Probleme bei der Bestimmung der Clusteranzahl bei hierarchischen Clusteranalysen

Seminararbeit

vorgelegt am: 7. September 2017

an der Rheinischen Friedrich-Wilhelms-Universität Bonn

Name:	Abdurahman Maarouf
Adresse:	Schulstrasse 31
PLZ Ort:	53332 Bornheim
Matrikelnummer:	2913095
Betreuer:	Dr. Heiko Wagner, Dominik Poß

Inhaltsverzeichnis

1 Einführung in die hierarchische Clusteranalyse

1.1 Problemstellung

Nachdem man die Daten für Untersuchungsobjekte, wie zum Beispiel Personen, Unternehmen oder Produkte, und deren Merkmalsausprägungen gesammelt hat, trifft man häufig auf folgende Problemstellung. Wie kann man die Untersuchungsobjekte anhand der Ergebnisse nun sinnvoll strukturieren oder sortieren? Eine beliebte Lösung ist es, Gruppen zu erstellen, wobei ähnliche Untersuchungsobjekte in eine Gruppe zusammengefasst werden. Hierbei möchte man möglichst alle Merkmalsvariablen gleichzeitig beachten, welche eine gewisse Bedeutung für die Untersuchung aufweisen.
Bei vielen Anwendungen, wie z.B. in der Medizin, Soziologie, Biologie und in den Wirtschaftswissenschaften, ist diese Situation von Bedeutung. Diese Problemstellung stellt sich auch im Alltag, wobei man dort natürlich auf multivariate Analysemethoden verzichtet. Folgendes Beispiel soll jedoch nur die Problemstellung verdeutlichen. Man möchte beispielsweise Bücher in fünf Regale sortieren und dabei ähnliche Bücher in einem Regal positionieren. Hierbei wären die Bücher die Untersuchungsobjekte und mögliche Merkmalsausprägungen sind Genre, Autor, Erscheinungsjahr, Seitenanzahl usw..
Die Stirling-Zahl zweiter Art [1, S. 3] gibt uns an, wieviele unterschiedliche disjunkte Aufteilungen der N Untersuchungsobjekte in M Gruppen insgesamt möglich sind. Sie ist definiert als:

$$S(N,M) = \frac{1}{M!} \cdot \sum_{i=0}^{M} (-1)^{M-i} \binom{M}{i} i^N \tag{1}$$

Bei auch nur 25 Büchern und 5 Regalen ergeben sich also 2.436.684.974.110.751 Möglichkeiten, diese zu sortieren. Es wird deutlich, dass ohne einer speziellen multivariaten Analysemethode, das Auffinden der besten Aufteilung, wobei die Untersuchungsobjekte in Gruppen so ähnlich wie möglich seien sollen, sehr schwer fällt. Hier kommt die Clusteranalyse ins Spiel.

1.2 Ziel und Funktion einer hierarchischen Clusteranalyse

Bei der beschriebenen Problemstellung bietet die Clusteranalyse eine Lösung. Das Ziel der Clusteranalyse ist es, die Grundgesamtheit der Untersuchungsobjekte vereinfacht und strukturiert darzustellen. Hierbei sollen die wichtigsten Eigenschaften aus der Grundgesamtheit erkennbar gemacht werden. Somit werden die Untersuchungsobjekte in Gruppen (Cluster) zusammengefasst, wobei die Clusteranalyse dadurch ausgezeichnet wird, dass sie alle Merkmalsvariablen der Untersuchungsobjekte gleichzeitig berücksichtigt.
Zunächst muss zwischen der partitionierenden und der hierarchischen Clusteranalyse unterschieden werden [2, S. 476] . Die partitionierende Clusteranalyse geht von einer gegebenen Gruppierung aus und somit ist die Clusteranzahl im Ausgangspunkt bekannt. Dies wäre für das vorherige Beispiel vorteilhaft, da dort die Anzahl an Regalen feststand. Die Untersuchungsobjekte werden den vorgegebenen Gruppen nun so zugeordnet, dass eine bestimmte Zielfunktion ihr Maximum erreicht. Da jedoch das hierarchische Verfahren in der Praxis häufiger zur Anwendung kommt, wird die-

se Arbeit sich auf die hierarchische Clusteranalyse beschränken.
Was die hierarchische Clusteranalyse von anderen Verfahren unterscheidet ist, dass die Gruppen im Ausgangspunkt unbekannt sind und eine einfache Zuordnung der Untersuchungsobjekte in vorgegebene Gruppen nicht möglich ist. Die Gruppen werden erst während des Verfahrens herbeigeführt, indem Untersuchungsobjektgruppen in der Grundgesamtheit mithilfe von Gruppierungsverfahren identifiziert werden. Das bedeutet bei unserem Beispiel, dass die optimale Anzahl an Regalen nicht unbedingt fünf sein muss. Dies lässt die hierarchische Clusteranalyse im Ausgangspunkt offen.
Zudem werden die Untersuchungsobjekte so gruppiert, dass folgende zwei Kriterien möglichst maximal erfüllt werden. Das erste Kriterium fordert Homogenität innerhalb der Gruppen. Also sollen die Objekte, welche zu einer Gruppe zusammengefasst werden, so ähnlich wie möglich sein. Das zweite Kriterium möchte eine maximale Heterogenität zwischen den Gruppen schaffen, d.h. Objekte unterschiedlicher Gruppen sollen möglichst unähnlich sein [3, S. 443].

