Prof. Dr. Reinhold Kosfeld
Fachbereich Wirtschaftswissenschaften
Universität Kassel

Lösungen zu den Aufgaben zur Multivariaten Statistik
Teil 4: Aufgaben zur Clusteranalyse

1. Erläutern Sie, wie das Konstrukt der Ähnlichkeit von Objekten bei qualitativen und quantitativen Merkmalen in der Clusteranalyse operationalisiert wird!

Lösung:

- Ähnlichkeit bei qualitativen Merkmalen

Bei qualitativen Merkmalen wird die Ähnlichkeit von Objekten typischerweise durch Ähnlichkeitsmaße i.e.S. operationalisiert, die auf dem Intervall [0,1] normiert sind. Bei dichotomen Merkmalen wird dabei im Prinzip die Anzahl der Merkmale ermittelt, die eine Ähnlichkeit zweier Objekte erzeugen, die dann zu der in Frage kommenden Anzahl der Merkmale in Beziehung gesetzt werden. Unterschiedliche Operationalisierungen (=Konkretisierungen) ergeben insbesondere daraus, dass eine Eigenschaft bei einem Merkmal als interessierend hervorgehoben wird oder aber beide Eigenschaften als gleichwertig betrachtet werden können.

Bei polytomen Merkmalen können im Kern dieselben Messkonzepte nach einer Dichotomisierung eingesetzt werden. Hierbei ist aber als Modifikation eine Gewichtung einzuführen, um zu verhindern, dass ein originäres Merkmal mit mehreren Ausprägungen übergewichtet wird. Durch die Gewichtung wird zudem das Problem der durch eine Dichotomisierung erzeugten "künstlichen" Übereinstimmungen stark abgeschwächt.

- Ähnlichkeit bei quantitativen Merkmalen

Die Operationalisierung von Ähnlichkeiten zwischen Objekten erfolgt bei quantitativen Merkmalen typischerweise unter Verwendung von Distanzmaßen. Die direkte Distanzmessung erfolgt auf der Basis der euklidischen Distanz, während die City-Block-Distanz die Entfernung zwischen zwei Objekten bei rechtwinkligen Umwegen misst. Die Tschebyscheff-Distanz gibt dagegen stets die maximale Entfernung zweier Objekte entlang einer Achse des m-dimensionalen Zahlenraums an.

Weiterführende Hinweise: Bei dieser Art der Distanzmessung bleiben die Merkmalskorrelationen unberücksichtigt. Ihnen kann durch eine vorgeschaltete Faktorenanalyse oder einer Verwendung der Mahalanobis-Distanz Rechnung getragen werden.

2. Bei verschiedenen Kunden sind 6 binäre Klassifikationsmerkmale erhoben worden. Welche Werte nehmen der Simple-Matching-Koeffizient, der Jaccard-Koeffizient und der RR-Koeffizient für die ersten beiden Kunden an, deren Objektvektoren durch

   $\mathbf{x}_1 = (1\ \ 1\ \ 0\ \ 1\ \ 1\ \ 1)'$ und $\mathbf{x}_2 = (1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1)'$

   gegeben sind?

   Lösung:

Vierfeldertafel

Objekt 1	Objekt 2: 1	Objekt 2: 0	$\sum$
1	a = 2	b = 3	a + b = 5
0	c = 0	d = 1	c + d = 1
$\sum$	a + c = 2	b + d = 4	m = 6

$$\text{SMC} = \frac{a+d}{m} = \frac{2+1}{6} = \frac{3}{6} = 0{,}5$$

$$J = \frac{a}{a+b+c} = \frac{2}{2+3+0} = \frac{2}{5} = 0{,}4$$

$$\text{RR} = \frac{a}{m} = \frac{2}{6} = 0{,}333$$

3. Bestimmen Sie die in Aufg. 2 genannten Ähnlichkeitsmaße für die beiden ersten Kunden für den Fall, dass die vier letzten Komponenten der Objektvektoren Dummy-Variablen zweier dreiwertiger Merkmale sind, wenn die originären Merkmale als gleichwertig zu betrachten sind!

