Clusteranalyse mit gemischtskalierten Merkmalen[1]

Günter Buttler und Norman Fickel

Friedrich-Alexander-Universität Erlangen-Nürnberg
Wirtschafts- und Sozialwissenschaftliche Fakultät
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Lange Gasse 20
D-90 403 Nürnberg

Abstract

Ziel einer Clusteranalyse ist es, eine Menge von Objekten anhand gegebener Merkmale in möglichst homogene Gruppen aufzuteilen. Probleme treten immer dann auf, wenn Merkmale unterschiedlichen Skalenniveaus vorliegen. Vorgeschlagen wird, durch eine geeignete Normierung vom Skalenniveau zu abstrahieren. Dies geschieht, indem die merkmalsspezifischen paarweisen Abstände durch die jeweiligen durchschnittlichen paarweisen Abstände dividiert werden. Ergebnis sind nicht nur dimensionslose, sondern auch in der Streuung vereinheitlichte Werte, die zu einem Gesamtabstand von je zwei Objekten addiert werden können. Als Maß für die Güte der Klassifikation kann der Anteil der paarweisen Distanzen zwischen den Klassen verwendet werden, und zwar sowohl insgesamt als auch gesondert für die einzelnen Merkmale.

[1] Diese Arbeit wurde mit Mitteln der Hans-Frisch-Stiftung, Lange Gasse 20, D-90 403 Nürnberg, gefördert.

1. Zielsetzung

Die Clusteranalyse[2] ist ein multivariates Analyseverfahren, das eine Menge von Objekten, die durch mehrere Variablen (Merkmale) beschrieben werden, in Teilmengen (sogenannte Cluster oder Klassen) zerlegt. Dabei sollen die zu demselben Cluster gehörigen Objekte möglichst ähnlich, die zu verschiedenen Clustern gehörigen dagegen möglichst unähnlich sein. Technisch geht die Clusteranalyse von dem numerisch beschreibbaren Ergebnis einer Erhebung aus, das sich tabellarisch in einer Datenmatrix darstellen läßt (Abb. 1). Zu jedem Objekt der Erhebung gehört eine Zeile der Tabelle und zu jedem Merkmal eine Spalte. Jede Zelle der Tabelle enthält den zugehörigen numerischen Wert, also eine bestimmte reelle Zahl. Zur Interpretation der Daten ist freilich noch die Kenntnis darüber notwendig, was die Zahlen inhaltlich aussagen. Angenommen, die Objekte der Erhebung seien Mitglieder einer bestimmten Personengruppe. Dann muß bekannt sein, ob etwa beim Merkmal Geschlecht ein Mann mit der Zahl Null und eine Frau mit Eins oder umgekehrt codiert worden ist, oder in welcher Maßeinheit, etwa Meter, Zentimeter oder Zoll, das Merkmal Körpergröße gemessen wurde.

Abb. 1: Datenmatrix für n Objekte und m Merkmale

	Merkmal 1	Merkmal 2	...	Merkmal m
Objekt 1				
Objekt 2				
⋮				
Objekt n				

Bleiben Zellen unbesetzt, gab es also beispielsweise bei einem Fragebogen Nichtbeantwortungen, so wird die Situation durch diese fehlenden Werte erschwert. Möchte man ein Objekt oder ein Merkmal, in dessen Zeile beziehungsweise Spalte eine Zelle nicht besetzt ist, dennoch mit in die Analyse aufnehmen, so muß entweder ein geeigneter Ersatzwert gefunden oder das verwendete Verfahren geeignet modifiziert werden. Im folgenden sei jedoch zur Vereinfachung stets angenommen, daß alle Werte vorhanden seien, gegebenenfalls, da die von fehlenden Werten betroffenen Merkmale oder Objekte bereits eliminiert wurden. Außerdem seien auch Merkmale, die keinerlei Variabilität besitzen, also deren Ausprägung bei jedem Objekt dieselbe ist, ebenfalls bereits ausgeschlossen.

Die einzelnen Objekte werden aufgrund der zugehörigen Merkmalsausprägungen in Klassen eingeteilt. Im Idealfall gelingt das so, daß sich die in einer gemeinsamen Klasse befindlichen Objekte bezüglich aller ihrer Merkmale nur sehr wenig unterscheiden, die Merkmalsauspräg-

---

[2] auch: automatische oder numerische Klassifikation, numerische Taxonomie

gungen also nur wenig streuen. Die zugehörigen Klassen sind dann in sich homogen. Gleichzeitig würden sich idealerweise je zwei Objekte aus unterschiedlichen Klassen deutlich unterscheiden, die Streuung zwischen Objekten verschiedener Klassen also vergleichsweise groß sein. Die Heterogenität zwischen den Klassen ist dann hoch.

Um die gerade verwendeten Begriffe der Streuung innerhalb einer Klasse und zwischen den Klassen zu operationalisieren, läßt sich das Konzept des Abstandes zweier Objekte einsetzen. Dazu wird jedem Paar von Objekten Nr. i und Nr. j eine Zahl d_{ij} zugeordnet, die um so größer sein soll, je mehr sich im sachlogischen Sinne das i-te und j-te Objekt unterscheiden. Man erhält dann eine Distanzmatrix, die soviele Spalten und auch Zeilen hat, wie es Objekte gibt (Abb. 2).

Abb. 2: Matrix der paarweisen Distanzen zwischen n Objekten

	Objekt 1	Objekt 2	...	Objekt n
Objekt 1	d_{11}	d_{12}	...	d_{1n}
Objekt 2	d_{21}	d_{22}	...	d_{2n}
⋮	⋮	⋮	⋱	⋮
Objekt n	d_{n1}	d_{n2}	...	d_{nn}

Dabei können die Einträge d_{ij} der Distanzmatrix im einfacheren Fall nur von den Merkmalsausprägungen der beiden Objekte Nr. i und Nr. j abhängen, es ist aber auch möglich, daß die gesamte Datenmatrix in die Berechnung des Abstandes d_{ij} eingeht. Je nachdem läßt sich dann von einer ex-ante- beziehungsweise ex-post-Distanzmatrix sprechen. Im ex-ante-Fall kann also der Abstand zweier Objekte ohne Kenntnis der übrigen Objekte bestimmt werden, wohingegen im ex-post-Fall im allgemeinen alle Objekte bekannt sein müssen, um nur einen einzigen Abstand bestimmen zu können. Das klingt ziemlich kompliziert, ergibt sich aber bei bestimmten Datenstrukturen ganz naheliegend, wie weiter unten noch gezeigt wird.

Das Vorgehen bei der Clusteranalyse läßt sich unter Verwendung von Abstandsmaßen mithin in zwei Abschnitte gliedern:

- Festlegung der Abstände zwischen je zwei Objekten basierend auf den numerischen Werten der Merkmalsausprägungen, wobei Merkmale abhängig vom Informationsgehalt der Werte (Skalenniveau) unterschiedlich behandelt werden.

- Bildung von Klassen nach Auswahl eines formalen Verfahrens der Clusteranalyse, welches automatisch aufgrund der durch die Abstände gegebenen Distanzmatrix alle Objekte in Klassen einteilt.

2. Festlegung der Abstände

2.1. Skalenniveaus

Die tabellarische Aufstellung in Abb. 3 enthält einige Beispiele für Merkmale der angegebenen Skalenniveaus.

Abb. 3: Skalenniveau von Merkmalen

	diskret	stetig
nominal[3]	Studienfach	keine sinnvollen Beispiele
ordinal[4]	Klausurnote	striktes Ranking
metrisch[5]	Semesterzahl	Körpergröße

Bei einem nominal-skalierten Merkmal besteht die Information beim Vergleich zweier Zahlen nur darin, ob diese unterschiedlich oder identisch sind. Bei ordinaler Skalierung kann aus einem Größenvergleich zweier Zahlen auch auf eine entsprechende realwissenschaftliche Beziehung geschlossen werden. Ist das Skalenniveau metrisch, so ist auch die Differenz der Zahlen interpretierbar.