1.3 Aufbau einer hierarchischen Clusteranalyse

Der Aufbau und die Durchführung einer Clusteranalyse werden im Folgenden anlehnend an den Lehrbüchern „Multivariate Statistik - Lehr- und Handbuch der angewandten Statistik“ [3] „Multivariate Analysemethoden - Eine anwendungsorientierte Einführung“ [2] und „Clusteranalyse - Einführung in Methoden und Verfahren der automatischen Klassifikation“ [4] erklärt.
Die hierarchische Clusteranalyse wird in drei Schritte aufgeteilt: Bestimmung der Ähnlichkeits- bzw. Distanzwerte, Bestimmung des Fusionierungsalgorithmus, Bestimmung der optimalen Clusteranzahl.
Bei der Bestimmung der Ähnlichkeits- bzw. Distanzwerte zwischen den Untersuchungsobjekten werden immer zwei Objekte auf Ähnlichkeiten bzw. Unterschiede bezüglich ihrer Merkmalsausprägungen überprüft. Dies wird mit einem Proximitätsmaß gemessen. Da die Wahl des Proximitätsmaß vom Skalenniveau der Merkmalsvariablen abhängt, wird diese Arbeit einen Fokus auf die Bestimmung von Ähnlichkeits- bzw. Distanzwerten bei binären und metrischen Merkmalsvariablen legen.
Bei der hierarchischen Clusteranalyse muss man zwischen divisiven und agglomerativen Verfahren unterscheiden. Diese Verfahren entscheiden den Ablauf des Fusionsalgorithmus. Bei divisiven Algorithmen geht man im Ausgangspunkt davon aus, dass alle Untersuchungsobjekte in einem gemeinsamen Cluster sind und mithilfe der Ähnlichkeits- bzw. Distanzwerte während des Verfahrens in kleinere Gruppen aufgeteilt werden. Die agglomerativen Algorithmen, denen in der Praxis die höhere Bedeutung zukommt, werden dadurch charakterisiert, dass sie von der kleinsten Partition ausgehen. Das bedeutet, dass im Ausgangspunkt die Anzahl an Untersuchungsobjekten und die Anzahl an Clustern gleich ist, und während des Verfahrens die Objekte anhand ihrer Ähnlichkeits- bzw. Distanzwerte zu größeren Gruppen zusammengefasst werden. Bei beiden Algorithmen müssen die zwei Kriterien der Gruppierung beachtet werden.
Der letzte Schritt ist die Bestimmung der optimalen Clusteranzahl. Dies ist eine Problemstellung für sich und ist somit auch der Schwerpunkt dieser Arbeit, denn die Bestimmung der optimalen Clusteranzahl trifft auf ein Trade-Off. Einerseits soll das

Endergebnis handhabbar und anschaulich sein aber andererseits soll die Homogenität innerhalb der Gruppen maximiert werden. Ersteres möchte die Clusteranzahl klein halten und Zweiteres wird maximal wenn es m=n (bei n Untersuchungsobjekten und m Gruppen) Cluster gibt.

2 Ähnlichkeits- und Distanzfunktionen

2.1 Definition

Wie schon erwähnt ist der erste Schritt einer hierarchischen Clusteranalyse die Bestimmung der Ähnlichkeits- bzw. Distanzwerte zwischen jedem möglichen Objektpaar der Untersuchungsmenge. Hierfür werden Proximitätsmaße verwendet, wobei Ähnlichkeitsmaße große Werte annehmen, wenn die Ähnlichkeit groß ist, und Distanzmaße große Werte annehmen, wenn zwei Objekte unterschiedlich voneinander sind.
Am Anfang einer Clusteranalyse befinden sich die Daten in einer $N \times J$ Rohdatenmatrix. Hier werden N Untersuchungsobjekte (Personen, Unternehmen, Bücher) durch J Merkmalsvariablen (Alter, Unternehmenswert, Erscheinungsjahr) bechrieben.

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ \vdots \\ Objekt_N \end{array} \begin{array}{c} \begin{array}{cccc} Merkmal_1 & Merkmal_2 & \dots & Merkmal_J \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1J} \\ x_{21} & x_{22} & \dots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NJ} \end{pmatrix} \end{array}$$

Rohdatenmatrix

Am Ende des ersten Schritts wird eine Ähnlichkeits- bzw. Distanzmatrix aufgestellt. Diese Matrix ist quadratisch ($N \times N$), da sie alle Ähnlichkeits- bzw. Distanzwerte der einzelnen Untersuchungsobjekte zu den anderen Untersuchungsobjekten zusammenfasst.

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ \vdots \\ Objekt_N \end{array} \begin{array}{c} \begin{array}{cccc} Objekt_1 & Objekt_2 & \dots & Objekt_N \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1N} \\ S_{21} & S_{22} & \dots & S_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ S_{N1} & S_{N2} & \dots & S_{NN} \end{pmatrix} \end{array}$$

Ähnlichkeits- bzw. Distanzmatrix

Dies ist eine Ähnlichkeitsmatrix, da im Inneren dieser Matrix die Werte mit einem Ähnlichkeitsmaß S_{nj} (Similarity zwischen Objekt n und j) bestimmt wurden.

2.2 Ähnlichkeitsfunktionen bei binären Merkmalen

Binäre Merkmale liegen vor, wenn sich die Merkmalsausprägung auf „Merkmal vorhanden"(x=1) und „Merkmal nicht vorhanden"(x=0) beschränkt. Im Folgenden wird ein fiktives Beispiel mit fünf Patienten und 4 Krankheiten betrachtet. Dementsprechend sieht die Rohdatenmatrix beispielsweise folgendermaßen aus:

$$\begin{array}{c} \\ Patient_1 \\ Patient_2 \\ Patient_3 \\ Patient_4 \\ Patient_5 \end{array} \begin{array}{c} \begin{array}{cccc} Krankheit_1 & Krankheit_2 & Krankheit_3 & Krankheit_4 \end{array} \\ \left(\begin{array}{cccc} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \end{array}$$