   Lösung:

   In diesem Fall müssen die Beiträge der beiden dreiwertigen Merkmale zu den Größen a, b, c und d der Vierfeldertafel des Objektvergleichs mit dem Faktor $\frac{1}{p-1} = \frac{1}{3-1} = \frac{1}{2}$ (p Anzahl der Merkmalskategorien) gewichtet werden. Man erhält dann

$$a = 1 + \frac{1}{2} \cdot 1 = 1{,}5$$

$$b = 1 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 2$$

$$c = 0, \quad d = \frac{1}{2} \cdot 1 = 0{,}5$$

$$\text{SCM} = \frac{a+d}{a+b+c+d} = \frac{1{,}5+0{,}5}{1{,}5+2+0+0{,}5} = \frac{2}{4} = 0{,}5$$

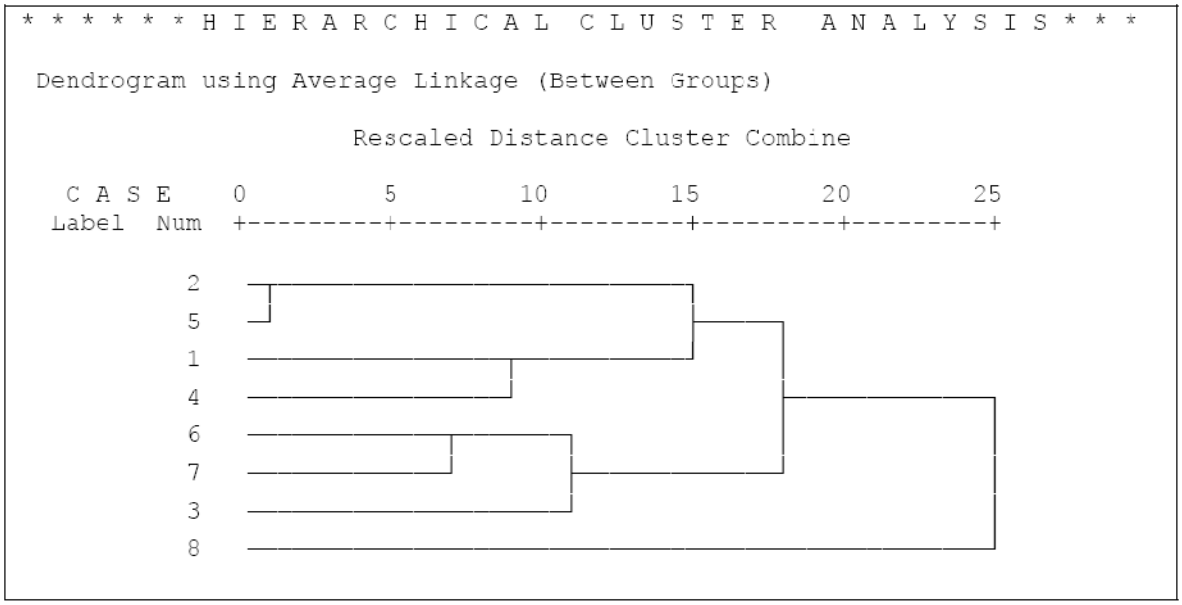
$$J = \frac{a}{a+b+c} = \frac{1{,}5}{1{,}5+2+0} = \frac{1{,}5}{3{,}5} = 0{,}429$$

$$\text{RR} = \frac{a}{a+b+c+d} = \frac{1{,}5}{1{,}5+2+0+0{,}5} = \frac{1{,}5}{4} = 0{,}375$$

4. Erläutern Sie den beiligenden SPSS-Output (Zuordnungsübersicht, Dendrogramm) einer hierarchischen Klassifikation von 8 Kunden! Wie lässt sich hieraus die Wahl einer adäquaten Clusterzahl begründen?

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koeffizienten (Jaccard)	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	5	1.000	0	0	5
2	6	7	.750	0	0	4
3	1	4	.667	0	0	5
4	3	6	.583	0	2	6
5	1	2	.417	3	1	6
6	1	3	.285	5	4	7
7	1	8	.000	6	0	0



Lösung:

Aus der Zuordnungsübersicht und dem Dendrogramm geht hervor, dass in den ersten drei Schritten der hierarchischen Klassifikation unter Einsatz des Average-Linkage-Verfahrens jeweils drei Zweiergruppen bei abnehmendem Ähnlichkeitsmaß (=Jaccard-Koeffizient) von 1, 0,750 und 0,667 gebildet werden. Im Dendrogramm ist der Ähnlichkeitskoeffizient in ein Unähnlichkeitsmaß (Heterogenitätsmaß) umgewandelt und auf das Intervall [0; 25] normiert worden. Das Cluster 1 besteht aus den Kunden 2 und 5, das Cluster 2 aus den Kunden 6 und 7 und das Cluster 3 aus den Kunden 1 und 4.