Charakteristisch für ein stetiges Merkmal ist, daß zwei Ausprägungen nie exakt gleich sein können, etwaige Gleichheiten in den Daten also entweder aus Meßungenauigkeiten oder als Rundungsfehler entstanden sind. Für stetig-ordinales Niveau bedeutet das, daß aus methodischer Sicht Bindungen nicht vorkommen können. Ein stetig-nominales Merkmal ist zwar theoretisch definiert, aber zwingenderweise sind alle zulässigen Maßzahlen ausschließlich von der Anzahl n der Objekte abhängig, da jede andere formale Information, die in den numerischen Werten der Ausprägungen vorhanden ist, nach Definition des Skalenniveaus nicht interpretierbar ist.

Demgegenüber geht man bei einem diskreten Merkmal von einer vorgegebenen Skala endlich vieler (oder höchstens abzählbar unendlich vieler) möglicher Ausprägungen aus. Man kennt also von vorneherein bestimmte vorstellbare Werte des Merkmals.

---

3 auch: qualitativ oder klassifikatorisch

4 auch: komparativ

5 auch: quantitativ oder kardinal, wobei als hier nicht näher behandelte Spezialfälle die Intervall-, Verhältnis- und Absolutskalierung unterschieden werden können

Zu beachten ist, daß für alle Skalenniveaus die Daten sowohl aus Einzelwerten bestehend als auch klassiert vorliegen können. Durch Klassierung geht Information verloren, wodurch eine daraus berechnete Maßzahl im allgemeinen nur noch die Näherung der entsprechenden aus den Einzelwerten gewonnenen Maßzahl ist. Das äußere Erscheinungsbild von klassiertem, stetigen Zahlenmaterial ist mit dem von diskretem identisch, obgleich inhaltlich ein ganz unterschiedliches Konzept dahinter steht.

In den folgenden Abschnitten wird nicht zwischen diskret- und stetig-metrischen Merkmalen getrennt, da sich für die benutzten Abstands- und Streuungsmaße kein Unterschied in Definition und Interpretation ergibt. Außerdem kann, da stetig-nominale Merkmale, wie bereits erwähnt, keine praktische Bedeutung haben, der Begriff "nominal" als gleichbedeutend für "diskret-nominal" verwendet werden. Damit bleiben bei Skalenniveaus nur noch vier Fälle, nämlich der nominale, diskret- und stetig-ordinale sowie der metrische, zu unterscheiden.

2.2. Abstandsmaße

Jedes Skalenniveau erfordert ein eigenes Abstandsmaß, das die Entfernung zweier Ausprägungen quantifiziert. Im folgenden wird bewußt durchgängig ein möglichst einfaches Maß ausgewählt, obgleich zum Teil auch andere, meist kompliziertere Möglichkeiten gebräuchlich sind. Das Maß, welches hier gewählt wird, ist aufgrund seiner Anschaulichkeit universell einsetzbar, wohingegen viele andere oft nur in speziellen Situationen sinnvoll sind.

Die Ausprägung des jeweils betrachteten Merkmals X beim i -ten Objekt werde mit x_i bezeichnet, wobei i die Nummern aller Objekte von 1 bis n durchlaufen kann. Auf einen eigenen Index für das Merkmal wird vorerst verzichtet, um die Notation nicht zu überladen.

2.2.1. Nominale Merkmale

Bei einem nominalen Merkmal ist nur die Information über Gleichheit und Ungleichheit der Ausprägungen sachlogisch bedeutsam. Daher steht, will man das Merkmal direkt verwenden, im wesentlichen nur die diskrete Metrik zur Wahl. Der Abstand ist null, falls die Ausprägungen gleich sind, und gleich der Konstanten Eins, falls sie verschieden sind, also:

$$d_{ij} = \begin{cases} 1, & x_i \neq x_j \\ 0, & x_i = x_j \end{cases} \qquad (i, j = 1,\ldots,n).$$

Dies läßt sich für ein nominales Merkmal mit nur zwei möglichen Ausprägungen 0 und 1 schreiben als

$$d_{ij} = \left| x_i - x_j \right| \qquad (i, j = 1,\ldots,n).$$

Hält man bei einem nominalen Merkmal mit mehr als zwei Ausprägungen die diskrete Metrik für zu wenig differenzierend, so ist es möglich, dieses in zwei oder mehr binäre[6], also andere nominale Merkmale mit nur zwei Ausprägungen, aufzuspalten. Allgemein kommt man stets mit weniger als $\log_2 k + 1$ binären Merkmalen aus. Jedoch kann es sinnvoll sein, trotzdem eine höhere Anzahl zu benutzen, wenn die neuen Merkmale inhaltlich besser zu interpretieren sind.

2.2.2. Diskret-ordinale Merkmale

Da bei einem ordinalen Merkmal generell nur Größenvergleiche interpretierbar sind, kann ein diskret-ordinales Merkmal mit k verschiedenen möglichen Ausprägungen ohne Informationsverlust auf die natürlichen Zahlen 1, ..., k transformiert werden. Mit r_i sei die so transformierte Ausprägung des Objekts Nr. i bezeichnet und der Abstand sei der Differenzbetrag

$$d_{ij} = \left| r_i - r_j \right| \qquad (i, j = 1,\ldots,n).$$

Der maximale Wert von d_{ij} ist offenbar *k*-1.

Die Benützung von Abständen zwischen den Ausprägungen ordinaler Merkmale mag auf Kritik stoßen. Sie scheint zu implizieren, daß gleiche Abstände zwischen den Ausprägungen unterstellt werden, was eigentlich erst bei metrischen Merkmalen möglich ist. Das Vorgehen läßt sich jedoch auch als Abzählung interpretieren, bei der lediglich die Anzahl aller dazwischen liegenden Ausprägungen bestimmt wird. Wenn zum Beispiel bei den Schulnoten 1, 2, ..., 6 zwei Schüler die Noten 1 und 4 haben, so beträgt der Abstand zwischen ihnen $\left|1-4\right|$ Notenstufen. Der schlechtere Schüler müßte sich um 3 Noten steigern, um mit dem besseren gleichzuziehen. Erhöht sich die Anzahl der Ausprägungen, erhöht sich ceteris paribus die Streuung, da jetzt stärker differenziert wird. Dabei kann die Anzahl *k* der Ausprägungen bei Vorliegen sehr vieler Bindungen erheblich kleiner als die Objektanzahl n sein. Etwa können $n = 100$ Studenten in einer Klausur die Noten 1, 2, 3, 4, 5 und 6 = *k* erhalten.

Ist die Zahl der Objekte dagegen recht klein, jedoch die verwendete Skala sehr differenziert, so kann k auch größer als n sein. So wird etwa eine Notenskala, die von 1,0; 1,25; 1,5; 1,75 usw. bis 4,75 reicht ($k = 16$), durch $n = 5$ Prüflinge nicht ausgeschöpft. Es bleiben also zwangsläufig einige Merkmalsausprägungen unbesetzt.

---

[6] auch: zweiwertige oder dichotome

Die zugehörigen Daten eines Merkmals lassen sich in einer Tabelle mit absoluten Häufigkeiten darstellen (Abb. 4). Die Tabelle enthält in der linken Spalte die transformierten Ausprägungen und in der rechten Spalte jeweils die Anzahl der Objekte, welche gerade diese Ausprägung besitzen.