Rohdatenmatrix: Beispiel 1

Um nun zwei Untersuchungsobjekte (Patienten) vergleichen zu können, muss man zwischen vier Fällen unterscheiden. Entweder beide Patienten besitzen die betrachtete Krankheit (a), oder nur der erste Patient besitzt sie (b), oder nur der zweite Patient sie (c), oder keiner der beiden Patienten leidet an dieser Krankheit (d). Die Buchstaben a,b,c und d messen die Anzahl des Eintretens dieser vier Fälle für alle Krankheiten. Vergleicht man beispielsweise Patient 1 mit Patient 2, dann würde gelten: a=0; b=1; c=3; d=0. Somit lässt sich eine einfache allgemeine Ähnlichkeitsfunktion herleiten:

$$S_{ij} = \frac{a + \delta d}{a + \delta d + \gamma(b + c)} \tag{2}$$

mit S_{ij} als Ähnlichkeit zwischen Patient i und j und δ und γ als mögliche Gewichtungsfaktoren. S_{ij} bewegt sich zwischen 0 und 1, wobei 0 die größtmögliche Unähnlichkeit und 1 die größtmögliche Ähnlichkeit ausdrückt. Distanzwerte können also durch $D_{ij} = 1 - S_{ij}$ berechnet werden. Die Gewichtungsfaktoren hängen von dem letztendlichen Ähnlichkeits- koeffizienten ab, von denen die drei wichtigsten im Folgenden vorgestellt und verglichen werden. Was bei jedem Ähnlichkeitskoeffizienten jedoch geich ist, ist $S_{12} = 0$, da bei Patienten 1 und 2 a=d=0 gilt, und es somit keine Ähnlichkeit zwischen diesen beiden Patienten gibt.

Der Jaccard-Koeffizient setzt $\delta = 0$ und $\gamma = 1$. Die Ähnlichkeitsfunktion vereinfacht sich auf:

$$S_{ij} = \frac{a}{a + b + c} \tag{3}$$

Daraus folgt, dass der Jaccard-Koeffizient den relativen Anteil der gemeinsamen Krankheiten bezüglich der Krankheiten, die bei mindestens einem der zwei Patienten auftreten, angibt. Vergleicht man Patient 1 mit Patient 3, dann ergibt sich $S_{13} = \frac{1}{1+0+1} = 0,5$ als Ähnlichkeitswert. Analog ist die Vorgehensweise für alle anderen Ähnlichkeitswerte. Mit allen Ähnlichkeitswerten erstellt man die 5×5-Ähnlichkeitsmatrix um den ersten Schritt der Clusteranalyse abzuschließen.

$$\begin{array}{c} \\ Patient_1 \\ Patient_2 \\ Patient_3 \\ Patient_4 \\ Patient_5 \end{array} \begin{array}{c} \begin{array}{ccccc} Patient_1 & Patient_2 & Patient_3 & Patient_4 & Patient_5 \end{array} \\ \left(\begin{array}{ccccc} 1 & & & & \\ 0 & 1 & & & \\ 0,5 & 0,25 & 1 & & \\ 0 & 0,33 & 0,5 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}$$

Ähnlichkeitsmatrix: Beispiel 1, Jaccard-Koeffizient

Bei dem Russel und Rao-Koeffizienten (RR-Koeffizienten) werden im Gegensatz zum Jaccard-Koeffizienten die Fälle, bei denen beide Patienten die Krankheit nicht aufweisen (d), mit in den Nenner aufgenommen. Die Ähnlichkeitsfunktion sieht also so aus:

$$S_{ij} = \frac{a}{a+b+c+d} = \frac{a}{M} \tag{4}$$

Der M-Koeffizient wird auch „Simple-Matching-Koeffizient“genannt, da er im Zähler alle übereinstimmenden Komponenten und im Nenner alle Fälle erfasst:

$$S_{ij} = \frac{a+d}{a+b+c+d} = \frac{a+d}{M} \tag{5}$$

Auffällig ist, dass wenn d=0 gilt, dass die drei Koeffizienten den gleichen Wert annehmen. Wenn jedoch d größer als 0 ist, dann wird der RR-Koeffizient am größten und der M-Koeffizient am kleinsten. Dementsprechend spielt d eine wichtige Rolle bei der Auswahl des Koeffizienten. Was hier zu beachten ist, ist die Bedeutung von d für die Problemstellung. Ist die Aussagekraft von „Merkmal nicht vorhanden“genau so groß wie die Aussagekraft von „Merkmal vorhanden“, dann ist die Wahl eines Koeffizienten, welcher d im Zähler und im Nenner genau so stark berücksichtigt wie a, b und c (M-Koeffizient), sinnvoll. Bei dem Beispiel „$Krankheit_j$ nicht vorhanden“und „$Krankheit_j$ vorhanden“, ist dies der Fall. Wenn jedoch die Aussagekraft nicht gleich ist, wie z.B. bei „Nicht-Deutsch“und „Deutsch“, dann liefert der Jaccard-Koeffizient genauere Ergebnisse für die Problemstellung.
Somit ist es bei Transformationen von einer nominalen Variable in eine binäre Variable wichtig, die Bedeutung der neuen Zerlegung gründlich zu überdenken. Denn wenn beispielsweise zwei „Nicht-Deutsche“Untersuchungsobjekte einen hohen Ähnlichkeitswert aufweisen, kann dies die Ergebnisse der gesamten Clusteranalyse verzerren. Also müsste man bei vielen möglichen Merkmalsausprägungen auf eine Transformation in eine binäre Variable verzichten und stattdessen eine andere Ähnlichkeitsbestimmung durchführen.

