de Gruyter Lehrbuch
Steinhausen/Langer · Clusteranalyse

Detlef Steinhausen · Klaus Langer

Clusteranalyse

Einführung in Methoden und Verfahren der automatischen Klassifikation

Mit zahlreichen Algorithmen, FORTRAN-Programmen, Anwendungsbeispielen und einer Kurzdarstellung der multivariaten statistischen Verfahren

Walter de Gruyter · Berlin · New York 1977

Dr. rer. nat., Dipl. math. *Detlef Steinhausen*,
Akademischer Oberrat am Rechenzentrum der Westfälischen Wilhelms-Universität Münster

Klaus Langer,
Dipl. psych., Dipl. päd. am Psychologischen Institut, Abt. Klinische Psychologie, der Westfälischen Wilhelms-Universität Münster

Mit 63 Abbildungen und 4 Tabellen

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Steinhausen, Detlef

Clusteranalyse: Einf. in Methoden u. Verfahren d. automatischen Klassifikation; mit zahlr. Algorithmen, FORTRAN-Programmen, Anwendungsbeispielen u. e. Kurzdarst. d. multivariaten statist. Verfahren / Detlef Steinhausen; Klaus Langer. – 1. Aufl. – Berlin, New York: de Gruyter, 1977.

(de Gruyter Lehrbuch)
ISBN 3-11-007054-5

NE: Langer, Klaus

Satz: IBM-Composer, Walter de Gruyter, Berlin.
Druck: Color-Druck, Berlin.
Bindearbeiten: Dieter Mikolai, Berlin.

Vorwort

Clusteranalysen werden mit Erfolg in nahezu allen Bereichen und Disziplinen angewandt, in denen größere Datenmengen auf wenige und überschaubare Interpretationseinheiten zu reduzieren sind. Bedingt durch die Transparenz und vielseitige Verwendbarkeit clusteranalytischer Verfahren im Vergleich zu den klassischen Methoden der Multivariaten Statistik hat in den letzten Jahren auf dem Hintergrund der zunehmenden Verbreitung mittlerer und größerer EDV-Anlagen eine immer rascher werdende Entwicklung eingesetzt, die einen ordnenden Überblick der diesem Problemkreis zuzurechnenden Ansätze, Methoden, Verfahren und Entwicklungstendenzen erforderlich macht. Pro Jahr dürften mehrere hundert einschlägige Arbeiten erscheinen, die entsprechend der weiten und oft unterschiedlichen Anwendung von Clusteranalysen auch unter Bezeichnungen wie Automatische Klassifikation, Numerische Taxonomie, Q-Analyse oder Unsupervised Learning zu finden sind.

Das vorliegende Buch stellt sich die Aufgabe einer einführenden Darstellung in die Grundlagen und Prinzipien der Clusteranalyse. Die wichtigsten clusteranalytischen Konzepte, Methoden und Verfahren werden ausführlich beschrieben. Für fast alle Verfahren werden Algorithmen und FORTRAN-IV Programme angegeben. Darüber hinaus werden zentrale Probleme der praktischen Durchführung einer Clusteranalyse sowie der Beurteilung und Interpretation der erreichten Ergebnisse diskutiert.

Das Buch wendet sich an Studierende der Wirtschafts- und Sozialwissenschaften, aber auch der Geographie, Biologie oder Medizin. Es dürfte ebenso für Studierende der Mathematik und Informatik, jedoch auch für den Praktiker von Interesse sein. Elementare statistische Kenntnisse sind nützlich. Die gebrauchten, allerdings wohl meist bekannten mathematischen Grundbegriffe werden im Anhang erläutert. Kenntnisse der Programmiersprache FORTRAN sind zum besseren Verständnis der Verfahren vorteilhaft, jedoch nicht notwendige Voraussetzung zum Studium des Textes. Da Einsatz und Interpretation der Clusteranalyse nicht zuletzt von der zugrunde liegenden inhaltlichen Fragestellung und damit vom speziellen Gegenstandsbereich abhängig sind, zielt dieses Buch auf eine Integration mathematisch-statistischer und anwendungsorientierter Grundlagen und Aspekte. Für den Leser soll dadurch zumindest ansatzweise jene künstliche Trennung aufgehoben werden, die zwischen den auf diesem Gebiet veröffentlichten mathematischen Arbeiten und rein pragmatisch angelegten Algorithmensammlungen einerseits sowie den ausgesprochen einführenden und stark disziplinspezifischen Beiträgen andererseits besteht.

Die hier verfolgte Konzeption einer möglichst systematischen Vermittlung theoretischer und anwendungsbezogener Momente führte zu folgendem Aufbau:

Im ersten Kapitel werden nach der Präzisierung der Problemstellung die relevanten Voraussetzungen und Begriffe geklärt.

Das zweite Kapitel enthält eine Übersicht über die bekanntesten multivariaten statistischen Verfahren und dient zur Einordnung der Clusteranalyse in diesen Kontext.