Abb. 4: Häufigkeitstabelle eines diskret-ordinalen Merkmals

r	n_r
1	n_1
2	n_2
⋮	⋮
k	n_k

2.2.3. Stetig-ordinale Merkmale

Bei ordinalen Daten eines stetigen Merkmals können, wenn keine Bindungen vorliegen, die Ausprägungen in eine eindeutige Reihenfolge gebracht werden. Es bezeichne $R(x_i)$ den Rang, also die Position des i -ten Objekts in dieser Folge. Dann ist $R(x_1)$, $R(x_2)$, ..., $R(x_n)$ eine Permutation der Zahlen 1, 2, ..., n. Damit sei der Abstand zwischen dem i -ten und j-ten Objekt erklärt als der Differenzbetrag der Ränge

$$d_{ij} = \left| R(x_i) - R(x_j) \right| \qquad (i, j = 1, \ldots, n).$$

Im Falle einer Bindung, die sich etwa durch eine Meßungenauigkeit erklären ließe, könnte eine willkürliche Rangfolge festgelegt werden, um trotzdem obige Abstandsdefinition zu verwenden. Systematischer ist es aber, die nun einmal aufgetretene Bindung beizubehalten und den betroffenen Objekten einen gemeinsamen mittleren Rang zuzuweisen. Wählt man dabei das arithmetische Mittel der möglichen Ränge, so erhält man dieselbe Rangsumme wie im Fall nicht vorliegender Bindungen, nämlich

$$\sum_{i=1}^{n} R(x_i) = \frac{n(n+1)}{2}.$$

Hat man viele Bindungen, ist es sinnvoll, die Daten wie im diskreten Fall als Häufigkeitstabelle darzustellen (Abb. 5), wobei es sich im stetigen Fall methodisch um eine Klassierung der Daten handelt, weil man jede Bindung als Verlust einer theoretisch existierenden Information über die exakte Rangfolge auffaßt. Die Verwendung des arithmetischen Mittels führt zu den in der letzten Spalte angegebenen mittleren Rängen.

Abb. 5: Häufigkeitstabelle eines stetig-ordinalen Merkmals bei Auftreten von Bindungen

r	n_r	R
1	n_1	$\frac{n_1+1}{2}$
2	n_2	$n_1+\frac{n_2+1}{2}$
$\vdots$	$\vdots$	$\vdots$
k	n_k	$n_1+n_2+\ldots+n_{k-1}+\frac{n_k+1}{2}$

Der Abstand d_{ij} zweier Objekte Nr. i und Nr. j bestimmt sich dann folgendermaßen (dabei sei, um die Betragsstriche weglassen zu können, $r_i > r_j$ angenommen):

$$
\begin{aligned}
d_{ij} &= \left|R(x_i)-R(x_j)\right| \\
&= \left(\sum_{l=1}^{r_i-1} n_l + \frac{n_{r_i}+1}{2}\right) - \left(\sum_{l=1}^{r_j-1} n_l + \frac{n_{r_j}+1}{2}\right) \\
&= \sum_{l=r_j+1}^{r_i-1} n_l + n_{r_j} + \frac{n_{r_i}-n_{r_j}}{2} \\
&= \sum_{l=r_j+1}^{r_i-1} n_l + \frac{n_{r_i}+n_{r_j}}{2} \qquad (i,j=1,\ldots,n).
\end{aligned}
$$

Der Abstand ist also gleich der Anzahl aller Objekte, deren Ausprägungen dazwischen liegen, zuzüglich dem arithmetischen Mittel aus der Anzahl der Objekte der beiden betroffenen Ausprägungen. Im Gegensatz zum diskret-ordinalen Merkmal gehen also auch die Ausprägungen bei anderen als den beiden betrachteten Objekten in die Berechnung des Abstandes ein. Daher ist dies eine ex-post-Distanz.

2.2.4. Metrische Merkmale

Als Abstand wird der Absolutbetrag der Differenz der Ausprägungen verwendet. Das liegt nahe, da nach Definition des metrischen Skalenniveaus eben diese Differenz inhaltlich interpretierbar ist. Also

$$d_{ij} = \left|x_i - x_j\right| \qquad (i,j=1,\ldots,n).$$

Es gilt dann die Dreiecksungleichung

$$d_{ij} \le d_{ih} + d_{hj} \qquad (h,i,j = 1,\ldots,n),$$

die besagt, daß der Abstand zwischen zwei Objekten stets kleiner oder gleich der Summe der Abstände beider Objekte zu einem dritten Objekt ist.

Eine mögliche Alternative zum Absolutbetrag ist das Quadrat der Differenz

$$d_{ij} = \left(x_i - x_j\right)^2 \qquad (i,j = 1,\ldots,n).$$

Durch die Quadrierung werden größere Differenzen gegenüber kleineren überproportional gewichtet. Allerdings ist bei Verwendung des Quadrats die Dreiecksungleichung nicht mehr gültig, denn man erhält beispielsweise für die Ausprägungen 1, 2 und 4

$$(1-4)^2 > (1-2)^2 + (2-4)^2 .$$

2.3. Mittlerer Abstand

Ein Maß für die Streuung eines Merkmals erhält man durch Bildung des arithmetischen Mittels zwischen allen geordneten Paaren verschiedener Objekte, also indem man die Summe aller Abstände durch deren Anzahl teilt:

$$D_G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} .$$

In der Notation D_G für diesen mittleren Abstand erinnert das tiefgestellte G an Corrado Gini, der für metrische Merkmale den mittleren Abstand eingeführt hat. Der mittlere Abstand ist gerade das arithmetische Mittel aller Einträge der Distanzmatrix außerhalb der Diagonale.

Bei allen erwähnten Abständen ist die Distanzmatrix symmetrisch und ist deren Hauptdiagonale nur mit Nullen besetzt, also

$$d_{ij} = d_{ji}, \qquad d_{ii} = 0 \qquad (i,j = 1,\ldots,n).$$

Daher genügt es, zur Berechnung von D_G nur diejenigen Paare von Indizes (i,j) zu summieren, für welche $j > i$ gilt. Davon gibt es genau $\binom{n}{2}$ Stück, weshalb gilt

$$D_G = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} .$$

Der mittlere Abstand läßt sich auch auf naheliegende Weise für eine Zufallsvariable X definieren. Seien nämlich X_1, X_2, ..., X_n weitere, wie X verteilte, stochastisch unabhängige Zufallsvariablen. Dann sei der mittlere Abstand von X als der Erwartungswert des mittleren Abstandes der einzelnen Zufallsvariablen erklärt, also

$$D_G(X) = E\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} d(X_i, X_j)\right)$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(d(X_i, X_j)\right),$$

wobei $d\left(x_i, x_j\right)$ der Abstand zweier Ausprägungen x_i und x_j bezüglich des vorliegenden Skalenniveaus ist.

Der mittlere Abstand hängt bei gegebener Anzahl n nur von der Verteilung der betrachteten Zufallsvariable ab, das heißt Variablen mit derselben Verteilung haben auch übereinstimmenden mittleren Abstand.

Für nominales, diskret-ordinales und metrisches Skalenniveau ist der mittlere Abstand sogar unabhängig von der Anzahl n, da wegen der ex-ante-Eigenschaft die Einzelabstände zwischen je zwei Merkmalsausprägungen nicht von der Menge aller betrachteten Objekte abhängen. Es ist dann für jede wie X verteilte, jedoch von X stochastisch unabhängige Zufallsvariable Y der mittlere Abstand gleich dem Erwartungswert des Abstandes von X und Y, also

$$D_G(X) = E(d(X,Y)) .$$

In diesem Kontext dient der mittlere Abstand einer Zufallsvariablen als Lieferant für abstrakte Beispiele. Zu jedem Skalenniveau wird er im folgenden für jeweils eine typische Verteilung bestimmt, so daß man Unterschiede und Gemeinsamkeiten mit anderen Streuungsmaßen erkennen kann. Die Verwendung konkreter Beispiele aus dem realwissenschaftlichen Bereich wäre zwar stärker an der praktischen Anwendung orientiert, hätte aber den Nachteil, daß man die Resultate nicht unmittelbar überprüfen könnte, sondern erst die Daten zu beschaffen und aufwendig auszuwerten wären. In diesem Sinne wird hier auf die Erfahrung der angewandten Statistik zurückgegriffen, daß die benutzten, als typisch betrachteten Verteilungen das Ergebnis vieler realer Erhebungen zutreffend beschreiben können.

