.

Korrigierte Matrix mit Entropiezuwächsen:

	1,2	3,4,5
1,2	0	7,32
3,4,5	7,32	0

Min ΔH = 7,32

	1,2,5	3,4
1,2,5	0	7,32
3,4	7,32	0

Min ΔH = 7,32

Fusion der beiden Gruppen 1,2 und 3, 4, 5 bzw. der Gruppen 1, 2, 5 und 3, 4.

Unterschiede zwischen den beiden Verfahren lassen sich insbesondere auf der Stufe 3 der hierarchischen Prozedur erkennen. Die Entropiezuwächse unterscheiden sich dort erheblich. Der Heterogenitätszuwachs, der durch die Fusion der Gruppe 3, 4 mit der Einheit 5 hervorgerufen wird, wird vom Verfahren für komparative Merkmale mit 2,96, beim Verfahren für klassifikatorische Merkmale dagegen mit 4,34 angegeben. Die Herleitung des Entropiezuwachses nach dem Verfahren von *Lance/Williams* soll für diesen Fall ausführlicher nachvollzogen werden:

Randverteilungen der Merkmale für die zu untersuchenden Gruppen:

Merkmal 1:

Ausprägung	absolute Häufigkeit in Gruppe 3,4	Einheit 5	Gruppe 3,4,5
1	1	0	1
2	0	1	1
3	1	0	1

Merkmal 2:

Ausprägung	absolute Häufigkeit in Gruppe 3,4	Einheit 5	Gruppe 3,4,5
1	0	0	0
2	1	1	2
3	1	0	1

Merkmal 3:

Ausprägung	absolute Häufigkeit in		
	Gruppe 3,4	Einheit 5	Gruppe 3,4,5
1	0	0	0
2	2	0	2
3	0	1	1

Entropiezuwächse bei der Fusion der Gruppen 3, 4 mit der Einheit 5

Merkmal Nr.	Verfahren für komparative Merkmale	Verfahren für klassifikatorische Merkmale
1	0,52	1,91
2	0,52	0,52
3	1,91	1,91

Vergleicht man die Entropiezuwächse, die auf die einzelnen Merkmale entfallen, mit den entsprechenden Entropiezuwächsen bei Anwendung des Verfahrens für klassifikatorische Merkmale, so erkennt man, daß die Unterschiede nur durch Merkmal 1 verursacht werden. Bei Fusion der Gruppe 3,4 mit der Einheit 5 entsteht bezüglich Merkmal 1 eine Gleichverteilung (jede Ausprägung ist einmal besetzt) mit der maximalen Entropie von 3 log 3 (= 3,2958). Werden die Merkmalsausprägungen hingegen dichotomisiert, entsteht eine Ungleichverteilung mit entsprechend geringerer Entropie. Durch das Dichotomisierungskonzept wird berücksichtigt, daß benachbarte Ausprägungen untereinander sehr ähnlich sein und im Grenzfall sogar als eine Kategorie betrachtet werden können. Da nicht bekannt ist, wie ähnlich sich die Ausprägungen sind, müssen sämtliche Zusammenfassungen dieser Art überprüft werden. Zur Messung des Homogenitätsverlustes ist dann diejenige Konstellation relevant, die zu einer maximalen Steigerung der Entropie führt.

Das Dichotomisierungskonzept hat sich übrigens auch in einem weiteren Bereich der multivariablen Analyse, der „Kontrastgruppenanalyse", bereits bewährt. Dort werden diejenigen Dichotomien der komparativen Prädiktorvariablen, die zu einer maximalen Streuungsreduktion der Prädikandenvariablen führen, zur Bildung der Kontrastgruppen herangezogen (*Gillo/Shelly* 1973).

Lance und *Williams* betonen mit besonderem Nachdruck, daß dieses Konzept nicht impliziert, daß Unterschiede zwischen benachbarten Ausprägungen des komparativen Merkmals als weniger wichtig erachtet werden als Unterschiede zwischen weiter entfernten Ausprägungen (*Lance/Williams* 1967B, S. 18). Es wird also nichts über die Unterschiede zwischen den Ausprägungen vorausgesetzt. Diese Eigenschaft des Verfahrens ist im Hinblick auf das Informationsniveau komparativer Merkmale besonders wichtig.

Das modifizierte Maß für den Entropiezuwachs bei komparativen Merkmalen hat im Gegensatz zum entsprechenden Maß für klassifikatorische Merkmale keine Additivitätseigenschaft, d.h. die Summe der Entropiezuwächse auf den Stufen der Klassifikationshierarchie ist nicht gleich der entsprechend definierten Entropie der Datenmatrix. Eine Aufspaltung in die Entropie innerhalb bzw. zwischen den Gruppen ist anhand dieses Maßes daher nicht möglich.