2.3 Ähnlichkeits-/ Distanzfunktionen bei metrischen Merkmalen

Die Herangehensweise an die Bestimmung von Ähnlichkeits- bzw. Distanzwerten bei metrischen Merkmalsvariablen ist anders als bei binären Merkmalsvariablen. Zwar wird am Anfang ebenfalls eine Rohdatenmatrix genutzt und am Ende eine Ähnlichkeits- bzw. Distanzmatrix aufgestellt, jedoch sind die Methoden aufgrund der Variablenstruktur nun anders.
Als Beispiel für die metrischen Merkmalsvariablen wird ein Beispieldatensatz genutzt, welcher für 8 unterschiedliche Berufe das Einkommen und den Wert des Markenbewusstseins (Marke) wiedergibt. Die Rohdatenmatrix sieht folgendermaßen aus:

	Einkommen	*Marke*
Arzt	6861	21765
Ingenieur	5150	28245
Professor	5152	24608
CEO	12810	27611
Anwalt	7203	21536
Koch	4162	24823
Lehrer	4311	14735
Reinigungspersonal	2132	8822

Rohdatenmatrix: Beispiel 2

Da hier nicht nur zwei Merkmalsausprägungen möglich sind, werden andere Methoden zur Bestimmung von Ähnlichkeits- bzw. Distanzwerten benötigt. Hierfür gibt es Methoden zur Bestimmung von Distanzwerten (Minkowski-Metrik, Einfache und quadrierte Euklidische Distanz) und Methoden zur Bestimmung von Ähnlichkeitswerten (Q-Korrelationskoeffizient). Wann welche Methode sinnvoller erscheint, wird nach der Vorstellung der unterschiedlichen Maße erläutert.
In der Praxis wird das Distanzmaß der Minkowski-Metrik häufig verwendet. Sie lässt sich wie folgt berechnen:

$$D_{g,h} = [\sum_{j=1}^{J} |x_{gj} - x_{hj}|^r]^{\frac{1}{r}} \tag{6}$$

$D_{g,h}$ ist die Distanz zwischen den Objekten g und h, wobei x_{gj} und x_{hj} die j-te Merkmalsausprägung (mit j=1,2,...,J) der Objekte g und h sind. $r \geq 1$ wird als Minkowski-Konstante definiert: Für r=1 heißt die Minkowski-Metrik auch „City-Block-Metrik“ oder „L1-Norm“ und gibt die Summe der absoluten Distanzwerte wieder. Falls r=2 ist, wird von einer Euklidischen Distanz (L2-Norm) gesprochen. Möchte man große Differenzwerte bei der Bestimmung der Distanzwerte stärker berücksichtigen nutzt man die quadrierte Euklidische Distanz ($D^2_{g,h}$), welche die einfache Euklidische Distanz wie der Name schon sagt quadriert. Die Berechnung der quadrierten Euklidischen Distanz zwischen einem Arzt und einem Ingenieur wäre also beispielsweise:

$$D^2_{Arzt,Ingenieur} = (6861 - 5150)^2 + (21765 - 28245)^2 = 44917921$$

Bei Minkowski-Metriken ist es wichtig, dass die zugrundeliegenden Maßeinheiten miteinander vergleichbar sind. Bei dem Beispiel mit Einkommen und Markenbewusstsein würde dies beispielsweise heißen, dass ein Unterschied zwischen einem Objekt mit einem Einkommen von 5000 und einem Objekt mit einem Einkommen von 6000 gleich zu interpretieren ist wie ein Einkommensunterschied von 16000-15000. Ist dies nicht der Fall, muss zuerst eine Standardisierung durchgeführt werden.
Zur Bestimmung von Ähnlichkeitswerten wird der Q-Korrelationskoeffizient genutzt. Er ist auch als Pearson-Korrelationskoeffizient bekannt.

$$r_{g,h} = \frac{\sum_{j=1}^{J}(x_{gj} - \overline{x}_g) \cdot (x_{hj} - \overline{x}_h)}{[\sum_{j=1}^{J}(x_{gj} - \overline{x}_g)^2 \cdot \sum_{j=1}^{J}(x_{hj} - \overline{x}_h)^2]^{\frac{1}{2}}} \tag{7}$$

$r_{g,h}$ misst die Ähnlichkeit zwischen den Objekten g und h. x_{gj} und x_{hj} sind wieder die j-te Merkmalsausprägung (mit j=1,2,...,J) der Objekte g und h und $\overline{x}_g$ und $\overline{x}_h$ sind die Durchschnittswerte aller Eigenschaften bei Objekt g und h.
Bei Beispiel 2 würde dieses Ähnlichkeitsmaß immer 1 sein, da es nur zwei Merkmale gibt und das Markenbewusstsein bei jedem Job höher ist als das Einkommen. Das bedeutet, der Koeffizient misst eine maximale Ähnlichkeit zwischen jedem Objektpaar. Da diese Ergebnisse für die Clusteranalyse nutzlos sind, muss auf den Q-Korrelationskoeffizienten bei zwei Merkmalsvariablen verzichtet werden.
Bei der Verwendung des Q-Korrelationskoeffizienten können allgemein ganz andere Ergebnisse als bei den Minkowski-Metriken auftreten. Es kann also sein, dass der Q-Korrelationskoeffizient eine hohe Ähnlichkeit zwischen zwei Objekten misst, bei denen die Euklidische Distanz eine hohe Distanz misst. Hier ist die Frage, wie man zwei Objekte als ähnlich bzw. unähnlich definieren möchte.
Allgemein gilt, dass ein Distanzmaß dann zu bevorzugen ist, wenn der absolute Abstand zwischen Objekten von Interesse ist. Ähnlichkeitsmaße basieren auf Korrelationswerten und sind somit dann geeignet, wenn die Distanz einzelner Merkmalsausprägungen unwichtiger sind als die Korrelation zweier Merkmale bei zwei Objekten. Für das Beispiel 2 hieße es, dass eine Verwendung von Distanzmaßen sinnvoller ist.
Mithilfe von Statistik-Programmen wie SPSS ist eine Bestimmung von Ähnlichkeits- bzw. Distanzmatrizen und auch eine gesamte Durchführung einer Clusteranalyse möglich. Es folgt die Distanzmatrix zu dem Beispiel 2, welche mit der Quadrierten Euklidischen Distanz bestimmt wurde.