2.3.1. Nominales Merkmal

Mit dem oben gewählten Abstand für nominale Merkmale ist der mittlere Abstand gleich dem Anteil derjenigen Paare von Objekten mit unterschiedlicher Merkmalsausprägung an allen Paaren, also in Formeln

$$D_G = \frac{\left|\left\{(i,j)\middle| x_i \neq x_j\right\}\right|}{n(n-1)} \qquad \text{bzw.} \qquad D_G = \frac{\left|\left\{(i,j)\middle|\ i<j,\quad x_i \neq x_j\right\}\right|}{\binom{n}{2}},$$

wobei die $\left|\{\ \}\right|$-Striche die Anzahl der Elemente mit der angegebenen Eigenschaft bezeichnen.

Betrachtet sei zunächst ein binäres Merkmal *X*, das Bernoulli-verteilt ist mit $P(X=1) = 1-P(X=0) = \pi$. Für den mittleren Abstand von *X* erhält man unter Verwendung einer weiteren, wie *X* verteilten, Zufallsvariablen *Y* die Gleichheit

$$\begin{aligned} D_G(X) &= E(|X-Y|) \\ &= 0\pi^2 + 1\pi(1-\pi) + 1(1-\pi)\pi + 0(1-\pi)^2 \\ &= 2\pi(1-\pi). \end{aligned}$$

Da die Varianz gleich $\sigma^2 = \pi(1-\pi)$ ist, stimmt ihr Doppeltes also mit dem mittleren Abstand überein: $D_G(X) = 2\sigma^2$. Er ist insbesondere genau dann am größten mit dem Wert ein Halb, falls erwartungsgemäß die Hälfte aller Objekte die eine Ausprägung aufweisen und die andere Hälfte die andere, also $\pi = \frac{1}{2}$ gilt.

Hat die nominale Zufallsvariable X mehr als zwei Ausprägungen, etwa in der Skala 1, 2, ..., k, so gilt mit $P(X=r) = \pi_r$ für r =1, ..., k die Formel (vgl. USCHNER 1989: 48)

$$D_G(X) = \sum_{r=1}^{k} \pi_r(1-\pi_r) = 1 - \sum_{r=1}^{k} \pi_r{}^2 .$$

2.3.2. Diskret-ordinales Merkmal

Hat das Merkmal X die diskret-ordinale Skala 1, 2, ..., k, so gilt mit der Notation $P(X=r) = \pi_r$ für r =1, ..., k die Gleichheit

$$D_G(X) = \sum_{q=1}^{k}\sum_{r=1}^{k} \pi_q \pi_r |q-r|.$$

Man beachte den Unterschied gegenüber dem Fall eines nominalen Merkmals!

Zur Vorbereitung der Bestimmung des mittleren Abstandes dient das

Lemma 1: Es gilt die Summenformel

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|i-j| = \frac{n\left(n^2-1\right)}{3}$$

Beweis: Ordnet man die Abstände $|i-j|$ in einer Tabelle an (Abb. 6), so läßt sich leicht eine Formel zur Berechnung der Doppelsumme finden.

Abb. 6: Hilfstabelle zum Beweis einer Summenformel

	1	2	3	...	*n*
1	0	1	2	...	*n*-1
2	1	0	1	...	*n*-2
3	2	1	0	...	*n*-3
⋮	⋮	⋮	⋮	⋱	⋮
n	*n*-1	*n*-2	*n*-3	...	0

Es genügt nämlich, da die Einträge symmetrisch zur Hauptdiagonalen sind, nur die oberen Nebendiagonalen aufzusummieren und das Ergebnis mit zwei zu multiplizieren. Man erhält mit Hilfe der Summenformeln (BRONSTEIN U. SEMENDJAJEW 1985: 114)

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

die Berechnung

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}|i-j| &= 2\sum_{k=1}^{n-1}(n-k)k \\
&= 2\left(n\sum_{k=1}^{n-1}k - \sum_{k=1}^{n-1}k^2\right) \\
&= 2\left(n\frac{(n-1)n}{2} - \frac{(n-1)n(2n-1)}{6}\right) \\
&= n(n-1)\frac{n+1}{3} = \frac{n\left(n^2-1\right)}{3}
\end{aligned}$$

q.e.d.

Zur spezielleren Analyse sei nun der Fall betrachtet, daß jede der k Ausprägungen bei gleich vielen Objekten angenommen wird, also $\pi_r = \frac{1}{k}$ für alle r = 1, ..., k gilt. Soll dies exakt erfüllt sein, muß die Objektzahl n durch k teilbar sein. Dann folgt mit Lemma 1 die Beziehung

$$D_G(X) = \sum_{q=1}^{k}\sum_{r=1}^{k}\frac{1}{k^2}|q-r|$$

$$= \frac{k\left(k^2-1\right)}{3k^2} = \frac{k}{3} - \frac{1}{3k}$$

Ist insbesondere die Anzahl k möglicher Merkmalsausprägungen groß, so gilt also im Spezialfall der Gleichverteilung näherungsweise $D_G(X) \approx \frac{k}{3}$, die mittlere Differenz wächst also linear mit k.

2.3.3. Stetig-ordinales Merkmal

Da Bindungen methodisch bei einem stetigen Merkmal nicht berücksichtigt werden müssen, folgt für den mittleren Abstand direkt aus Lemma 1

$$D_G(X) = \frac{1}{n(n-1)}\frac{n\left(n^2-1\right)}{3} = \frac{n+1}{3}.$$

Insbesondere wächst die mittlere Differenz linear mit n. Liegen also keine Bindungen vor, so hängt der mittlere Abstand bei stetig-ordinalem Skalenniveau nur von der Anzahl n der Objekte ab und ist damit unabhängig von den numerischen Werten der Ausprägungen.

2.3.4. Metrisches Merkmal

Für die Normalverteilung erweist sich der mittlere Abstand - bis auf einen konstanten Faktor - als identisch mit der Standardabweichung:

Lemma 2: Ist X normalverteilt mit der Standardabweichung σ, so gilt für den mittleren Abstand

$$D_G(X) = \frac{2}{\sqrt{\pi}}\sigma$$

(das ist ungefähr 1,128 σ).

Beweis: Seien X und Y zwei unabhängige, normalverteilte Zufallsvariablen mit gemeinsamem Erwartungswert μ und Standardabweichung σ. Da X und Y unabhängig sind, ist die Differenz $X-Y$ normalverteilt mit Erwartungswert Null und Varianz $2\sigma^2$. Damit ist die Zufallsvariable $Z := \frac{X-Y}{\sqrt{2\sigma^2}}$ standardnormalverteilt. Wird mit f die Dichtefunktion der Standardnormalverteilung bezeichnet, so erhält man

$$E(|Z|) = \int_{-\infty}^{+\infty} |x| f(x)\,\mathrm{d}x = 2\int_{0}^{+\infty} x f(x)\,\mathrm{d}x$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} x e^{-\frac{1}{2}x^2}\,\mathrm{d}x = \frac{2}{\sqrt{2\pi}} \left[-e^{-\frac{1}{2}x^2} \right]_{x=0}^{\infty}$$

$$= \frac{2}{\sqrt{2\pi}}$$

Daraus folgt

$$E(|X - Y|) = \sqrt{2\sigma^2}\, E(|Z|) = \frac{2}{\sqrt{\pi}}\sigma \qquad \text{q.e.d.}$$

Die bewiesene Gleichheit ist erstaunlich, da in der üblichen Berechnung der Standardabweichung die Quadrate, und nicht die Absolutbeträge der Differenzen, summiert werden. Dennoch unterscheiden sich beide Streuungsmaße nur um einen konstanten Faktor.