Die verschiedenen Vorgehensweisen, Ähnlichkeiten und Unähnlichkeiten zwischen Elementen und Elementgruppen zu definieren, werden im dritten Kapitel aufgegriffen.

Die dort erörterten Distanz- und Ähnlichkeitsfunktionen werden im vierten Kapitel, das einen wesentlichen Teil dieses Buches ausmacht, benötigt. Hier werden die einzelnen clusteranalytischen Verfahren eingehender beschrieben.

Spezielle Probleme der praktischen Durchführung oder der Beurteilung und Interpretation werden im fünften Kapitel diskutiert.

Kapitel sechs enthält einen zusammenfassenden Überblick, während das siebte Kapitel den Anhang bildet, in dem Grundbegriffe der Mengenlehre und Linearen Algebra erläutert sind.

Die am Ende des zweiten, dritten und vierten Kapitels zusammengestellten Übungen und Ergänzungen sollen zum einen zur Vertiefung des Stoffes beitragen. Zum anderen enthalten sie Herleitungen und weitere Überlegungen, auf die im laufenden Text aus Gründen der besseren Lesbarkeit verzichtet worden ist.

Die Verfasser sind zahlreichen Benutzern des Rechenzentrums der Westfälischen Wilhelms-Universität Münster und Studierenden zu Dank verpflichtet, die uns im Rahmen ihrer speziellen Anwendungsprobleme zur grundsätzlicheren Auseinandersetzung mit der Problematik der Clusteranalyse anregten. Die in diesen Projekten, Forschungsbereichen, Untersuchungsreihen und empirischen Arbeiten auftauchenden Fragen und Probleme führten zu instruktiven Hinweisen. Wir hoffen, die daraus gewonnenen Erfahrungen und Kenntnisse hier einem größeren Leser- und Benutzerkreis zugänglich machen zu können.

Das Manuskript korrigierten in dankenswerter Weise unsere Kollegen Dr. *M. Bestehorn*, Dr. *Hörmann*, Dipl. Päd. *M. Groth*, Dr. *H. Kamp* und Dr. *H. Pudlatz*. Neben wertvollen Anmerkungen steuerte Herr Dr. *H. Pudlatz* das Programm zur Erstellung der geographischen Karten bei. Insbesondere ist an dieser Stelle Herrn Prof. Dr. *W. Oberwittler* von der Medizinischen Universitätsklinik Münster für das medizinische Anwendungsbeispiel zu danken. Herrn *H. Mecke* danken wir für die sorgfältige Anfertigung der graphischen Abbildungen und Herrn *W. Herden* für die Hilfe bei der Implementierung einiger Programme. Davon unabhängig

liegt die Verantwortung für die Darstellung und deren Mängel bei den Verfassern, die Kritik, Anregungen und Hinweise aus dem Kreis der Leser gerne entgegennehmen.

Münster, im Juni 1977

Detlef Steinhausen
Klaus Langer

Inhalt

1. Einleitung

1.1 Problemstellung

In diesem Buch werden Ansätze, Verfahren und Algorithmen zur multivariaten Datenanalyse beschrieben und diskutiert, die zum Teil recht unterschiedlich bezeichnet werden (*Cluster-Analysis, Automatic Classification, Grouping* or *Clumping Strategies, Numerical Taxonomy, Q-Analysis* usw.). Ihre Gemeinsamkeit besteht jedoch in dem Ziel, Objekte nach bestimmten Prinzipien möglichst zweckmäßig oder optimal in Klassen, Gruppen oder Teilgesamtheiten aufzuteilen.

Dies Problem der möglichst sinnvollen oder nützlichen Gruppierung von Objekten stellt sich sowohl in alltäglichen Situationen als auch im Rahmen wissenschaftlicher Untersuchungen. Man versuche etwa, seine Schallplatten zu ordnen, was sicherlich geringe Schwierigkeiten bereitet, wenn man sich nach einem einzigen oder nur nach wenigen Kriterien richtet. Sollen allerdings Musikform, Epoche, Dirigent, Komponist, Orchester, Instrumentenart, Beliebtheitsgrad oder andere Aspekte gleichzeitig berücksichtigt werden, wird man nicht ohne weiteres eine angemessene Gruppierung finden und unter Umständen nach einer geeigneten Kombination der Merkmale (Musikform, Komponist, Orchester o. ä.) vorgehen.

Während man selbst bei vielen Schallplatten noch relativ leicht Gruppen bilden kann, entstehen nicht unerhebliche Schwierigkeiten bei dem Versuch, z. B. die Gesamtheit aller Patienten einer größeren Klinik anhand zahlreicher Meßwerte, die Labortests, Krankheitsbefunde, anamnestische und diagnostische Informationen einschließen, in jeweils ähnliche Teilgesamtheiten aufzugliedern, um gezielte Behandlungsmethoden, Symptombereiche oder effiziente Therapieprogramme entwickeln zu können. Analoge Schwierigkeiten tauchen auf, wenn man etwa die durch eine Vielzahl von Einzeldaten beschriebenen wirtschaftlichen Perioden so ordnen will, daß ähnliche Zeitabläufe jeweils einer bestimmten Gruppe zugewiesen werden.