Bei einem Wert des Jaccard-Koeffizienten von 0,583 wird im 4. Schritt der Kunde 3 dem Cluster 3 zugeordnet. Dieses Cluster ist in der Zuordnungsübersicht durch ein ihm zugehöriges Objekt mit der bisher kleinsten Objektnummer, d.h. hier 6, gekennzeichnet. Im 5. Schritt werden die beiden Cluster 1 und 3, die durch die Objekte 2 und 1 gekennzeichnet sind, bei einem Ähnlichkeitskoeffizienten von 0,417 fusioniert. Im vorletzten Schritt wird das größer gewordene Cluster 1 (Label Objekt 1) mit

dem Cluster 2 (Label Objekt 3) bei einer Ähnlichkeit von 0,285 vereinigt, bevor der Kunde 8 im letzten Schritt diesem Cluster zugeordnet wird.

Der Kunde 8, der keinerlei Ähnlichkeit mit den anderen Kunden aufweist (J = 0), kann als Ausreißer eingestuft werden. Für die verbleibenden Objekte geht aus dem Dendrogramm der größte Anstieg der Unähnlichkeit im vorletzten Schritt bei einer Vereinigung des Clusters 1 (Kunden 1, 2, 4 und 5) mit dem Cluster 2 (Kunden 3, 6 und 7) hervor. Die hierarchische Klassifikation deckt damit eine Zwei-Gruppen-Struktur mit einem Ausreißer auf.

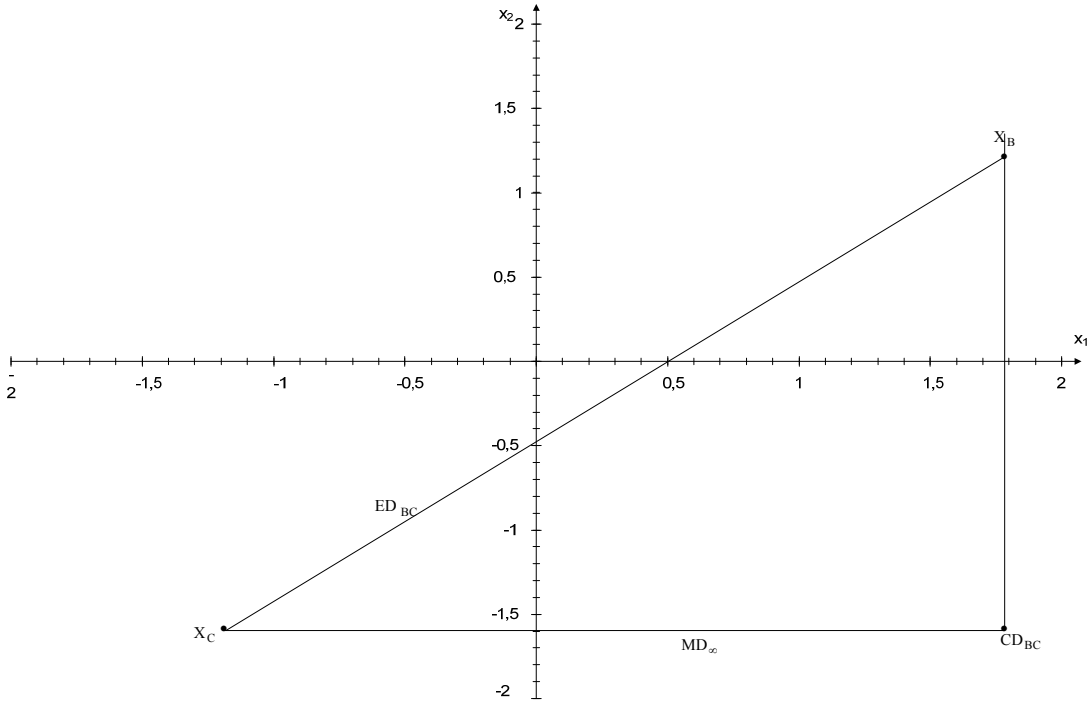
5. Die Objektvektoren der beiden Regionen B und C sind im Hinblick auf die Merkmale Bevölkerungsdichte (X_1) und BIP pro Einwohner (X_2) durch

$\mathbf{x}_B = (1{,}709 \quad 1{,}254)'$ und $\mathbf{x}_C = (-1{,}343 \quad -1{,}653)'$

gegeben. Visualisieren Sie euklidische Distanz, die City-Block-Distanz und die Tschebyscheff-Distanz zwischen den beiden Objekte im zweidimensionalen Merkmalsraum!