Ein weiterer Nachteil ist, daß die Folge der Entropiezuwächse auf den Stufen der Hierarchie häufig nicht monoton wachsend ist. Es treten Reversals auf, die sich im Dendrogramm recht störend bemerkbar machen können. Bei den durchgeführten Datensimulationen (vgl. Abschn. 4) traten in der Regel etwa 2 bis 3 Reversals vornehmlich auf den unteren Stufen der Hierarchie auf. Auch *Lance/Williams* beschreiben dieses Phänomen, weisen aber darauf hin, daß diese Erscheinung im Gegensatz zu den Verfahren „Centroid Sorting" oder „Median Group Average Sorting" für metrische Merkmale nur recht harmlose Formen annimmt. Dies kann im wesentlichen bestätigt werden.

4. Leistungsfähigkeit des Verfahrens

Das vorgestellte hierarchische Klassifizierungsverfahren für komparative Merkmale wurde einem Leistungstest unterzogen, um zu überprüfen, inwieweit das von *Lance* und *Williams* vorgeschlagene Konzept die Ordnungsinformation der Merkmale ausnutzt. Dazu wurden verschiedene Datensimulationen auf der CDC-Rechenanlage der Universität zu Köln vorgenommen.

Um den Aufwand möglichst gering zu halten, wurden Datensätze mit relativ kleiner Anzahl von Einheiten und Merkmalen (n = 50, m = 4) untersucht. Zur Erzeugung der Merkmalswerte wurden mit einem CDC-spezifischen Zufallszahlengenerator im Intervall [0, 1] rechteckverteilte Zufallzahlen generiert, die anschließend mittels der Box-Muller-Transformation in standardnormalverteilte Zufallszahlen überführt wurden (vgl. *Box/Muller* 1958). Durch Addition von Konstanten für jedes Merkmal wurde die Hälfte der Einheiten mittelwertverschoben, so daß die Zufallsvektoren zwei bezüglich der Merkmalsmittelwerte unterschiedlichen Gesamtheiten entstammten. Komparative Merkmalswerte wurden aus diesen Daten durch Einteilung in Klassen und Zuweisung der Klassennummern erzeugt. Die Zufallszahlen wurden vor der Klassifizierung standardisiert. Insgesamt wurden 100 solcher Datensätze erstellt und anschließend mit dem FORTRAN IV-Programm YGROUP klassifiziert. Die Übereinstimmung der resultierenden Klassifizierung mit der a-priori bekannten Aufteilung wurde mit Hilfe des Concordanzmaßes nach W. *Rand* (1971) gemessen.

Für die erste Simulationsuntersuchung wurden mit Ausnahme der Randklassen die standardisierten Merkmalswerte in 6 gleichbreite Klassen aufgeteilt:

Tabelle 1: Klassengrenzen in Simulationsuntersuchung 1

Klassen Nr.	Klassengrenzen
1	$\leqslant -2$
2	> -2 bis -1
3	> -1 bis 0
4	> 0 bis $+1$
5	$> +1$ bis $+2$
6	$> +2$

Die Datensätze wurden mit folgenden Verfahren klassifiziert:

Tabelle 2: Angewendete Klassifizierungsverfahren

Nr.	Verfahren	Kurzbezeichnung
1	Verfahren nach *Ward*	WARD
2	Verfahren für komparative Merkmale nach *Lance/Williams*	KOMP
3	Verfahren für klassifikatorische Merkmale	KLAS

Tabelle 3 zeigt die Mittelwerte und Standardabweichungen für das Concordanzmaß:

Tabelle 3: Ergebnisse der Simulationsuntersuchung 1

Verfahren	Mittelw. d. Concordanzmaßes	Stand.abw. d. Concordanzmaßes	Laufzeit für 100 Datensätze in sec.
WARD	0,63	0,07	27
KOMP	0,59	0,08	176
KLAS	0,54	0,06	130