2.4. Aggregierung der Abstände

Die Abstände $d_{ij;h}$ zwischen je zwei Objekten bezüglich des h-ten Merkmals werden zu einem Abstand d_{ij}^{norm} zusammengefaßt, der die Ausprägungen aller Merkmale berücksichtigt. Dazu wird eine bestimmte lineare Aggregierung verwendet, das heißt es werden Gewichte α_1, ..., α_m gewählt, so daß sich der aggregierte Abstand d_{ij}^{norm} zwischen dem i-ten und j-ten Objekt ergibt als

$$d_{ij}^{\text{norm}} = \sum_{h=1}^{m} \alpha_h \; d_{ij;h} \,.$$

Die Koeffizienten α_h werden so bestimmt, daß jeweils der mittlere Abstand des h-ten Merkmals auf den Wert Eins normiert ist, also

$$D_{G;h}^{\text{norm}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_h \; d_{ij;h} = 1 \,.$$

Wenn man obige Gleichung nach α_h auflöst, erhält man

$$\alpha_h = \frac{n(n-1)}{\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij;h}} = \frac{1}{D_{G;h}} \,.$$

Damit ist das Gewicht eines Merkmals gerade der Kehrwert seines mittleren Abstandes. Die Aggregierung der Einzelabstände ergibt sich somit als

$$d_{ij}^{\text{norm}} = \sum_{h=1}^{m} \frac{d_{ij;h}}{D_{G;h}}.$$

Für den mittleren Abstand über alle Merkmale gilt insbesondere nach Wahl der Gewichte

$$D_G^{\text{norm}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^{\text{norm}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{h=1}^{m} \frac{d_{ij;h}}{D_{G;h}} = \sum_{h=1}^{m} \frac{D_{G;h}}{D_{G;h}} = m.$$

Er ist also gleich der Anzahl der Merkmale und damit insbesondere unabhängig von den Ausprägungen und von der Zahl der Objekte. Jedes Merkmal erhält im aggregierten Abstand dasselbe Gewicht in dem Sinne, daß seine Streuung, gemessen durch den mittleren Abstand, vereinheitlicht wird. Die Normierung führt überdies zu dimensionslosen Distanzen, die folglich auch bei Merkmalen unterschiedlichen Skalenniveaus und unterschiedlicher Maßeinheiten aggregiert werden können.

Da die Streuung eines Merkmals vom Gewicht abhängt, kann der Abstand zweier Objekte im allgemeinen nur bestimmt werden, wenn die Streuungen aller Merkmale bereits bekannt sind. Weil aber die Streuung von der Gesamtheit abhängt, gehen alle Objekte in die Berechnung jedes einzelnen Abstandes ein. Die Abstände können damit nicht ex-ante festgelegt werden, gehören also zu einer ex-post-Distanzmatrix.

3. Bildung von Klassen

3.1. Streuungszerlegung

Eine gegebene Klassifikation, bestehend aus s Klassen C_1, ..., C_s, gibt Anlaß zu der im folgenden beschriebenen Zerlegung der normierten Gesamtstreuung $D = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^{\text{norm}} = mn(n-1)$.

Dazu sei die Summe aller Abstände innerhalb der k-ten Klasse

$$DI(C_k) = \sum_{i,j \in C_k} d_{ij}^{\text{norm}} \qquad (k = 1, \ldots, s)$$

als die Streuung innerhalb dieser Klasse bezeichnet und Streuung zwischen der k-ten Klasse mit den anderen Klassen sei die Summe

$$DZ(C_k) = \sum_{\substack{i \in C_k \\ j \notin C_k}} d_{ij}^{\text{norm}} \qquad (k = 1, \ldots, s)$$

genannt. Es gilt dann nämlich

Lemma 3: Ist $DI = \sum_{k=1}^{s} DI(C_k)$ die Streuung innerhalb der Klassen und $DZ = \sum_{k=1}^{s} DZ(C_k)$ die Streuung zwischen den Klassen, so ist die Summe aus beiden gerade die Gesamtstreuung D, das heißt, man hat die Streuungszerlegung $D = DI + DZ$.

Beweis: Da die Klassen disjunkt sind, je zwei verschiedene Klassen also keine gemeinsamen Objekte besitzen, gilt

$$\begin{aligned}
D &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^{\text{norm}} = \sum_{k=1}^{s}\sum_{i \in C_k}\sum_{j=1}^{n} d_{ij}^{\text{norm}} \\
&= \sum_{k=1}^{s}\sum_{i \in C_k}\left(\sum_{j \in C_k} d_{ij}^{\text{norm}} + \sum_{j \notin C_k} d_{ij}^{\text{norm}} \right) \\
&= \sum_{k=1}^{s}\sum_{i,\, j \in C_k} d_{ij}^{\text{norm}} + \sum_{k=1}^{s}\sum_{\substack{i \in C_k \\ j \notin C_k}} d_{ij}^{\text{norm}} \\
&= \sum_{k=1}^{s} DI(C_k) + \sum_{k=1}^{s} DZ(C_k) = DI + DZ \qquad \text{q.e.d.}
\end{aligned}$$

Mittels der Streuungszerlegung läßt sich auf naheliegende Weise die Trenneigenschaft einer Klassifizierung beurteilen. Faßt man nämlich die Streuung innerhalb einer bestimmten Klasse als Maß für die Heterogenität dieser Klasse auf, so liefert die Summe aller Innerklassenverschiedenheiten DI ein Maß für die Güte der gesamten Klassifikation in Bezug auf die gegebenen Abstände: Je kleiner DI, desto höher ist die Güte (vgl. STEINHAUSEN u. LANGER 1977: 100).

Dabei ist $DI = 0$, falls alle Klassen nur jeweils ein Objekt umfassen, also $n = s$ gilt. Falls $s = 1$, also eine einzige Klasse alle Objekte enthält, so hat man $DZ = 0$ und daraus $D = DI$, da die Gesamtstreuung gleich der Streuung innerhalb der einzigen Klasse ist. Da stets $DI \geq 0$ gilt, erhält man für die Streuung zwischen den Klassen DZ die Grenzen $0 \leq DZ \leq D$. In Analogie zur Streuungszerlegung bei einer linearen Regression kann also der Quotient

$$R^2 = \frac{DZ}{D}$$

als Bestimmtheitsmaß der Klassifikation aufgefaßt werden. Es gibt den Anteil der Zwischenklassenstreuung an der Gesamtstreuung an, insbesondere gilt $0 \leq R^2 \leq 1$. Das Bestimmtheitsmaß R^2 ist null, falls es nur eine Klasse gibt, und gleich Eins, falls es soviele Klassen wie Objekte gibt. Generell ist es hoch, falls die Streuung zwischen den Klassen groß gegenüber der Streuung in den Klassen ist, also falls die Klassen selbst relativ homogen sind, die Klassen untereinander sich jedoch deutlich unterscheiden.

3.2. Klassifizierungsverfahren

Der Prozeß der Klassenbildung unterscheidet sich bei Verwendung der beschriebenen Distanzmessung nur wenig von den bekannten Vorgehensweisen. So sind bei den hierarchisch-agglomerativen Verfahren ohne weitere Modifikation Single-Linkage[7], Complete-Linkage[8] und Average-Linkage möglich (vgl. NORUŠIS 1993: 97ff.).