<u>Lösung:</u>

Visualisierung von Distanzen im Merkmalsraum



6. Berechnen Sie die drei in Aufg. 5 genannten Distanzmaße für die beiden Regionen B und C!

Lösung:

Euklidische Distanz:

$$\begin{aligned} ED_{BC} &= \sqrt{[1{,}709-(-1{,}343)]^2+[1{,}254-(-1{,}653)]^2} \\ &= \sqrt{3{,}052^2+2{,}907^2} = \sqrt{9{,}3147+8{,}4506} \\ &= \sqrt{17{,}7653} = 4{,}215 \end{aligned}$$

City-Block-Distanz:

$$\begin{aligned} CD_{BC} &= |1{,}709-(-1{,}343)|+|1{,}254-(-1{,}653)| \\ &= |3{,}052|+|2{,}907| = 5{,}959 \end{aligned}$$

Tschebyscheff-Distanz:

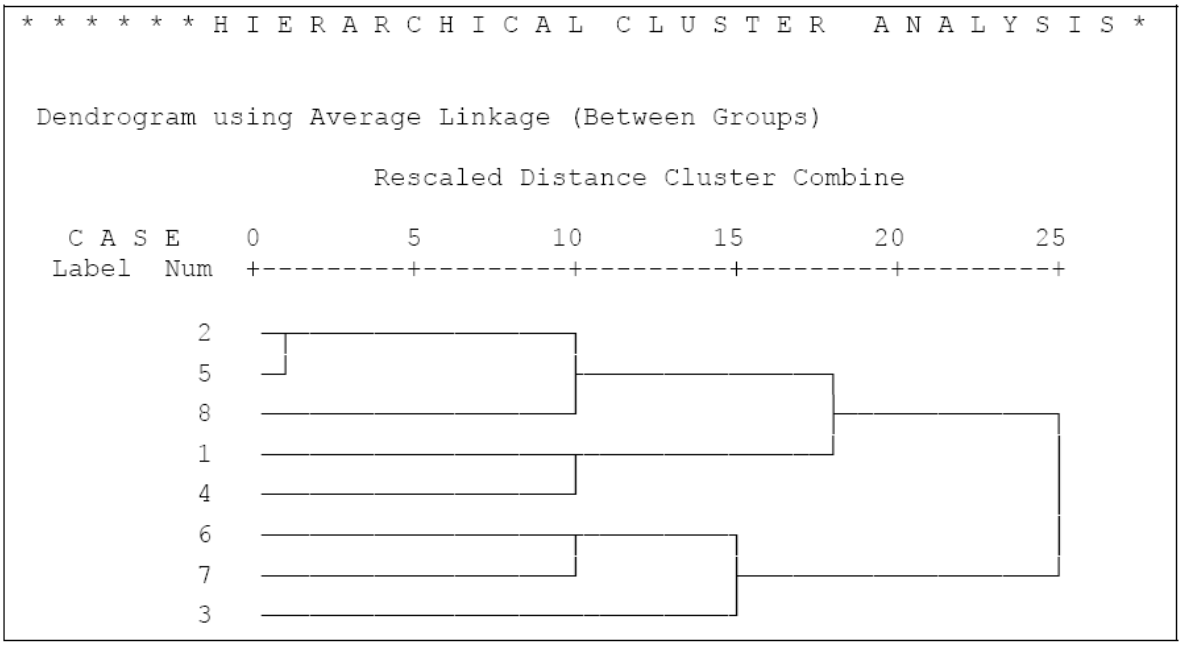
$$\begin{aligned} MD_{\infty} &= \max\{\,|1{,}709-(-1{,}343)|\,,\,|1{,}254-(-1{,}653)|\,\} \\ &= \max\{\,|3{,}052|\,,\,|2{,}907|\,\} \\ &= 3{,}052 \end{aligned}$$

7. Wodurch unterscheiden sich die Partitionsverfahren von den hierarchischen Klassifikationsverfahren?

Lösung.