	Arzt	*Ingenieur*	*Professor*	*CEO*	*Anwalt*	*Koch*	*Lehrer*	*Reinigungspersonal*
Arzt	,000							
Ingenieur	44917921	,000						
Professor	11003330	13227773	,000					
CEO	69566317	59077556	67662973	,000				
Anwalt	169405	49225490	13643785	68344074	,000			
Koch	16635965	12686228	1026325	82560848	20052050	,000		
Lehrer	55923400	183224021	98183410	238024377	54617265	101789945	,000	
Reinigungspersonal	189884690	386361253	258318196	467046205	187360837	260152901	39711610	,000

Distanzmatrix: Beispiel 2

3 Clusteranalysealgorithmen

3.1 Auswahl des Fusionierungsalgorithmus

Die gewonnene Ähnlichkeits- bzw. Distanzmatrix bildet den Ausgangspunkt der Clusteranalysealgorithmen. Diese Algorithmen möchten letztendlich mithilfe der Ähnlichkeits- bzw. Distanzwerte die Gruppierung durchführen. Wie schon erwähnt gibt es unterschiedliche Clusteralgorithmen. Diese Arbeit wird sich auf die agglomerativen hierarchischen Verfahren beschränken, da sie in der Praxis häufig zur Anwendung kommen. Das bedeutet, dass man mit der feinsten Partition beginnt (jedes Objekt stellt eine Gruppe dar) und während des Verfahrens Objekte oder Gruppen zu größeren Gruppen zusammenfasst, bis alle Untersuchungsobjekte in einer gemeinsamen Gruppe (Ein-Cluster-Lösung) sind.

3.2 Agglomerative Hierarchische Verfahren

Das agglomerative Verfahren lässt sich in 6 Ablaufschritte unterteilen. Für Beispiel 2 wurden die ersten zwei Schritte bereits erledigt und somit wird das Beispiel für das weitere Verfahren genutzt.
Die ersten zwei Schritte beinhalten einmal die Aufteilung der Untersuchungsobjekte in die feinste Partition und als zweites die paarweise Bestimmung der Ähnlichkeits- bzw. Distanzwerte und der Aufstellung der Ähnlichkeits- bzw. Distanzmatrix.
Im 3. Schritt beginnt die Gruppierung. Hierfür werden die zwei Objekte/Gruppen gesucht, welche die höchste Ähnlichkeit bzw. niedrigste Distanz aufweisen. In Beispiel 2 wären das die Jobs Arzt und Anwalt mit einer Distanz von nur 169405. Im 4. Schritt werden diese zwei Objekte/Gruppen zu einer neuen Gruppe zusammengefasst.
Da nun eine neue Gruppe entstanden ist, muss die „alte" Ähnlichkeits- bzw. Distanzmatrix aktualisiert werden. Es werden also die Ähnlichkeits- bzw. Distanzwerte zwischen der neu entstandenen Gruppe und den übrigen Objekten/Gruppen berechnet und man gelangt zu einer „reduzierten" Ähnlichkeits- bzw. Distanzmatrix. Objekte/Gruppen, die zusammengefasst wurden, werden aus der Matrix entnommen und die neue Gruppe wird eingesetzt. Die Berechnung der neuen Ähnlichkeits- bzw. Distanzwerte wird noch erläutert.
Die Gruppenanzahl nimmt nach einem Durchlauf der Schritte 3-5 immer um 1 ab. Es werden nur 2 Objekte/Gruppen pro Durchlauf zusammengefasst. Der 6. Schritt ist die Wiederholung dieses Durchlaufs, bis die Ein-Cluster-Lösung erreicht wurde.

Für die Bestimmung der neuen Distanzwerte gibt es unterschiedliche Methoden. Die einfachsten Methoden nehmen die kleinere (Single-Linkage), größere (Complete-Linkage) oder mittlere (Average-Linkage) Distanz der beiden zusammengefassten Objekte/Gruppen zu den übrigen Objekten/Gruppen als neuen Distanzwert. Nachteile dieser Methoden sind zum Einen, dass die Single-Linkage-Methode dazu neigt, wenige große und viele kleine Gruppen zu bilden, und zum Anderen, dass die Complete-Linkage-Methode etwa gleich große, aber viele Gruppen herbeiführt. Die Average-Linkage-Methode ist zwar ein Mittelmaß dieser zwei Extrema, jedoch führt es meist nicht zu den gewünschten Ergebnissen.
Für metrische Skalenniveaus stellt das Ward-Verfahren eine in der Praxis häufig verwendete und die leistungsstärkste [5, S. 96 f.] Methode dar. Diese Methode basiert nicht nur auf einer anderen Bestimmung von neuen Distanzwerten, sondern auch auf einen anderen Fusionierungsvorgang. Zunächst werden die neuen Distanzwerte folgendermaßen bestimmt:

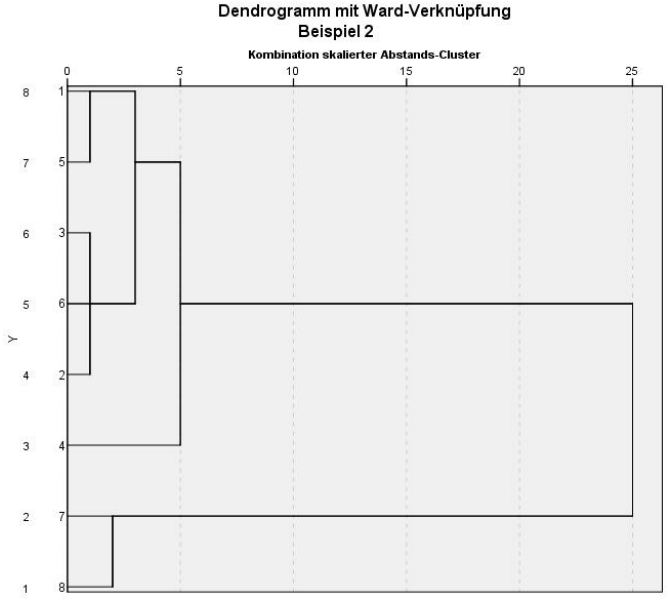
$$D(X,Y+Z) = \frac{1}{NX+NY+NZ} \cdot [(NX+NY)D(X,Y)+(NX+NZ)D(X,Z)-NX \cdot D(Y,Z)] \tag{8}$$

Es wird also der neue Distanzwert zwischen der neuen Gruppe, bestehend aus den zusammengefassten Objekten/Gruppen Y und Z, und einer übrigen Gruppe X bestimmt. Hierbei sind NX, NY und NZ die Anzahl an Objekten in den jeweiligen Gruppen. Somit bekommen die ursprünglichen Distanzwerte eine Gewichtung, die davon abhängt, wieviele Objekte bereits in der Gruppe vorhanden sind.
Anders als bei den Linkage-Methoden, vereint die Ward-Methode diejenigen Objekte/Gruppen, welche eine bestimmte Fehlerquadratsumme am wenigsten vergrößern.

Dadurch werden die Gruppen im Endergebnis möglichst homogen. Die Fehlerquadratsumme für eine bestimmte Gruppe i errechnet sich wie folgt:

$$V_i = \sum_{k=1}^{K} \sum_{j=1}^{J} (x_{kji} - \overline{x}_{ji})^2 \tag{9}$$

wobei x_{kji} der Wert der Merkmalsvariable j bei Objekt k in der Gruppe i ist, und $\overline{x}_{ji}$ der Mittelwert der Merkmalsvariable j von allen k Objekten in der Gruppe i ist. Es lässt sich zeigen, dass bei einer Benutzung der quadrierten Euklidischen Distanz, die errechneten Distanzwerte genau der doppelten Zunahme der Fehlerquadratsumme bei Fusionierung zweier Objekte entspricht.
Die Fehlerquadratsumme ist im Ausgangspunkt in jeder Gruppe 0, da jedes Objekt eine Gruppe darstellt. Für das Beispiel 2 bedeutet das, dass im ersten Schritt die Jobs Arzt und Anwalt vereinigt werden, da sie die geringste Distanz in Höhe von 169405 aufweisen und somit die Fehlerquadratsumme am wenigsten erhöhen. Die Zunahme der Fehlerquadratsumme beträgt also $\frac{169405}{2}$ und liegt bei $0 + \frac{169405}{2}$.
Nach der Vereinigung müssen die neuen Distanzwerte mithilfe von (8) bestimmt werden, welche wieder die theoretische doppelte Zunahme der Fehlerquadratsumme bei Fusionierung der betrachteten Objekte/Gruppen darstellen. Im nächsten Schritt werden wieder die zwei Objekte/Gruppen vereinigt, welche die Fehlerquadratsumme am wenigsten erhöhen und so weiter. Die Fehlerquadratsumme nimmt bis zur Ein-Cluster-Lösung immer weiter zu.
Graphisch kann man diesen Vorgang mit einem Dendrogramm verdeutlichen:



wobei die Beschriftung 1,2,...,8 für die Jobs Arzt, Ingenieur, ..., Reinigungspersonal steht, und das Dendrogramm folgendermaßen schrittweise zu lesen ist:

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Fehlerquadratsumme	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	Arzt	Anwalt	84702,500	0	0	5
2	Professor	Koch	597865,000	0	0	3
3	Ingenieur	Professor+Koch	9064811,167	0	2	5
4	Lehrer	Reinigungspersonal	28920616,17	0	0	7
5	Arzt+Anwalt	Ingenieur+Professor+Koch	56373459,40	1	3	6
6	Arzt+Anwalt+...+Koch	CEO	108155811,7	5	0	7
7	Arzt+Anwalt+...+CEO	Lehrer+Reinigungspersonal	381299768,8	6	4	0