Bei Single-Linkage werden die zwei Klassen C_k und C_l fusioniert, die am nähesten beieinander liegende Objekte besitzen, also bei denen der Abstand $\min_{i \in C_k;\, j \in C_l} d_{ij}^{\text{norm}}$ minimal ist. Im Gegensatz dazu fusioniert Complete-Linkage die beiden Klassen, bei denen das Abstandsmaximum $\max_{i \in C_k;\, j \in C_l} d_{ij}^{\text{norm}}$ unter allen Paaren von Klassen den niedrigsten Wert hat. Das Verfahren Single-Linkage neigt dazu, Ketten zu bilden, wohingegen bei Complete-Linkage eher kugelförmige Klassen entstehen können.

Werden die Klassen fusioniert, bei denen der absolute klasseninterne Streuungszuwachs minimal ist, besteht die Tendenz, gleichgroße Klassen zu bilden. Dieses Vorgehen entspricht dem Verfahren von Ward (vgl. BACKHAUS u.a. 1993: 298). Soll dagegen stärker auf die Homogenität in den Klassen abgestellt werden, empfiehlt es sich, den klassendurchschnittlichen Streuungszuwachs zu minimieren, also die Summe der mittleren Abstände innerhalb der Klassen

$$\overline{DI} = \sum_{k=1}^{s} \frac{DI(C_k)}{n_k (n_k - 1)} = \sum_{k=1}^{s} D_G(C_k)$$

als Kriterium zu verwenden (dabei bezeichnet $n_k = |C_k|$ die Anzahl der Objekte in der k-ten Klasse). Dies ist das Verfahren Average-Linkage (innerhalb der Gruppen).

Average-Linkage ist in gewisser Weise ein Kompromiß zwischen Single- und Complete-Linkage, mit dem bei der Klassenbildung extreme Effekte eher vermieden werden können. Es

---

7 auch: Methode des nächstgelegenen Nachbarn

8 auch: Methode des entferntesten Nachbarn

berücksichtigt bei der Auswahl der zu fusionierenden Klassen nicht Minima oder Maxima von Abständen, sondern Durchschnittswerte von mehreren Objektpaaren. Konkret werden bei Average-Linkage (innerhalb der Gruppen) die beiden Klassen fusioniert, bei denen der mittlere Abstand innerhalb der entstehenden Klasse

$$D_G(C_k \cup C_l) = \frac{DI(C_k \cup C_l)}{(n_k + n_l)(n_k + n_l - 1)} \qquad (k, l = 1, \ldots, s)$$

minimal ist (vgl. SPSS INC. 1991: 32). In der Formel bezeichnet n_k bzw. n_l die Anzahl der Objekte in der Klasse C_k bzw. C_l, womit wegen deren Disjunktheit $n_k + n_l$ die Gesamtzahl der Objekte in der fusionierten Klasse $C_k \cup C_l$ ist.

Die gesamte Innerklassenstreuung DI wird durch die Zusammenfassung der Klassen C_k und C_l gerade um die Differenz

$$DI(C_k \cup C_l) - \left(DI(C_k) + DI(C_l)\right) = 2 \sum_{\substack{i \in C_k \\ j \in C_l}} d_{ij}^{\text{norm}}$$

aus gemeinsamer Innerklassenstreuung und den einzelnen Innerklassenstreuungen erhöht. Dies ist gerade die Streuung zwischen den beiden Klassen, wobei der Faktor Zwei auf der rechten Seite in obiger Formel nötig ist, da bei der Innerklassenstreuung Paare doppelt gezählt werden. Das Bestimmtheitsmaß R^2 nimmt entsprechend ab, das heißt um die mit dem Faktor $1/D$ multiplizierte Differenz.

Bei einem Schritt von Average-Linkage erfolgt nicht notwendigerweise die Fusion derjenigen beiden Klassen, bei welchen das Bestimmtheitsmaß nur geringstmöglich abnimmt. Durch die Mittelung der Streuung zwischen den Klassen werden nämlich solche mit vielen Elementen eher zusammengefaßt, da dann die Streuungssumme durch einen größeren Klassenbesetzungswert $(n_k + n_l)(n_k + n_l - 1)$ geteilt wird. Ausreißer bleiben damit länger isoliert und es entsteht keine Tendenz zur Bildung von gleichgroßen Klassen.

Auch ein dem Centroid-Sorting entsprechendes Verfahren ist denkbar (vgl. ANDERBERG 1973: 160ff.). Allerdings müßte dafür für jede Klasse jeweils ein Zentroid bestimmt werden. Dazu wären dem Skalenniveau entsprechende Klassenmittelwerte zu bestimmen, also etwa der Modus für nominale, der Median für ordinale und der arithmetische Mittelwert für metrische Merkmale. Sind Iterationszyklen zur Verbesserung einer ersten Zuordnung geplant, erscheint es zweckmäßig, vor jedem neuen Durchgang geeignete Startgrößen zu bestimmen. Dies können unter anderem die Klassenzentroide sein, oder besser noch, die Objekte, die diesen Zentroiden am nächsten liegen. Dies hat den Vorteil, daß die Abstände nicht jeweils neu berechnet

werden müssen. Außerdem ist das Ergebnis anschaulicher, da man den Zentroid als typischen Vertreter seiner Klasse interpretieren kann.

Auch hierfür ist es ratsam, die vorgegebene Klassenanzahl zu variieren, um durch einen Vergleich die optimale Einteilung zu finden. Entscheidungshilfe leistet hierbei das weiter vorne angeführte Bestimmtheitsmaß.

Ist eine nicht-erschöpfende Klassifikation beabsichtigt, sollen also eventuelle Ausreißer isoliert werden, empfiehlt sich vor der Klassifizierung eine einfache Ausreißerdiagnose. Zu diesem Zweck sind lediglich die Zeilen der Ausgangsmatrix aufzusummieren. Diese Zeilensummen geben dann das $(n-1)$-fache des mittleren Abstandes des betreffenden Objekts zu allen übrigen Objekten an. Summen, die weit über denen der meisten anderen liegen, weisen auf potentielle Ausreißer hin. Es ist Ermessenssache festzulegen, ab wann ein Objekt als Ausreißer bezeichnet wird.

Im übrigen läßt die Distanzmatrix auch erkennen, ob es sich um isolierte Ausreißer handelt oder ob etwa zwei benachbarte Ausreißer vorliegen. Bei isolierten Ausreißern sind alle Einzeldistanzen groß. Einzelne kleine Abstände deuten dagegen auf benachbarte Objekte hin.

3.3. Eine konkrete Klassifikation

Zur Demonstration des beschriebenen Vorgehens soll ein aus der Literatur bekanntes empirisches Beispiel verwendet werden: die Klassifikation von Staaten nach ihrem Entwicklungsstand, wie sie VOGEL (1992, 1993) präsentiert hat. Insgesamt 112 Länder werden anhand von 57 Variablen unterschiedlichen Skalenniveaus aus den vier Bereichen

- Wirtschaft,
- Demographie,
- Politik,
- Grundbedürfnisse und soziale Faktoren

klassifiziert, um eine möglichst vielfältige Erfassung des Entwicklungsstands zu gewährleisten. Da diese vier Bereiche jedoch unterschiedlich stark durch Merkmale repräsentiert werden, und außerdem die Merkmale nicht alle die gleiche Relevanz besitzen, nimmt VOGEL eine Gewichtung vor, die hier jedoch nicht nachvollzogen werden soll. Aus diesem Grund können und sollen die Ergebnisse hier nicht mit denen von VOGEL verglichen werden.

Die Klassifikation erfolgte mit Hilfe des Programmpakets SPSS/PC+. Für die Ermittlung der Distanzmatrix sowie für die Klassendiagnose wurde von Karlheinz Wunner ein ergänzendes Programm geschrieben.

Von den verschiedenen Klassifikationsalgorithmen ergab Average-Linkage (innerhalb der Gruppen), also das Verfahren, bei dem die klasseninterne Streuung minimiert wird, die besten, das heißt anschaulichsten, Ergebnisse.