Während die Clusterzahl bei den Partitionsverfahren vorgegeben wird und unverändert bleibt, verändert sie sich bei den hierarchischen Verfahren von Stufe zu Stufe. Bei den praktisch relevanten agglomerativen hierarchischen Verfahren wird ausgehend von n einelementigen Clustern in n–1 Stufen durch Fusion der beiden jeweils "ähnlichsten" Cluster eine aus einem einzigen Cluster bestehende Partition erstellt. Dagegen erfolgt bei den Partitionsverfahren in jeder Stufe bei unveränderter Clusterzahl eine Umgruppierung von Objekten, wenn die Partition dadurch homogener wird. Wenn keine Verbesserung der Homogenität mehr erreichbar ist, ist eine "optimale" Partition ermittelt worden. Bei Einsatz hierarchischer Verfahren muss die für eine Anwendung am besten geeignete Partition aus der Clusterhierarchie herausgefunden werden. Sprunghafte Erhöhungen der Heterogenität (Unähnlichkeit, Distanz) bei einer Verschmelzung zweier Cluster werden hierbei zur Bestimmung der Gruppenzahl herangezogen.

8. Erläutern Sie die Gruppierung von 8 Konsumenten anhand des vorliegenden Dendrogramms!



Lösung:

Interpretation wie Aufg. 4 ohne Kenntnis der ermittelten Ähnlichkeitskoeffizienten (→ Ergebnis: Zwei-Gruppen-Struktur)

9. Gegeben ist die die Distanzmatrix

$$\mathbf{D} = \begin{bmatrix} 0 & & & & \\ 0{,}8 & 0 & & & \\ 0{,}3 & 0{,}7 & 0 & & \\ 0{,}2 & 0{,}6 & 0{,}9 & 0 & \\ 0{,}1 & 0{,}4 & 0{,}85 & 0{,}75 & 0 \end{bmatrix}$$

von 5 Konsumenten. Ermitteln Sie hierarchische Klassifikationen nach dem

- Single-Linkage-Verfahren,
- Complete-Linkage-Verfahren,
- Average-Linkage-Verfahren!

Lösung:

Angaben über die Vereinigung von Clustern zu den jeweiligen (Cluster-)Distanzen in den vier Stufen des hierarchischen Klassifikationsprozesses:

	Single-Linkage-Verf.	Complete-Linkage-Verf.	Average-Linkage-Verf.
1.Stufe	$D(C_1,C_5) = d(A,E)=0,1$	$D(C_1,C_5) = d(A,E) = 0,1$	$D(C_1,C_5) = d(A,E) = 0,1$
2. Stufe	$D(C_1,C_4) = 0,2$	$D(C_2,C_4) = 0,6$	$D(C_1, C_4) = 0,475$
3. Stufe	$D(C_1,C_3) = 0,3$	$D(C_1, C_2) = 0,8$	$D(C_1,C_2) = 0,6$
4. Stufe	$D(C_1,C_2) = 0,4$	$D(C_1,C_2) = 0,9$	$D(C_1,C_2) = 0,688$

10. Für 4 Käufer liegen die standardisierten Werte des Einkommens und Alters und eine Klassifikation vor:

i (Käufer)	Einkommen	Alter	C_g
1	1,2	0,8	C_1
2	-0,6	-1,3	C_2
3	-0,8	0,2	C_1
4	0,2	0,3	C_2

Ermitteln Sie die Minimal-Distanz-Partition unter Anwendung des K-Means-Verfahrens!

Lösung:

Im 1. Durchlauf (1. Iteration) sind zunächst die quadrierten euklidischen Distanzen zwischen den vier Käufern und den beiden Clusterzentroiden (Clusterschwerpunkten) zu bestimmen. Da die beiden Käufer 3 und 4 von ihrem jeweiligen Clusterzentrum weiter entfernt liegen als vom Schwerpunkt des anderen Clusters, werden sie umgruppiert. Am Ende des ersten Durchlaufs ergibt sich folgende Partition:

Cluster 1			Cluster 2		
Käufer	X_1 (Eink.)	X_2 (Alter)	Käufer	X_1 (Eink.)	X_2 (Alter)
K_1	1,2	0,8	K_2	-0,6	-1,3
K_4	0,2	0,3	K_3	-0,8	0,2
	$\bar{z}_{11} = 0,7$	$\bar{z}_{12} = 0,55$		$\bar{z}_{21} = -0,7$	$\bar{z}_{22} = -0,55$

Bestimmt man nun im 2. Durchlauf (2. Iteration) die Distanzen der vier Käufer zu den neuen Clusterzentroiden, zeigt sich, dass sich keine Distanz eines Käufers zu einem Cluster durch eine weitere Umgruppierung vermindern lässt. Daher ist die sich am Ende des ersten Durchlaufs ergebende Partition eine Minimal-Distanz-Partition.