4 Analyse und Interpretation

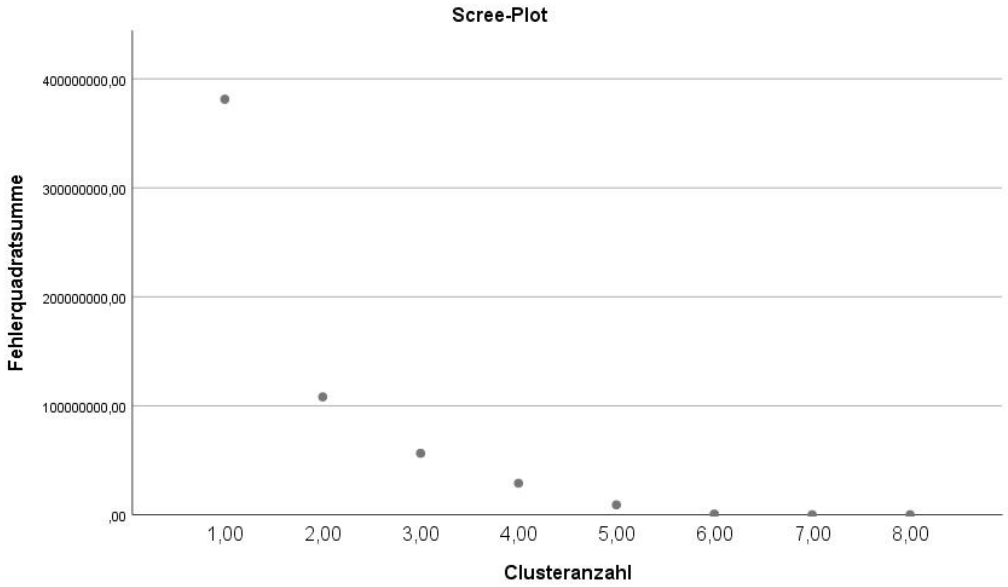
4.1 Spezielle Probleme

Bei der Clusteranalyse stößt man auf bestimmte Probleme der praktischen Durchführung. Wie bei anderen multivariaten Analysemethoden erschwert eine große Anzahl an Untersuchungsobjekten die Durchführung, da sehr viele Ähnlichkeits- bzw. Distanzwerte bestimmt werden müssen und man auf eine dementsprechend große Ähnlichkeits- bzw. Distanzmatrix kommt. Die Aufstellung der reduzierten Ähnlichkeits- bzw. Distanzmatrizen nach jedem Durchlauf des Fusionierungsverfahrens fällt also auch schwer. Somit ist eine Benutzung von Statistik Programmen wie SPSS besonders bei vielen Untersuchungsobjekten sinnvoll. Es empfiehlt sich in diesem Fall auch auf partitionierende Verfahren zurückzugreifen.
Eine hohe Anzahl an Merkmalsvariablen erschwert nicht nur die Durchführung einer Clusteranalyse, sondern auch die Interpretation der Ergebnisse. Zwar ermöglicht die Clusteranalyse eine Gruppierung anhand aller Merkmalsvariablen, jedoch ist es meist nicht nützlich viele Merkmalsvariablen für eine Untersuchung zu beachten. Die erschwerte Interpretation tritt beispielsweise dann auf, wenn den Gruppen im Endergebnis Namen oder Überschriften zugeordnet werden müssen.
Ein weiteres Problem der Clusteranalyse ist die Wahl der optimalen Clusteranzahl, also der Anzahl an Gruppen. Wie schon angesprochen gibt es hier ein Konflikt zwischen der Homogenitätsanforderung, welche die Anzahl der Gruppen maximiert, und der Handhabbarkeit der Ergebnisse, welche die Anzahl der Gruppen möglichst klein halten möchte. Dieses Problem und mögliche Lösungskonzepte werden mit dem Elbow-Kriterium [2, S. 495 f.], der Silhouetten-Statistik [6] und der Slope-Statistik [7] im Folgenden vorgestellt.

4.2 Bestimmung der Clusteranzahl

Ein wichtiges und bis heute nicht vollständig gelöstes Problem der Clusteranalyse ist die Bestimmung der optimalen Clusteranzahl k*. Es soll gesagt sein, dass die Bestimmung sich nicht zu sehr auf statistische Methoden beruhen muss, da für die Problemstellung meist eine bestimmte Anzahl ohne Rechnungen sinnvoll erscheint. Es gibt jedoch Ansätze und Lösungskonzepte für die statistische Bestimmung der optimalen Clusteranzahl k*, die im Folgenden vorgestellt werden.

Die erste Methode beruht auf eine optische Bestimmung und wird Elbow-Kriterium [2, S. 495 f.] genannt. Man sucht einen Sprung/Elbow in der Entwicklung der Fehlerquadratsumme für die jeweilige Gruppenanzahl. Dies lässt sich beispielsweise in der Zuordnungsübersicht für Beispiel 2 bestimmen. Einfacher wird es, den Sprung oder Elbow zu identifizieren, wenn die Fehlerquadratsumme gegen die Gruppenanzahl in einem sogenannten Scree-Plot aufgetragen wird:



Hier ist zu sehen, dass der Sprung/Elbow bei der Anzahl von etwa 3 Gruppen liegt. Somit würde das Elbow-Kriterium auf eine optimale Anzahl von k*=3 Gruppen kommen. Da diese Methode sich an einer optischen Bestimmung der optimalen Anzahl orientiert sollte sie nicht benutzt werden, wenn eine exakte Lösung gewünscht ist.

Im Jahr 1986 schlug Peter J. Rousseeuw [6] eine Methode vor, welche überprüft, wie gut jedes Objekt in einer bestimmten Gruppierung gruppiert wurde und damit das optimale k* erkundet. Diese Methode konstruiert sogenannte Silhouetten im Datensatz und wird somit auch Silhouetten-Methode genannt. Für die Konstruktion dieser Silhouetten werden die Distanzmaße zwischen allen Objekten und eine zu überprüfende Gruppierung benötigt.
Für das Beispiel 2 wurden die Distanzmaße in Form von quadrierten Euklidischen Distanzen bereits berechnet. Mit der Silhouette-Methode kann also überprüft werden, ob der Gruppierungsvorschlag des Elbow-Kriteriums geeignet ist oder nicht. Hierfür wird getestet, ob ein Objekt i in die passende Gruppe sortiert wurde. Dem Elbow-Kriterium nach sollten die Untersuchungsobjekte in 3 Cluster sortiert werden: Cluster A (Arzt, Anwalt, Ingenieur, Professor, Koch); Cluster B (Lehrer, Reinigungspersonal); Cluster C (CEO).
Als erstes wird für das Objekt i die durchschnittliche Distanz zu den übrigen Objekten der zugehörigen Gruppe X bestimmt. Sie ist also definiert als:

$$a(i) = \frac{1}{n} \sum_{j \neq i}^{n} D^2(Xi, Xj) \tag{10}$$

mit n als Anzahl der Untersuchungsobjekte in Gruppe X und $D^2(Xi, Xj)$ als quadrierte Euklidische Distanz des Objekts i in Cluster X und des Objekts j im selben