Folgende fünf Klassen wurden gebildet:

- *Klasse 1 (27 Länder):* Ägypten, Birma, Elfenbeinküste, Guyana, Haiti, Hongkong, Indien, Indonesien, Kamerun, Kenia, Kongo, Liberia, Madagaskar, Malawi, Malaysia, Marokko, Mauretanien, Pakistan, Papua-Neuguinea, Sambia, Senegal, Sierra Leone, Simbabwe, Tansania, Togo, Tunesien, Zaire

- *Klasse 2 (14 Länder):* Äthiopien, Bangladesh, Benin, Burundi, Ghana, Jordanien, Nepal, Nigeria, Obervolta (Burkina Faso), Ruanda, Sudan, Tschad, Uganda, Zentralafrikanische Republik

- *Klasse 3 (32 Länder):* Argentinien, Bahamas, Barbados, Bolivien, Brasilien, Chile, Costa Rica, Dominikanische Republik, Equador, El Salvador, Fidschi, Guadeloupe, Guatemala, Honduras, Jamaika, Kolumbien, Martinique, Mexiko, Nicaragua, Panama, Paraguay, Peru, Philippinen, Seychellen, Sri Lanka, Südkorea, Surinam, Syrien, Thailand, Tonga, Türkei, Uruguay

- *Klasse 4 (8 Länder):* Algerien, Gabun, Irak, Iran, Kuwait, Libyen, Saudi-Arabien, Trinidad/Tobago

- *Klasse 5 (31 Länder):* Australien, Belgien/Luxemburg, Bundesrepublik Deutschland, Dänemark, Finnland, Frankreich, Griechenland, Großbritannien, Irland, Israel, Italien, Japan, Jugoslawien, Kanada, Libanon, Malta, Neukaledonien, Neuseeland, Niederlande, Norwegen, Österreich, Portugal, Schweden, Singapur, Spanien, Südafrika, Tschechoslowakei, Ungarn, USA, Venezuela, Zypern

Auf den ersten Blick erscheint es negativ, daß es drei große und zwei kleine Klassen gibt. Man sollte allerdings bedenken, daß eine gleichmäßige Klassenaufteilung kein Selbstzweck sein kann, sondern daß die Homogenität der Klassen und damit die Möglichkeit der Typisierung im Vordergrund steht. Bleibt zu prüfen, wie gut die gefundenen Klassen diesem Ziel entsprechen.

Über alle 57 Merkmale ergibt sich ein Bestimmtheitsmaß von 82 %. Lediglich 18 % der Streuung liegt in den fünf Klassen, die folglich auf den ersten Blick als recht homogen angesehen werden können.

Die merkmalsspezifischen Bestimmtheitsmaße streuen um diesen Wert, doch beträgt die Spannweite gerade 14 Prozentpunkte. Das höchste Bestimmtheitsmaß weist die Variable "Anteil Einwohner/Arzt" mit 90 % auf, den niedrigsten Wert hat die Einwohnerdichte mit 76 %.

Die Aussage ist jedoch zu relativieren. Von den insgesamt 6216 Paarvergleichen der 112 Länder entfallen gerade 1431 auf die fünf Klassen, das sind also 23 %. Merkmale mit einem Bestimmtheitsmaß von 77 % oder weniger tragen also nichts zur Klassifikation bei. Das sind hier die Variablen "Verhältnis Importe/Bruttoinlandsprodukt", Anteil der Nahrungsmittelimporte und die Einwohnerdichte. Weitere 20 Merkmale haben ein Bestimmtheitsmaß von weniger als 80 %.

Da nur 23 % der Paarvergleiche in den vorliegenden fünf Klassen erfolgen, wäre bei zufälliger Aufteilung der Länder auf die Klassen folglich ein Bestimmtheitsmaß von 77 % zu erwarten. Das erreichte Bestimmtheitsmaß von 82 % bedeutet also gerade eine Steigerung der Homogenität durch die Klassierung von 5 %. Es fragt sich daher, ob ein solches Resultat den Aufwand lohnt.

Allerdings ist zu bedenken, daß ein Bestimmtheitsmaß von 100 % nach dieser Berechnung eine Verbesserung von 23 % gegeben hätte. Als Qualitätsmaßstab ist es daher wohl besser, den erreichten Homogenitätszuwachs zu beziehen auf den maximal möglichen. Das ergibt hier einen Wert von rund 20 %. Auch das ist weniger als zu erwarten war.

Möglich ist, daß im vorliegenden Falle die erwähnten, für eine Klassifikation eigentlich ungeeigneten Merkmale dazu führen, daß eine Reihe von Ländern mehr oder minder zufällig angeordnet wird. Das würde aber das Resultat verschlechtern. Wenn diese Überlegung richtig ist, müßte folglich eine Reduzierung des Merkmalkatalogs zu besseren, das heißt homogeneren Klassen führen.

Trotz dieser skeptischen Aussagen sollen die Ergebnisse der Analyse im Folgenden kurz vorgestellt werden. Zunächst zu den Klassen im einzelnen:

- *Klasse 1* umfaßt Entwicklungsländer mit geringen wirtschaftlichen Fortschritten. So beträgt das Pro-Kopf-Inlandsprodukt 493 US-Dollar, davon werden 32 % in der Landwirtschaft erzeugt, wo auch der weitaus größere Teil der Erwerbstätigen beschäftigt ist. Ein knappes Viertel der Exporte entfallen auf Erze und Metalle. Sie sorgen dafür, daß der Exportanteil am Inlandsprodukt mit 32 % etwa dem weltweiten Durchschnitt entspricht. Die Einwohnerdichte ist mit 199 recht hoch. Ebenso überdurchschnittlich ist auch die rohe Sterberate. Bei den politischen Verhältnissen zeigen sich vereinzelte Ansätze zu einer De-

mokratisierung. Bei den Grundbedürfnissen schließlich ergeben sich durchwegs unterdurchschnittliche Werte. So ist die Arztdichte gering und die Analphabetenquote recht hoch.

- *Klasse 2* umfaßt die ärmsten Länder der Welt . Das Bruttoinlandsprodukt pro Einwohner beträgt im Durchschnitt gerade einmal 339 US-Dollar im Jahr, das sind 15 % des Weltdurchschnitts. Davon stammen 50 % aus der Landwirtschaft. Zusätzliche Einnahmen durch Rohstoffexporte haben sie kaum. Einer wirtschaftlichen Entwicklung steht auch die sehr schlechte Verkehrsinfrastruktur entgegen. Ein Anteil von 81 % der Bevölkerung lebt auf dem Lande. Die Sterblichkeit ist hoch, die Verbrauch von Kalorien beträgt nur 80 % des Weltdurchschnitts und auch bei der ärztlichen Versorgung bilden diese Länder mit Abstand das Schlußlicht. Auch die politischen Verhältnisse sind weit von einer Demokratie entfernt.