Formal gesehen, besteht das Problem in all diesen Situationen darin, meist sehr viele Objekte, Einheiten oder Elemente in kleinere und homogene Gruppen, Klassen oder *Cluster* (engl. = Haufen, Traube) aufzuteilen. Die zu gruppierenden Elemente werden in der Regel durch zahlreiche Eigenschaften, Merkmale oder Variablen charakterisiert. Die auf diese Problemstellung bezogenen mathematisch-statistischen und heuristischen Verfahren der multivariaten Datenanalyse werden im folgenden zusammenfassend als *Clusteranalyse* bezeichnet.

Da Clusteranalysen in nahezu allen Disziplinen, in denen größere Datensätze auf wenige und überschaubare Interpretationseinheiten reduziert werden sollen, Ver-

wendung finden, kann ihr Gegenstand sehr verschieden sein. So kann es sich bei den Elementen z. B. um Personen, soziale Gruppen, Firmen, Produkte, Handschriften, Dokumente, Radarsignale, Rohstoffe, Aktien, Bakterien oder Insekten handeln.

Entsprechend verschieden sind die zur Gruppierung benutzten Variablen, deren Auswahl ebenso wie die der Elemente von dem betreffenden Untersuchungsziel abhängt. Durch die (empirisch ermittelten) Ausprägungen auf diesen meist sehr zahlreichen Variablen werden die Elemente näher gekennzeichnet (multivariate Information). Diese Daten sind der Ausgangspunkt des Gruppierungs- bzw. Klassifikationsprozesses.

Die Aufgabe, Elemente nach angemessenen mathematisch-statistischen und heuristischen Kriterien zu ordnen, ist klar von dem Problem zu unterscheiden, bereits vorgegebene Gruppen in einem noch näher zu definierenden Sinne optimal zu diskriminieren. Der erste Fall wird hier Klassifikationsproblem, der zweite Diskriminationsproblem genannt, obwohl beide Begriffe mitunter synonym gebraucht werden.

a) Beim *Diskriminationsproblem* geht es darum, schon vorgegebene (a priori) Gruppen oder Klassen möglichst optimal zu diskriminieren bzw. die Gruppenzugehörigkeit noch nicht eingeordneter Elemente mit möglichst hoher Wahrscheinlichkeit anzugeben. Die Definition und Existenz bestimmter Gruppen, die mit Hilfe eines Außenkriteriums, der sogenannten Gruppierungsvariablen erfolgt, wird dabei kennzeichnenderweise vorausgesetzt.

b) In Erweiterung der diskriminanzanalytischen Fragestellung werden beim *Klassifikationsproblem* die Klassen oder Gruppen erst gesucht. Weder Anzahl, Homogenität oder Lokalisation der Gruppen sind bekannt, noch besitzt man Informationen über die Zuordnung einzelner Elemente zu den Gruppen. Üblicherweise sind ‚Klassen' elementfremde (disjunkte), Gruppen hingegen nicht notwendig disjunkte Teilmengen. Da die gesuchten Teilmengen disjunkt oder nicht disjunkt sein können, kann man auch allgemeiner vom *‚Gruppierungsproblem'* sprechen.

Diskriminanzanalytische Verfahren sind insbesondere in der medizinischen Diagnostik verbreitet und werden ausführlicher in der Literatur behandelt [Cacoullos 1973]. Auf die Kombination der statistischen Verfahren der Diskriminanzanalyse und der Clusteranalyse wird im weiteren Verlauf noch Bezug genommen.

In diesem Buch wird von verschiedenen Aspekten aus das Klassifikations- oder Gruppierungsproblem diskutiert. Voraussetzung dazu ist zunächst eine einführende Klärung des Begriffs Clusteranalyse.

1.2 Zum Begriff „Clusteranalyse"

Die bei clusteranalytischen Verfahren benutzten Prinzipien lassen sich, wenn auch zum Teil in modifizierter Form, bereits im Anfangsstadium wissenschaftlicher Arbeit nachweisen, zumal die Bildung von Gruppen, Klassen oder Teilmengen sehr oft ein notwendiges und gleichzeitig ökonomisches Mittel zur Informationsreduktion ist. Im Jahr 1939 wurde der Begriff Clusteranalyse ausdrücklich von Tyron verwandt, diente allerdings zur Kennzeichnung eines speziellen Verfahrens zur Gruppierung von Variablen, das in enger Anlehnung an das faktorenanalytische Modell konzipiert worden war [Tyron 1939, Tyron & Bailey 1970].