Das Simulationsergebnis zeigt deutlich die Überlegenheit des Verfahrens KOMP gegenüber dem Verfahren KLAS. Die Berücksichtigung der Ordnungsinformation führt im Mittel zu einer wesentlichen Verbesserung des Klassifizierungsergebnisses. Gemessen an dem Simulationsergebnis, das sich für die nicht klassierten Ursprungsdaten unter Verwendung des Verfahrens von Ward ergibt, Mittelwert 0,65 und Standardabweichung 0,08, nimmt das KOMP-Verfahren die vom geringeren Meßniveau der Merkmale her zu erwartende mittlere Position ein. Der zeitliche Mehraufwand (hier von ca. 35 %) gegenüber dem Verfahren KLAS lohnt sich offensichtlich. Das überraschend gute Abschneiden des Verfahrens von Ward – hier auf die komparativen Merkmalswerte angewendet – läßt sich vermutlich damit erklären, daß bei der Klassierung der Merkmalswerte im relevanten Bereich $[-2, +2]$ gleiche Klassenbreiten verwendet wurden. In diesem Falle ist die Annahme, daß die Unterschiede zwischen den Klassennummern gleich groß sind, gerechtfertigt. Insofern ist es plausibel, daß das Verfahren von Ward, das die Abstände zwischen den Klassennummern berücksichtigt, erfolgreicher sein muß.

Ein wesentlich anderes Bild ergibt sich, wenn man bei der Klassierung der Merkmalswerte extrem ungleiche Klassenbreiten verwendet. In einer zweiten Simulationsuntersuchung wurden folgende Klassengrenzen vorgegeben:

Tabelle 4: Klassengrenzen in Simulationsuntersuchung 2

Klassennummer	Klassengrenzen
1	$\leqslant -2$
2	> -2 bis $-1,8$
3	$> -1,8$ bis $-1,4$
4	$> -1,4$ bis $-0,5$
5	$> -0,5$ bis $+2$
6	$> +2$

In Tabelle 5 sind die Simulationsergebnisse für diesen Fall zusammengestellt.

Tabelle 5: Ergebnisse der Simulationsuntersuchung 2

Verfahren	Mittelwert des Concordanzmaßes	Stand.abw. des Concordanzmaßes
WARD	0,52	0,04
KOMP	0,58	0,06
KLAS	0,55	0,06

Man erkennt, daß das Verfahren von Ward sehr deutlich scheitert, weil die Voraussetzung gleich großer Unterschiede bzw. Abstände zwischen den Klassennummern nicht erfüllt ist. Demgegenüber reagieren die beiden Verfahren KOMP und KLAS erwartungsgemäß unempfindlich gegenüber der Änderung in den Klassengrenzen.

Zusammenfassend ist festzustellen, daß das Verfahren für komparative Merkmale dem Verfahren für klassifikatorische Merkmale überlegen ist, weil es zusätzlich einen großen Teil der Ordnungsinformation in den Merkmalen ausnutzt. Weiterhin ist vor der Anwendung des Verfahrens von Ward auf komparative Merkmale zu warnen, weil es sehr sensibel auf die Verletzung der Annahme gleich großer Unterschiede zwischen den Ausprägungen der Merkmale reagiert. Es empfiehlt sich daher, in den Fällen, in denen keine weiteren Informationen über zugrundeliegende Klassenbreiten und -unterschiede vorliegen – das dürfte der Regelfall sein –, dem skalenadäquaten Verfahren KOMP den Vorzug zu geben.

Quellen:

(1) Anderberg, M. R.: Cluster analysis for applications. New York–London: Academic Press 1973.

(2) Bock, H. H.: Automatische Klassifikation. Göttingen–Zürich: Vandenhoeck & Ruprecht 1974.

(3) Box, G. E. P., Muller, M. E.: A note on the generation of random normal deviates. In: Ann. Math. Statistics 29 (1958) p. 610–611.

(4) Dale, M. B., Lance, G. N., Albrecht, L.: Extensions of information analysis. In: Australian Comp. J. 3 (1971) No. 1.

(5) Gillo, M. W., Shelly, M. W.: A technique for predictive modelling of multivariable and multivariate data. Techn. Report 45–73, 1973.

(6) Lance, G. N., Williams, W. T.: A general theory of classificatory sorting strategies I: hierarchical systems. In: Comp. J. (1967) p. 373–380 (A).

(7) Lance, G. N., Williams, W. T.: Mixed-data classificatory programs, I. Agglomerative systems. In: Australian Comp. J. 1 (1967) p. 15–25 (B).

(8) Rand, W. M.: Objective criteria for the evaluation of clustering methods. In: J. Amer. Statistical Assoc. 66 (1971) p. 846–850.

(9) Shannon, C. E.: A mathematical theory of communication. In: Bell System Techn. J. (1948) p. 370–423, 623–656.

(10) Sodeur, W.: Empirische Verfahren zur Klassifikation. Stuttgart: Teubner Studienskripten 1974.

(11) Vogel, F.: Probleme und Verfahren der numerischen Klassifikation. Göttingen: Vandenhoeck & Ruprecht 1975.