§ 1. Problemstellung, Grundlagen

a) Einführung

In vielen Lebensbereichen stellt sich das Problem, daß eine Vielzahl ungeordneter Objekte (Personen, Gegenstände, Krankheiten oder Dokumente) aufgrund von Ähnlichkeit und sachlicher Verwandtschaft in kleinere, homogene und praktisch nützliche Klassen (Gruppen) zerlegt werden muß. Dieses Problem kennt jeder, der einmal seine Bücher getrennt nach der Art ihres Inhalts aufstellen wollte und hierfür geeignete „Sachgruppen" einrichten mußte; oder der seine Literaturkartei systematisch anordnen will und hierfür ein passendes Klassifikationsschema sucht. In größerem Rahmen tritt dasselbe Problem aber auch in der Medizin, der Psychologie und der Biologie auf, wo etwa Bakterien- oder Krankheitsarten zu gruppieren sind und jede der betreffenden Gruppen gesondert behandelt werden muß. Andere Anwendungen erstrecken sich auf Fragen der Mustererkennung (automatisches Lesen von Handschriften, Auswertung von Radarsignalen), auf Informationssysteme (Datenbanken), ferner auf technische, volkswirtschaftliche und organisatorische Fragen. Dabei sind typischerweise sehr viele Objekte zu klassifizieren und überdies sehr viele Merkmale und Eigenschaften dieser Objekte zu berücksichtigen.

Das Ziel einer Klassifikation kann sehr unterschiedlich sein: Dadurch, daß man gleichartige oder ähnliche Dinge in homogene Klassen zusammenfaßt und die Merkmale jeder Klasse als repräsentativ für die betreffenden Objekte ansieht, ist es zunächst möglich, die Struktur der betrachteten Objektmenge vereinfacht darzustellen und die Vielfalt der beobachteten Erscheinungsformen auf ein erträgliches, überschaubares Maß zu reduzieren. Das Prinzip der Klassenbildung erweist sich somit als eine Methode der Datenreduktion und insofern – ähnlich wie die Abstraktion – als ein nützliches Hilfsmittel zur Erkenntnis neuer und unbekannter Zusammenhänge.

Bei praktischen Anwendungen geht es sehr häufig um die Erkennung von „wahren" oder „natürlichen" Gruppierungen: Man weiß oder vermutet von vornherein, daß die betrachtete Objektmenge in mehrere kleine, prinzipiell gut unterscheidbare Gruppen (Populationen) zerfällt; indessen sind weder die Zuordnung der Objekte zu den einzelnen Gruppen noch die kennzeichnenden Charakteristika dieser Gruppen bekannt. Man sucht dann die unbekannte Gruppierung dadurch zu rekonstruieren, daß man ähnliche Objekte in ho-

mogene Klassen zusammenfaßt, und erwartet, daß jede gefundene Klasse eine Population bzw. einen Objekt-„Typ" repräsentiert.

Aber auch dann, wenn eine Objektmenge keine natürliche Gruppenstruktur aufweist, kann eine Klassenbildung nützlich sein: Die Klassifikation erfolgt dann – wie etwa die Einführung von Güteklassen oder Lohngruppen – lediglich aufgrund ihrer praktischen oder organisatorischen Zweckmäßigkeit und hat rein deskriptiven Charakter. Das Problem besteht hier primär in der Festlegung einer „zweckmäßigen" Klassifikation.

„Automatische Klassifikation" ist ein Sammelbegriff für eine Reihe mathematischer und statistischer Verfahren mit dem Ziel, in einer gegebenen Objektmenge homogene Klassen ähnlicher Objekte zu entdecken und eine optimale oder möglichst zweckmäßige Gruppierung zu konstruieren[1].

Hierzu sind gewisse Ausgangsdaten erforderlich: Die genannten Verfahren gehen sämtlich von der Annahme aus, daß die „Ähnlichkeit" oder die „Zusammengehörigkeit" zweier Objekte – wie immer diese auch interpretiert werden mag – numerisch durch Zahlenwerte erfaßbar ist oder daß die Ähnlichkeit verschiedener Objektpaare wenigstens verglichen werden kann (vgl. Abschnitt d). Aufgrund dieser (und nur dieser) Information wird die gesuchte Gruppierung der Objekte konstruiert, wobei ausschließlich mathematisch-statistische Kriterien über die Zusammensetzung der Klassen entscheiden („objektive" im Gegensatz zu „subjektiver" Klassifikation).