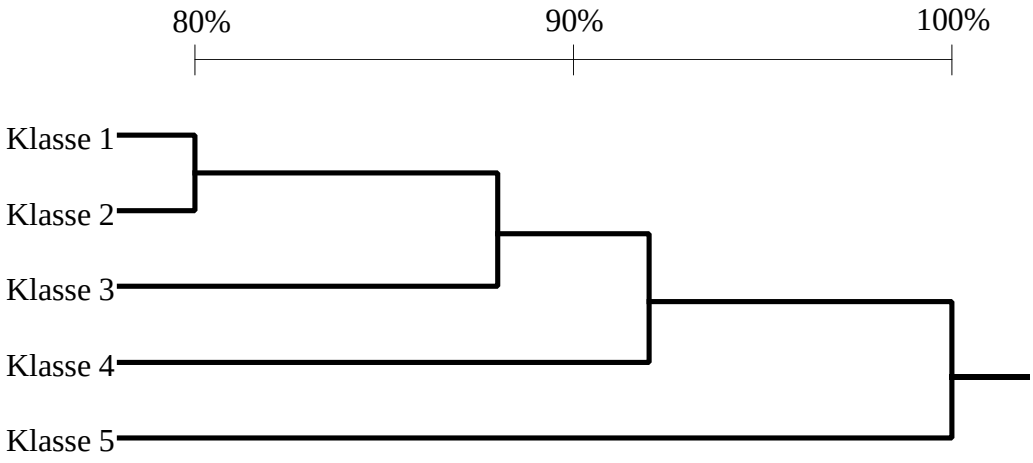
- *Klasse 3* enthält die Länder, die bereits ein höheres Niveau der wirtschaftlichen Entwicklung erreicht haben. Hierzu gehören die Schwellenländer wie Chile, Südkorea oder die Türkei, von denen einige in den letzten Jahren weitere erhebliche Fortschritte gemacht haben. Die Daten stammen aus den 80er Jahren. Nur noch 22 % des Inlandsprodukts wird in der Landwirtschaft erwirtschaftet. Das ist der weltweite Durchschnitt. Auch der Anteil der Beschäftigten in der Industrie entspricht ungefähr dem Weltmaßstab. Wie überhaupt diese Länder in vielen Bereichen "mittlere" Verhältnisse aufweisen. Allerdings ist die rohe Sterberate am niedrigsten, Folge einer jungen Bevölkerung bei gleichzeitig weit überdurchschnittlicher Arztdichte. Auch die politischen Verhältnisse sind vergleichsweise gut, der Bildungsstand der Bevölkerung ist recht hoch.

- *Klasse 4* sammelt im wesentlichen die erdölexportierenden Länder des mittleren Ostens. Durch ihre Bodenschätze haben sie mit 5.200 US-Dollar das höchste Einkommensniveau. Ihr Wohlstand beruht auf den Erdölexporten, die 94 % des Gesamtexports ausmachen. Die Einkommensverteilung fällt allerdings sehr ungleichmäßig aus, da 45 % der Bevölkerung auf dem Lande leben. Die politischen Verhältnisse sind autoritär, die Oppostion genießt nur geringe oder gar keine Einflußmöglichkeiten. Die medizinische Versorgung ist überdurchschnittlich. Das Bildungsniveau der Bevölkerung ist schlecht, 51 % der Einwohner sind Analphabeten.

- *Klasse 5* erfaßt die hochentwickelten Länder vorwiegend aus Europa. Das Bruttoinlandsprodukt beträgt mit 5.050 US-Dollar ungefähr das Doppelte des Weltniveaus. Die Sterblichkeit ist dank hoher Lebenserwartung und guter medizinischer Versorgung niedrig. Die politischen Verhältnisse sind, von Ausnahmen abgesehen, demokratisch.

Das Dendrogramm in Abb. 7 läßt die Nähe der gefundenen fünf Klassen zueinander, die mögliche weitere Fusionen bestimmen würde, erkennen. Die Achsenbeschriftung gibt den Abstand der auf dieser Stufe zusammenzufassenden Klassen bezogen auf den des letzten Schritts an. Danach würden bei 80 % des Maximalabstandes die Entwicklungsländer mit geringen wirtschaftlichen Fortschritten (Klasse 1) und die ärmsten Länder der Welt (Klasse 2) vereinigt. Dazu kämen der Reihe nach Klasse 3 und Klasse 4 hinzu, womit die 112 Länder in 81 nichthochentwickelte Länder einerseits und die 31 hochentwickelten Länder (Klasse 5) andererseits aufgeteilt wären. Wie jedes hierarchisch-agglomerative Verfahren, so würde auch Average-Linkage im letzten Schritt alle Objekte in einer einzigen Klasse vereinen.

Abb. 7: Dendrogramm



Die Zusammenfassung der Länder zu Klassen und die Nähe dieser Klassen zueinander ist im großen und ganzen plausibel, das heißt sie entspricht in etwa unseren Vorstellungen. Zu bedenken ist aber generell, daß die Daten zum Teil zehn Jahre alt sind. In der Zwischenzeit hat sich bei einer Reihe von Ländern eine erhebliche Veränderungen zum Besseren (Südkorea, Chile) oder zum Schlechteren (Jugoslawien, Irak) ergeben. Auch sind Datenfehler nicht auszuschließen. Der Umstand, daß die Daten von renommierten internationalen Institutionen veröffentlich wurden, sagt noch nichts über deren Qualität. In vielen Ländern, nicht nur in den Entwicklungsländern, weist die Statistik erhebliche Mängel auf. Grobe Schätzungen dürften die Regel sein. Sie sind oft nicht als solche zu erkennen und täuschen so eine nicht gewährleistete Genauigkeit vor. Bewußte Manipulationen zum Zwecke einer verbesserten Selbstdarstellung, etwa zur Hebung der Kreditwürdigkeit, sind nicht auszuschließen. Wie weit derartige Eingriffe gehen können, lehrt uns das Beispiel der ehemaligen Ostblockländer, deren Wirtschaftsdaten, wie man heute weiß, samt und sonders geschönt waren. Das erklärt beispielsweise auch, weshalb Ungarn unter den Industrieländern erscheint.

Das eigentliche Ziel der Beispiele zu demonstrieren, daß mit dem oben geschilderten Vorgehen Merkmale beliebigen Skalenniveaus zusammengefaßt werden können, wurde voll erreicht. So hat beispielsweise das nominale Merkmal "Art der Amtsübernahme des derzeitigen Chief-Executive" ein Bestimmtheitsmaß von 82 %, das ordinale Merkmal "Verhältnis Regierung und Parlament/Justiz" eines von 84 %.

Als positiv darf auch vermerkt werden, daß man mit dem neuen Verfahren ungeschminkt erkennt, welche Steigerung der Homogenität durch die Klassierung erreicht wird. Dieses gilt sowohl für die Klassen als ganzes als auch für die einzelnen Merkmale.

4. Literatur

ANDERBERG, M. R. *Cluster Analysis for Applications.* New York [u.a], 1973.

BACKHAUS, Klaus, [u.a.] *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung.* Berlin [u.a.], 1993.

BRONSTEIN, I. N., u. K. A. SEMENDJAJEW. *Taschenbuch der Mathematik.* Thun [u.a.], 1985.

DOBBENER, Reinhard. *Grundlagen der Numerischen Klassifikation anhand gemischter Merkmale.* Göttingen, 1983.

MUCHA, Hans-Joachim. *Clusteranalyse mit Mikrocomputern.* Berlin, 1992.

NORUŠIS, Marija J. *SPSS for Windows, Professional Statistics, Release 6.0.* Chicago, 1993.

SPÄTH, Helmuth. *Cluster-Formation und -Analyse:* Theorie, FORTRAN-Programme und Beispiele. München [u.a.], 1983.

SPSS INC., Hg. *SPSS Statistical Algorithms.* Chicago, 1991.

STEINHAUSEN, Detlef, u. Klaus LANGER. *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation.* Berlin [u.a.], 1977.

USCHNER, Helmut. *Streuungsmessung nominaler Merkmale mit Hilfe von Paarvergleichen.* Diss. Erlangen-Nürnberg, 1989.

VOGEL, Friedrich. "Some Remarks on a Classification of the Countries of the World According to their Stage of Development". *Jahrbücher für Nationalökonomie und Statistik* 211 (1993): 306-323.

VOGEL, Friedrich. "Underdevelopment - Development: Report on a Study of Classification of the Countries of the World According to Their Stage of Development". *Acta Demographica* 1992: 237-252.

VOGEL, Friedrich. *Probleme und Verfahren der numerischen Klassifikation.* Göttingen, 1975.

WÖRDENWEBER, Martin. *Clusteranalysen bei gemischtskalierten Datensätzen unter besonderer Berücksichtigung der Gewichtung von Variablen und Variablengruppen.* Diss. Münster, 1985.