Cluster X.
Im nächsten Schritt wird für das Objekt i der Gruppe X die durchschnittliche Distanz zu den Objekten einer anderen Gruppe Y errechnet:

$$d(i, Y) = \frac{1}{m} \sum_{k=1}^{m} D^2(Xi, Yk) \tag{11}$$

mit m als Anzahl der Untersuchungsobjekte in Gruppe Y und $D^2(Xi, Yk)$ als quadrierte Euklidische Distanz des Objekts i in Cluster X und des Objekts k im Cluster Y. Dieser Schritt wird auch für Objekt i und allen übrigen Gruppen Z durchgeführt. Für die nächsten Schritte ist die geringste durchschnittliche Distanz von Objekt i zu den Objekten einer anderen Gruppe von Bedeutung, da überprüft wird, ob Objekt i auch in das nächstgelegene Cluster gruppiert werden könnte. Nimmt man an, dass die Objekte der Gruppe Y diese Eigenschaft erfüllen, dann gilt:

$$b(i) = \min_{Z \neq X} d(i, Z) = d(i, Y) \tag{12}$$

Mit den Werten a(i) und b(i) kann folgende Formel aufgestellt werden, welche die Qualität der Gruppierung von Objekt i in Cluster X misst:

$$s(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]} \tag{13}$$

Klar ist, dass sich s(i) in einem Bereich zwischen -1 und 1 bewegt. Bei einem Wert von s(i) nahe 1 gilt, dass die durchschnittliche Distanz des Objekts i zu den Objekten des nächstgelegenen Clusters Y b(i) relativ groß ist, und die Distanzen innerhalb der Gruppe von i relativ klein sind. Somit würde die gegebene Gruppierung von i sinnvoll erscheinen. Bei einem s(i) nahe -1 wäre das Objekt i also nicht gut gruppiert, da es höhere Ähnlichkeiten zu den Objekten aus Gruppe Y aufweist. Ist s(i) ungefähr gleich 0, dann ist es nicht klar, zu welcher Gruppe das Objekt besser zugeordnet wäre.
Um nun die optimale Clusteranzahl bestimmen zu können sucht man die Clusteranzahl, welche das höchste durchschnittliche s(i) der Untersuchungsobjekte aufweist. Mit anderen Worten wird die Clusteranzahl gesucht, wo die Untersuchungsobjekte durchschnittlich am besten gruppiert sind. Hierfür wird die Silhouette-Statistik genutzt, welche für jede Anzahl an Clustern k=2,3,...,n das durchschnittliche s(i) wiedergibt.

$$s(k) = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{14}$$

Im letzten Schritt wird also das k gesucht, was (14) maximiert.

Im Jahre 2014 haben die Forscher A. Fujita, D. Takahashi und A. Patriota [7] erkannt, dass die Silhouette-Methode für homogene Gruppen hervorragend geeignet ist. Mit homogenen Gruppen ist hier gemeint, dass die Gruppen eine ähnliche innere Variabilität zeigen und eine etwa gleich große Anzahl an Objekten beeinhalten. Doch erkannten sie auch, dass die Silhouette-Methode bei nicht-homogenen Gruppen unzuverlässig wird und große Gruppen aufteilt, die nicht aufgeteilt werden sollten.

Man nehme beispielsweise an, dass eine Gruppe eine starke innere Variabilität aufzeigt und mehr Objekte in dieser Gruppe sind als in anderen, es jedoch auch keinen Sinn machen würde, diese Gruppe aufzuteilen. s(k) sollte den drei Forschern nach trotzdem hohe Werte annehmen können, auch wenn es nur wenige Gruppen gibt. Somit entwickelten sie eine Methode, welche auch für den Fall von nicht-homogenen Gruppen auf zuverlässige Ergebnisse kommt.
Hierfür wird folgender Schätzer für die optimale Anzahl an Gruppen k* aufgestellt:

$$\hat{k} = arg \max_{k \in \{2,\dots,n-1\}} -[s(k+1) - s(k)]s(k)^p \tag{15}$$

Der Schätzer beachtet also jetzt auch die Differenz zwischen s(k*) und s(k*+1). Hintergrund dieser Überlegung ist, dass s(k*) der Silhouetten-Methode nach den höchsten Wert annimmt und der Wert für k*+1,2,... und für k*-1,2,... bei homogenen Gruppen signifikant abnimmt. Wenn es jedoch eine Gruppe gibt, welche viel mehr Objekte beinhaltet als andere, dann ist der Wert von s(k) stark abhängig von dieser Gruppe. Aus Abweichungen von k folgen keine so signifikanten Abnahmen wie bei homogenen Gruppen. Wenn diese Gruppe mit vielen Objekten jedoch fälschlicherweise aufgeteilt wird, dann fällt auch die Differenz zwischen s(k) und s(k+1) stärker aus und der gesamte Term für den Schätzer wird größer. Deswegen heißt diese Methode auch Slope-Statistik.

Literatur

[1] M. R. Anderberg, *Cluster Analysis for Applications - Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Amsterdam, Boston: Academic Press, 2014.

[2] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, *Multivariate Analysemethoden - Eine anwendungsorientierte Einführung*. Berlin Heidelberg New York: Springer-Verlag, 2015.

[3] J. Hartung and B. Elpelt, *Multivariate Statistik - Lehr- und Handbuch der angewandten Statistik*. München: Oldenbourg Verlag, 2007.

[4] D. Steinhausen and K. Langer, *Clusteranalyse - Einführung in Methoden und Verfahren der automatischen Klassifikation*. Berlin: Walter de Gruyter, 1977.

[5] S. Bergs, *Optimalität bei Clusteranalysen - Experimente zur Bewertung numerischer Klassifikationsverfahren*. Linz: na, 1981.

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, pp. 53 – 65, 1987.

[7] A. Fujita, D. Y. Takahashi, and A. G. Patriota, "A non-parametric method to estimate the number of clusters," *Computational Statistics & Data Analysis*, vol. 73, pp. 27 – 39, 2014.