Zur Illustration der behandelten Fragestellung mögen die folgenden, anschaulichen Beispiele dienen. Sie zeigen, welche Ausgangsdaten einem Klassifikationsproblem zugrunde liegen können, und umreißen den Anwendungsbereich der späteren Gruppierungsverfahren.

b) Einige Anwendungsbeispiele

Beispiel 1.1.: Die Bekleidungsindustrie will neue Konfektionsmaße für Herrenbekleidung festlegen. Hierbei will man erreichen, daß ein möglichst großer Teil der (männlichen) Bevölkerung zufriedengestellt wird, d.h. „passende" Anzüge kaufen kann. Aus Rationalisierungsgründen können jedoch nur $m = 30$ Größen geführt werden. Um die 30 zugehörigen Schnittmustermaße optimal festzulegen, kann man so vorgehen: Man wählt aus der Bevölkerung eine hinreichend große, zufällige Stichprobe aus, etwa $N = 1000$ Männer, und mißt deren Körpermerkmale: Größe, Hals-, Bauch-, Brustumfang, Armlänge usw.,

[1]) Vgl. Bemerkung 1.2.

insgesamt etwa p Merkmale. Diese Maße werden nach Art von Abb. 1.1 in einer $N \times p$-Tabelle festgehalten, wobei die Zeilen den N Personen (Objekte $0_1, \ldots, 0_N$) und die Spalten den Merkmalen $(M_1, \ldots, M_p)$ entsprechen. Es

	M_1	M_2	$M_3 = M_p$	
0_1	63,0	16,5	9,0	x_1'
0_2	48,8	23,0	14,0	x_2'
0_3	50,3	14,2	2,5	x_3'
0_4	58,5	14,0	12,0	x_4'
$0_5 = 0_N$	81,8	2,2	4,5	x_5'

Abb. 1.1: Datenmatrix $X = (x_{kj})$ für quantitative Daten ($N = 5$, $p = 3$).

ist anschaulich klar, daß die Ähnlichkeit bzw. Unähnlichkeit zweier Personen (hinsichtlich des Körperbaus) durch Vergleich der beiden entsprechenden Zeilen dieser Matrix festgestellt werden kann. Man teilt nun die $N = 1000$ Personen derart in $m = 30$ Klassen auf, daß innerhalb jeder Klasse die Körpermaße möglichst wenig variieren, die Ähnlichkeit also groß ist (Klassifikationsproblem). Für jede der gefundenen Gruppen wird dann eine eigene Konfektionsgröße eingeführt, deren Schnittmustermaße sich z. B. aus den mittleren Körpermaßen der betreffenden Personen ergeben. – Da den meisten Menschen eine einzige Konfektionsgröße genügt, darf hier jede Person nur einer einzigen Gruppe angehören; wir sprechen dann von einer *disjunkten* Gruppierung (Abb. 1.2).

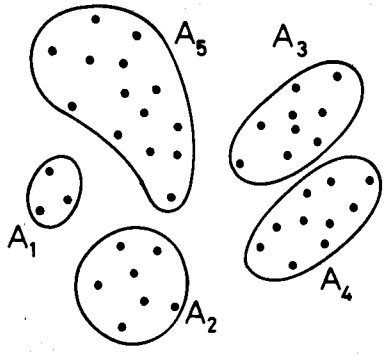

Abb. 1.2: Disjunkte, exhaustive Gruppierung von $N = 43$ Objekten (Punkten) in $m = 5$ Klassen.

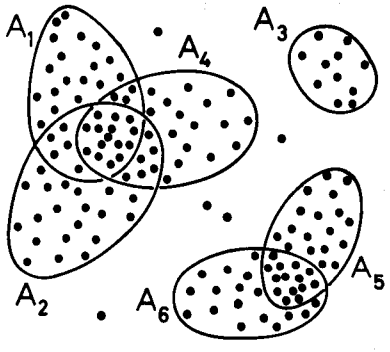

Abb. 1.3: Nichtdisjunkte Gruppierung mit $m = 6$ Klassen.

Beispiel 1.2: Ein Psychologe untersucht experimentell das menschliche Sozialverhalten. Er will feststellen, ob und welche typischen, sozialen Verhaltensmuster es gibt[1]). Hierzu hat er N Personen (= Objekte) ausgewählt und jede davon in jeweils p Testsituationen beobachtet, wobei die Reaktion bei jedem Test als „positiv" (= 1) oder „negativ" (= 0) beurteilt wurde. Das Resultat der Versuchsreihe läßt sich dann nach Art von Abb. 1.4 in einer $N \times p$-Tabelle zusammen-

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	$M_{10} = M_p$
0_1	1	1	1	0	1	1	1	1	1	0
0_2	1	0	1	0	1	1	1	1	1	0
0_3	0	1	1	1	1	1	1	1	1	0
0_4	1	0	1	0	0	1	1	1	1	0
$0_5 = 0_N$	1	0	0	0	0	1	0	1	0	1