Während dieser Terminus, der sich im weiteren Verlauf primär auf Verfahren zur Gruppierung von Objekten bezog, nur vereinzelt in der Literatur vorkommt, ist seit Anfang der 70er Jahre ein verstärktes Interesse festzustellen, das sich auch im raschen Anwachsen der einschlägigen Literatur bemerkbar macht (als Übersicht [Duran & Odell 1974]).

So erschienen umfangreichere Arbeiten mit Titeln wie *‚Cluster-Analysis for Applications'*, *‚Clustering Algorithms'* bzw. *‚Cluster-Analyse-Algorithmen'* [Anderberg 1973, Hartigan 1975, Späth 1975], was nicht nur die Aktualität des Themas, sondern zugleich den Schwerpunkt der meisten Publikationen auf diesem Gebiet zum Ausdruck bringt, in denen praktikable Algorithmen und nicht etwa ‚highly sophisticated' mathematisch-statistische Konzepte dargestellt werden.

Sucht man umfassende und theoretisch ausgerichtete Analysen des Gruppierungsproblems, wird man auf Arbeiten zur *‚Automatischen Klassifikation'* und *‚Computer Klassifikation'* verwiesen, die für den deutschen Sprachraum präzis und ausführlicher von [Bock 1974] behandelt worden sind.

Methoden und Techniken der Clusteranalyse sind vor allem in biologischen Wissenschaften, zum geringeren Teil in der Psychologie entwickelt worden. Daher finden sich im Rahmen der biologischen Typenkonstruktion und -interpretation wesentliche Ansätze und Verfahren der Clusteranalyse. Weil bei einer Typenkonstruktion disziplinspezifische Aspekte eine große Rolle spielen, spricht man in diesem Kontext von (biologischer) *Taxonomie*, bzw. unter Betonung methodischer Aspekte von *Numerischer Taxonomie* und *Mathematischer Taxonomie* oder kurz von *Taxonometrie* (vgl. die Standardwerke von [Sokal & Sneath 1963], [Jardine & Sibson 1971]). Bezogen auf die biologische Klassifikationseinheit *Taxon*, wird Taxonomie als das theoretische Studium der Klassifikation, einschließlich ihrer Voraussetzungen, Prinzipien, Prozeduren und Regeln, bestimmt. Die Begriffe Taxonomie und Taxonometrie werden inzwischen auch in anderen Wissenschaften benutzt [Cattell & Coulter 1966, Goronzy 1969].

Die Unterschiedlichkeit der bestehenden Konzepte zur Datengruppierung ist mit ein Grund dafür, daß die zugehörigen statistischen Verfahren nicht einheitlich abgegrenzt werden und eine stringent abgeleitete, exakte Definition von Clusteranalyse, wie dies mittlerweise für andere multivariate Verfahren gelungen ist, noch aussteht.

Insofern ist die hier vorgeschlagene Begriffsbestimmung vorläufig. Clusteranalyse wird verstanden als ein zusammenfassender Terminus für eine Reihe unterschiedlicher mathematisch-statistischer und heuristischer Verfahren, deren Ziel darin besteht, eine meist umfangreiche Menge von Elementen durch Konstruktion homogener Klassen, Gruppen oder Cluster optimal zu strukturieren. Die gesuchten Cluster sollen jeweils nur ähnliche Elemente enthalten, während Elemente verschiedener Gruppen möglichst unähnlich sein sollen. Bei dieser Aufteilung der Elemente wird davon ausgegangen, daß die Ähnlichkeit der Elemente untereinander quantifizierbar ist und sich durch (reelle) Zahlenwerte ausdrücken läßt. Diese allein bilden die Grundlage der Gruppierung, die somit ausschließlich nach mathematisch-statistischen und heuristischen Prinzipien, nicht jedoch auf Grund intuitiver, substanzwissenschaftlicher oder anderer Kriterien erfolgt. Wie bereits oben durch die Adjektive mathematisch, numerisch bzw. automatisch erkennbar, handelt es sich bei der Clusteranalyse um ‚objektive' Gruppierungen im Unterschied zu ‚subjektiven' Gruppierungen.

1.3 Ziel und Funktion

Ziel und Funktion von Clusteranalysen können sehr unterschiedlich sein. Generell besteht das Ziel clusteranalytischer Verfahren in einer vereinfachenden Darstellung der Struktur der vorgegebenen Menge von Elementen. Das Prinzip systematischer Informationsverdichtung wird angewandt, um aus einer Fülle von Einzeldaten wesentliche Charakteristika der Struktur der Objektmenge erkennen zu können.