Abb. 1.4: Datenmatrix $X = (x_{ki})$ für binäre Merkmale ($N = 5$, $p = 10$).

fassen. Zur Analyse dieser Daten zerlegt der Psychologe die Menge der N Prüflinge aufgrund ihrer Testergebnisse in möglichst homogene Klassen; er sucht also solche Personengruppen, deren Mitglieder weitgehend ähnliche Reaktionen zeigen bzw. anders reagieren als die Angehörigen anderer Gruppen. Jede dieser Gruppen charakterisiert dann einen eigenen „Verhaltenstyp". Wegen des Auftretens „gemischter" Verhaltensweisen erscheint es hier realistisch, eine gewisse Überschneidung der Klassen (Typen) zuzulassen; man spricht dann von einer *nichtdisjunkten* Klassifikation (Abb. 1.3).

Beispiel 1.3: Es gibt sehr viele Arten von Schnupfenviren. Zur vorbeugenden Bekämpfung des Schnupfens ist es notwendig, die Menge aller dieser Virenarten in mehrere Gruppen einzuteilen und für jede Virengruppe einen eigenen, artspezifischen Impfstoff zu entwickeln. Hierzu stellt etwa ein Bakteriologe experimentell die verschiedenen Eigenschaften dieser Viren fest (Resistenz, Vermehrung etc.) und faßt – bei N Virenarten und p untersuchten Merkmalen – das Untersuchungsergebnis ähnlich wie früher in einer $N \times p$-Tabelle zusammen. Anhand dieser Daten sollen dann die gesuchten Gruppen konstruiert werden. Man interessiert sich dabei einerseits für sehr feine Gruppierungen, bei denen kleine und sehr homogene Klassen entstehen (Feinstruktur), und andererseits auch für gröbere Klassifikationen, bei denen einzelne Virenstämme zu größeren Einheiten zusammengefaßt sind (Makrostruktur). Es emp-

[1]) Zur Lösung solcher Probleme sind die Methoden der automatischen Klassifikation oft besser geeignet als die bei Psychologen häufig benutzte Faktoranalyse.

fiehlt sich deshalb eine *hierarchische* Klassifikation, bei der die Gruppen einander nach Art eines „Stammbaums" über- bzw. untergeordnet sind (Abb. 1.5). Die homogensten Klassen stehen dabei am unteren Ende der Hierarchie und verschmelzen sukzessiv zu größeren Gruppen. Wenn diese Hierarchie die Ähnlichkeitsstruktur der Objekte hinreichend berücksichtigt, so kann sie u. U. zur Erstellung eines Diagnoseschlüssels dienen. – Allgemein ist die Klassifikation von Krankheiten, Bakterien o. ä. ein wichtiger Schritt zur Entwicklung automatischer Diagnosesysteme, weil dort die einzelnen Krankheitstypen (-klassen) und ihre charakteristischen Symptome bereits bekannt sein müssen.

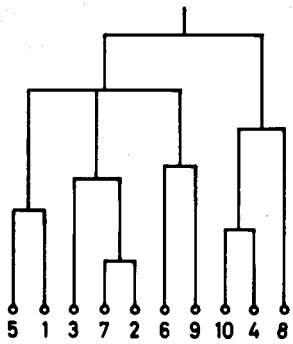

Abb. 1.5: Hierarchie mit $N = 10$ Objekten.

Wir erwähnen noch einige weitere Situationen, in denen ein Klassifikationsproblem auftritt:

- Patente oder Dokumente müssen nach (zunächst unbekannten und meist überschneidenden) Sachgruppen gegliedert werden, um rasches Wiederauffinden zu garantieren („information retrieval", Datenbanken).
- Eine Werbeagentur ermittelt homogene Verbrauchergruppen, um ihre Werbemittel gezielt einzusetzen.
- Die Einzelteile (Transistoren, Widerstände usw.) eines Computers müssen so in einzelne Baugruppen (gedruckte Platten) aufgeteilt werden, daß zwischen den Baugruppen möglichst wenig Verdrahtungen nötig sind.
- Eine Versicherung will ihre Auto-Haftpflicht-Versicherten aufgrund unfallspezifischer Merkmale (Wagentyp, Unfallhäufigkeit, Unfallkosten etc.) in homogen zusammengesetzte Tarifgruppen einteilen.
- Volkswirtschaftliche Daten müssen zwecks einfacherer Beschreibung zu Klassen zusammengefaßt werden (Aggregationsproblem).

c) Die Datenmatrix; quantitative und qualitative Merkmale

Ein Klassifikationsproblem kann nur dann sinnvoll gelöst werden, wenn über die Eigenschaften der zu klassifizierenden Objekte hinreichende Information vorliegt. Wir wollen in diesem und dem folgenden Abschnitt angeben, welcher Art die Daten sind, von denen man bei der automatischen Klassifikation ausgeht. Die hierbei eingeführten Begriffe sind grundlegend für alle weiteren Betrachtungen.