Die Identifizierung der Struktur einer vorgegebenen Menge erlaubt jedoch keine Aussagen über die Beziehung der erstellten Lösung zu einer bestimmten Grundgesamtheit. Clusteranalysen haben die Funktion der *Datenstrukturierung*, nicht jedoch die der Schätzung spezieller Parameter einer Population. In Abgrenzung zu entsprechenden inferenzstatistischen Verfahren kann man Clusteranalysen als deskriptive Verfahren betrachten, obgleich die Organisation der Daten auf einem vergleichsweise höheren Niveau erfolgt als die Berechnung normaler statistischer Kenngrößen.

Von dort werden Clusteranalysen auch oft zum *faktorenanalytischen Modell* in Beziehung gesetzt, dessen Prinzip bekanntlich in der Reduzierung einer Variablenvielfalt auf wenige, als Faktoren oder Dimensionen bezeichnete, hypothetische

Größen besteht. Da Faktorenanalyse und Clusteranalyse gewisse Gemeinsamkeiten besitzen, auf die im zweiten Kapitel spezieller eingegangen wird, könnte man etwas ungenau Clusteranalysen auch zu den dimensionsstatistischen Verfahren rechnen. Man vernachlässigt hierbei jedoch, daß Clusteranalysen nicht zur Grundkomponenten-, sondern zur Grundmusterzerlegung führen, was noch ausführlich dargestellt wird. Clusteranalysen befinden sich stärker auf der Ebene der Beobachtungsdaten. Damit ist verbunden, daß Clusteranalysen sehr oft geringere Anforderungen an die Datenqualität stellen und in vielen Situationen angemessener sein dürften als das häufig benutzte multivariate Standardmodell der Faktorenanalyse.

Die konkrete Funktion von Clusteranalysen wird einsichtig, wenn man die praktischen Anwendungen analysiert, von denen wesentliche ohne Anspruch auf eine geschlossene Systematik kurz skizziert werden sollen:

- Verhaltensweisen, Einstellungen, Testergebnisse, Organisationen, soziale Einheiten, Sprachen, soziale Prozesse usw. sind Gegenstand von Clusteranalysen in den *Sozialwissenschaften*. Die Anwendungen in der Soziologie, Psychologie, Kriminologie, Pädagogik, Publizistik, Politikwissenschaft usw. sind ausgesprochen zahlreich und vielfältig. Mittels clusteranalytischer Verfahren werden z. B. gruppenspezifische Urteilstendenzen bei Lehrern (Lehrertypen) gesucht, Erziehungsstile nachgewiesen, Studenten nach ihren gesellschaftlich-politischen Attitüden gruppiert, Kommunikationsformen in sozialen Gruppen klassifiziert oder psychologische Testprofile strukturiert. Im schulischen Sektor interessiert man sich etwa für homogene Gruppen von ‚leistungsschwachen' Schülern, um differenzierte Förderungsmaßnahmen einleiten zu können. Ebenso lassen sich mittels clusteranalytischer Verfahren informelle soziale Subgruppen in der Schulklasse oder in anderen sozialen Verbänden und Kollektiven aufdecken. Zur Identifizierung dieser Interaktionsstrukturen würde man auf der Basis soziometrischer Daten die Schüler schrittweise danach zu Clustern zusammenfassen, inwieweit sie ähnliche Präferenzen abgeben. Ein anderes Gruppierungskriterium wäre etwa die Ähnlichkeit der Schüler hinsichtlich der erhaltenen positiven Wahlen. Zieht man die verschiedenen Konzepte der Clusteranalyse in Betracht, dürfte man in Zukunft bei derartigen Untersuchungen in einem weit größeren Maße als bisher zu pädagogisch, psychologisch oder soziologisch relevanten Ergebnissen kommen.

- In den *biologischen* und *medizinischen Wissenschaften* sollen z. B. Pflanzen, Tiere, Mikroorganismen oder Patienten, Krankheiten, Symptome, Laborergebnisse usw. gruppiert werden. Analysiert wird die Strukturierung zellulärer Einheiten, die anhand ihrer Stoffwechselfunktion, Größe, Färbbarkeit und ihres Proteingehalts o. ä. bestimmten Clustern zugeteilt werden. Gerade aus der Biologie ist die Problematik einer adäquaten Systematisierung der Lebewesen geläufig. Umfassende taxonometrische Systeme werden entworfen, die

man mit Hilfe von Clusteranalysen konstruiert. Vorläufer derartiger Klassifikationen sind die bekannten Systeme von Linne (1735) und Adanson (1757). In der Medizin werden Symptomklassen gebildet, um Diagnose, Therapie und Prognose zu verbessern.

- Ebenso vielfältig sind die Anwendungen in den *Wirtschaftswissenschaften*. Gegenstand sind hier Firmen, Produkte, Konsumenten, Verkaufsprogramme oder Aufträge. Relativ verbreitet ist die Strategie der Marktsegmentierung mittels clusteranalytischer Verfahren, um möglichst homogene Absatzmärkte (Segmente) herauszufinden.

- *Mathematisch-naturwissenschaftliche* und *technische* Anwendungen sind gleich verstreut wie unterschiedlich. Speziell werden Clusteranalysen im Rahmen der Mustererkennung (Pattern Recognition), Künstlichen Intelligenz, Informationswiedergewinnung oder allgemeiner der angewandten Systemtheorie eingesetzt. Sehr differenziert sind Verfahren und Methoden zur Klassifikation von Handschriften, Sprachsignalen, Fingerabdrücken oder etwa Radarsignalen.

Allen Beispielen ist gemeinsam, daß Einsatz und Interpretation der Clusteranalyse entscheidend vom Untersuchungsziel determiniert werden. Wesentliche Fragen und Schwierigkeiten entstehen jedoch nicht nur auf dem mathematischen und statistischen Gebiet, sondern – wie noch deutlich wird – im Bereich der Anwendung der Algorithmen.

1.4 Das Clusteranalyseproblem

Das breite Spektrum der Anwendungsmöglichkeiten sollte dennoch nicht darüber hinwegtäuschen, daß die mittels clusteranalytischer Verfahren erzielten Lösungen meist keine eindeutigen Lösungen sind und nur zu lokalen Extrema führen. Man vergegenwärtige sich zunächst, wieviele Gruppierungen einer Menge existieren.

Sei $E = \{e_1, ..., e_n\}$ die nichtleere, endliche Menge der zu gruppierenden n Elemente und **G** eine *Gruppierung* (Zerlegung) von E in k elementfremde Teilmengen oder Gruppen g_i, deren Vereinigung E ergibt, also

$$E = \bigcup_{i=1}^{k} g_i \qquad \text{und} \qquad g_i \cap g_j = \emptyset \qquad (i, j = 1, ..., k; i \neq j) .$$

Die Anzahl $S(n, k)$ der bei k Gruppen möglichen disjunkten Zerlegungen (*Partitionen*) **G** von E läßt sich über die Formel berechnen

$$S(n,k) = \frac{1}{k!} \cdot \sum_{i=0}^{k} (-1)^i \cdot \binom{k}{i} \cdot (k-i)^n$$

bzw. im Anwendungsfall einfacher rekursiv über

$$S(n+1,k) = S(n,k-1) + k \cdot S(n,k) \quad \text{mit} \quad S(n,1) = S(n,n) = 1$$
$$S(n,k) = 0 \quad \text{für} \quad n < k\,.$$

Speziell für $k = 2$ ergibt sich

$$S(n,2) = 2^{n-1} - 1\,.$$

Dies wird insbesondere für die hierarchische Clusteranalyse (Kap. 4.2) noch von Bedeutung sein.

Die $S(n, k)$ werden auch *Stirling'sche Zahlen zweiter Art* genannt und besitzen etwa für $n = \{1, 2 \ldots, 10, 15, 20, 25, 50, 100\}$ und $k = \{2, 3, 4, 5, 10\}$ folgende Werte:

STIRLING'SCHE ZAHLEN ZWEITER ART

N \ K	1	2	3	4	5	10
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	3	1	0	0	0
4	1	7	6	1	0	0
5	1	15	25	10	1	0
6	1	31	90	65	15	0
7	1	63	301	350	140	0
8	1	127	966	1701	1050	0
9	1	255	3025	7770	6951	0
10	1	511	9330	34105	42525	1
15	1	16383	2375101	42355950	2108 E09	12662650
20	1	524287	580606446	4523 E07	7492 E08	5918 E09
25	1	1678 E04	1412 E08	4677 E10	2437 E12	1203 E15
50	1	5629 E11	1196 E22	5282 E25	7401 E29	2615 E40
100	1	6338 E28	8590 E43	6704 E55	6574 E64	2756 E90

YY

ANMERKUNG: XXXX EYY BEDEUTET :XXXX*10

Abb. 1.4.1 Stirling'sche Zahlen zweiter Art

Wird die Anzahl k der Gruppen nicht vorherbestimmt, kann man die Anzahl B_n aller möglichen Partitionen rekursiv ermitteln, indem man folgende Formel benutzt

$$B_n = \sum_{i=1}^{n} S(n, i) ,$$

woraus sich für $n \geqslant 0$ ergibt:

$$B_{n+1} = \sum_{i=0}^{n} \binom{n}{i} \cdot B_i \quad \text{mit} \quad B_0 := 1; B_1 = 1 .$$

Diese Zahlen werden *Bell'sche Zahlen* oder *Exponentialzahlen* genannt.

BELL'SCHE ZAHLEN

N	B(N)	N	B(N)
1	1	15	1383 E06
2	2	20	5172 E10
3	5	25	4639 E14
4	15	30	8467 E20
5	52	35	2816 E26
6	203	40	1575 E32
7	877	45	1393 E38
8	4140	50	1857 E44
9	21147	60	9769 E56
10	115975	70	1807 E70

Abb. 1.4.2 Bell'sche Zahlen

Das der Intuition nachgestaltete Vorgehen, alle möglichen Zerlegungen von E zu bilden und diejenige auszuwählen, die nach einem bestimmten Kriterium die beste aller Zerlegungen darstellt, ist bereits für kleine n zu aufwendig (*Methode der totalen Enumeration*). Wenn man etwa für jede Gruppierung den Wert der entsprechenden Gütefunktion in einer Mikrosekunde berechnen könnte, würde man zum Auffinden des Optimums dieser Funktion bei $n = 20$ und $k = 5$ rund 208 Stunden, bei $n = 50$ und $k = 3$ rund $4 \cdot 10^{12}$ Jahre benötigen.

Selbst wenn durch eine neue Computergeneration die Berechnungszeit um den Faktor 10^2 oder 10^3 zu reduzieren wäre, ist die Methode der Berechnung aller Zerlegungen, die sicher zum wahren Extremum führt, praktisch nicht realisierbar, so daß die Notwendigkeit zur Entwicklung effizienter Methoden und Verfahren entsteht.

In Anlehnung an Methoden der Optimierung wie der *dynamischen Programmierung* [Jensen 1969; Rao 1971] oder der *Branch and Bound-Verfahren* lassen

sich Strategien der Clusteranalyse entwickeln [Koontz et al 1975], durch die eine hohe Anzahl überflüssiger, schlechter Gruppierungen ausgeschaltet und somit das Optimum aus einer geringeren Zahl besserer Gruppierungen gefunden werden kann. Sie sind allerdings bisher noch nicht soweit entwickelt, daß sie eine wirtschaftliche Alternative zu den heuristischen Verfahren darstellen.

Letztere sind dadurch charakterisiert, daß sie ausgehend von einer mehr oder weniger plausiblen Optimierungsstrategie meist sehr rasch zu Lösungen kommen, welche nur *lokal optimal* oder *suboptimal* sind. Sie spielen in der Praxis bei begrenzter Rechenzeit und großen Fallzahlen derzeit noch die größte Rolle, da man sich dort oft mit suboptimalen Lösungen zufrieden gegeben hat. Das eventuell lediglich um wenige Prozent bessere globale Optimum ist oft nur durch einen unrealistisch hohen Mehraufwand an Rechenzeit zu erreichen.

Bei der Beurteilung von Clusterlösungen üben daher subjektive und vom jeweiligen Untersuchungskontext abhängige Faktoren einen nicht unwesentlichen Einfluß aus.

1.5 Ablaufschema

Die Vermittlung formaler und inhaltlicher, vom Untersuchungsziel bestimmter Gesichtspunkte wird ebenso deutlich, wenn man die zur Durchführung einer Clusteranalyse notwendigen Schritte oder Phasen zu gliedern versucht. Vereinfachend ergeben sich acht solcher Phasen, an denen zugleich der Aufbau dieses Buches zu erkennen ist.

(1) Präzisierung der Untersuchungsfragestellung

(2) Auswahl der Elemente und Variablen

(3) Aufbereitung der Daten

(4) Festlegung einer angemessenen Ähnlichkeitsfunktion

(5) Bestimmung des geeigneten Algorithmus zur Gruppierung

(6) Technische Durchführung

(7) Analyse der Ergebnisse (Postanalyse)

(8) Interpretation der Ergebnisse

zu (1)
Obwohl dieser Punkt oft vernachlässigt wird, halten wir die vorherige Präzisierung der Untersuchungsfragestellung für außerordentlich wichtig, da von ihr ausgehend alle nachfolgenden Schritte zu bearbeiten sind. Clusteranalysen werden häufig in Kombination oder Ergänzung zu anderen multivariaten Verfahren eingesetzt. Man sollte vor der Berechnung das inhaltliche Problem, das zur Verwen-

dung der Clusteranalyse führt, so weit wie möglich spezifizieren, um überhaupt Stellenwert und Funktion der Gruppierung angeben zu können.

zu (2)
Die ausgewählten Elemente und Variablen sollen sich eindeutig auf das Untersuchungsziel beziehen und für den untersuchten Bereich hinreichend repräsentativ sein.

zu (3)
Die Datenaufbereitung besteht zunächst darin, die ermittelten Meßwerte in Form einer *Rohdatenmatrix* $\mathbf{X}^R = (x_{ij}^R)$ anzuordnen, bei der x_{ij}^R in der i-ten Zeile und in der j-ten Spalte den Meßwert der j-ten Variable für das i-te Element darstellt.

Entsprechend der inhaltlichen Zielsetzung sind die Rohdaten eventuell zu korrigieren, indem man fehlende Daten (‚missing data') berücksichtigt, bestimmte Kenngrößen berechnet oder z. B. die Daten normiert bzw. standardisiert. Grundlage der Gruppierung bildet dann die *Datenmatrix* $\mathbf{X} = (x_{ij})$. Vor der eigentlichen Konstruktion der Gruppen kann man diese Matrix unter Umständen in einem weiteren Schritt mittels bekannter multivariater Verfahren (z. B. Faktoren-, Hauptkomponentenanalyse) auf eine Matrix niederen Ranges reduzieren und diese zum Ausgangspunkt der Gruppierung machen.

zu (4)
Die vielen Ansätze, aufbauend auf der in dieser Matrix $\mathbf{X}$ enthaltenen Information, Ähnlichkeiten zu definieren, werden im dritten Kapitel beschrieben. Die Ähnlichkeitsfunktion ist danach zu bestimmen, inwieweit sie den Variablen und dem Variablentyp sowie dem Untersuchungsziel angemessen ist. Oft gelangt man zu einer (symmetrischen) *Ähnlichkeitsmatrix* $\mathbf{S} = (s_{ij})$. Da Ähnlichkeitsfunktionen leicht in Unähnlichkeitsfunktionen transformierbar sind, ergibt sich auch die *Unähnlichkeitsmatrix* (*Distanzmatrix*) $\mathbf{D} = (d_{ij})$. Die Zahl s_{ij} bzw. d_{ij} gibt die Ähnlichkeit bzw. Distanz des i-ten Elements mit dem j-ten Element wieder.

zu (5)
Wie auch die Wahl einer adäquaten Distanzfunktion ist die Bestimmung eines geeigneten Clusteranalysealgorithmus, die zumeist die Festlegung eines Gütekriteriums zur Beurteilung der erstellten Lösung impliziert, nicht nur von formalen Aspekten aus zu klären. Sind in der Praxis oft externe Kriterien wie Rechenzeit, Speicherplatzbedarf oder Verfügbarkeit der Programme ausschlaggebend für den Einsatz eines Verfahrens, sollte man vielmehr den speziellen Algorithmus nach der Bedeutung auf das Gruppierungsresultat hin auswählen. Im vierten Kapitel wird erläutert, daß je nach implementiertem Algorithmus für dieselben Elemente verschiedene und zum Teil ganz unterschiedliche Gruppierungen gebildet werden können.

zu (6)
Bei den in der Praxis vorkommenden Fall- und Variablenzahlen ist die technische Durchführung einer Clusteranalyse nur mit Hilfe eines Computers möglich. Da bestimmte Algorithmen nur bei geschickter Implementierung anwendbar sind, sind Fragen der effektiven Programmierung sehr wichtig. Sie werden so weit wie möglich jeweils bei der Darstellung der einzelnen Algorithmen behandelt.

zu (7)
Hierunter ist zunächst die formale Analyse des Gruppierungsresultats zu verstehen, die sich etwa auf die Beurteilung richtet

- der Homogenität der gebildeten Cluster,
- der Differenz der Clustermittelpunkte,
- des Einflusses bestimmter Variablen und Elemente oder
- der Bedeutung der Startnäherung.

An diesem Punkt wird also die Angemessenheit bzw. Optimalität der gefundenen Lösung von statistischer Seite aus beurteilt.

zu (8)
In Bezug auf die im ersten Punkt (s. o.) geleistete Präzisierung der Untersuchungsfragestellung werden die Cluster genauer analysiert und interpretiert. Hier werden etwa die clusteranalytischen Ergebnisse mit anderen im Datensatz vorhandenen, sog. *‚passiven'* Variablen verglichen, wodurch weitere inhaltliche Aufschlüsse über die Cluster gewonnen werden können.

Während im anschließenden zweiten Kapitel dieses Buchs die Prinzipien anderer multivariater Verfahren skizziert und gegeneinander abgegrenzt werden, um dem Leser eine weitere Orientierungshilfe zum Thema Clusteranalyse zu geben, läßt sich der weitere Verlauf der Darstellung an diesem Schema verfolgen. Der Aufbau dieses Buchs richtet sich nach den bei der praktischen Durchführung einer Clusteranalyse notwendigen Schritten.

Da Punkt 1 und 2 zweckmäßigerweise nur von der konkreten inhaltlichen Untersuchung her zu beantworten sind, werden sie an dieser Stelle nicht eigens erörtert. Soweit es sinnvoll erscheint, werden im Text allgemeine Hinweise zur Lösung entsprechender Fragen zu finden sein.

Im dritten Kapitel werden *Ähnlichkeits- bzw. Unähnlichkeitsfunktionen (Distanzfunktionen)* zwischen Elementen und Elementgruppen diskutiert. Dies entspricht den Punkten 3 und 4.

Den Schwerpunkt des Buchs, der in den Punkten 5 und 6 zu finden ist, bilden die *Clusteranalysealgorithmen*, die im vierten Kapitel dargestellt werden. Nach der Beschreibung des speziellen Algorithmus werden Anwendbarkeit, Optimali-