Ausgangspunkt unserer Überlegungen sind N Objekte $0_1, \ldots, 0_N$ (z. B. N Patienten), deren Gesamtheit wir als *Objektmenge*

$$S = \{0_1, \ldots, 0_N\} \tag{1.1}$$

bezeichnen; oft schreiben wir kürzer $S = \{1, \ldots, N\}$. Gesucht ist generell eine Klassifikation von S derart, daß jede Klasse möglichst homogen ist und nur Objekte mit ähnlichen Eigenschaften enthält.

Die Beispiele 1.1, 1.2 und 1.3 zeigen, daß die zur Klassifikation relevanten Eigenschaften der Objekte häufig durch gewisse Merkmale $M_1, \ldots, M_p$ erfaßt werden können, bei Patienten etwa durch Temperatur, Blutdruck usw. Durch Befragung, Test oder Experiment wird dann festgestellt, wie sich jedes Objekt 0_k $(k = 1, \ldots, N)$ bzgl. aller p Merkmale M_i verhält $(i = 1, \ldots, p)$. Wir setzen immer voraus, daß sich dieses Verhalten numerisch durch eine reelle Zahl x_{ki} beschreiben läßt.

Als Grundlage zur Lösung des Klassifikationsproblems ergibt sich dann eine $N \times p$-Matrix $X = (x_{ki})$, die wir als *Datenmatrix* bezeichnen und die nach Art von Abb. 1.6 in einer Tabelle dargestellt werden kann („Rohdaten"). Die Matrix X enthält in der k-ten Zeile den p-stelligen Vektor

Merkmale / Objekte	M_1	M_2	...	M_p	
0_1	x_{11}	x_{12}	...	x_{1p}	x'_1
0_2	x_{21}	x_{22}	...	x_{2p}	x'_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
0_N	x_{N1}	x_{N2}	...	x_{Np}	x'_N

Abb. 1.6: Die Datenmatrix $X = (x_{ki})$.

$$x_k = \begin{pmatrix} x_{k1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_{kp} \end{pmatrix} \qquad k = 1, \ldots, N. \tag{1.2}$$

Dieser Vektor beschreibt die Eigenschaften des Objekts 0_k und ist insofern ein Repräsentant für 0_k. Deshalb schreiben wir gelegentlich auch $S = \{x_1, \ldots, x_N\}$ anstelle von (1.1) und sprechen von einer Klassifikation der Vektoren $x_1, \ldots, x_N$.

Im folgenden bezeichne $\mathfrak{X}_i$ den Wertebereich der N Beobachtungen $x_{1i}, \ldots, x_{Ni}$ des i-ten Merkmals M_i ($i = 1, \ldots, p$). Das kartesische Produkt $\mathfrak{X} = \mathfrak{X}_1 \times \mathfrak{X}_2 \times \ldots \times \mathfrak{X}_p$ heiße der *Merkmalsraum.* Es ist dann $x_k \in \mathfrak{X}$ für alle k.

Es ist einleuchtend, daß jedes vernünftige Klassifikationsverfahren die Form der Daten x_{ki} in angemessener Weise berücksichtigen muß. Je nach dem Wertebereich $\mathfrak{X}_i$ der Zahlen $x_{1i}, \ldots, x_{Ni}$ unterscheiden wir quantitative und qualitative Merkmale:

Ein Merkmal M_i heißt *quantitativ,* wenn die zugehörigen Zahlen x_{ki} ($k = 1, \ldots, N$) prinzipiell jeden Wert eines (endlichen oder unendlichen) Intervalls annehmen dürfen (Abb. 1.1); Beispiele sind etwa Körpergröße, Lebensdauer, Temperatur. Ohne Beschränkung der Allgemeinheit wählen wir in diesem Fall für $\mathfrak{X}_i$ immer die Menge R aller reellen Zahlen: $\mathfrak{X}_i = R = \{x \mid -\infty < x < \infty\}$ und $\mathfrak{X} = R^p$.[1])

Ein Merkmal M_i heißt *qualitativ,* wenn das Verhalten jedes Objekts bzgl. dieses Merkmals durch eine endliche (i. a. kleine) Zahl von Alternativen oder Zuständen beschrieben wird; die Anzahl dieser Zustände werde mit m_i bezeichnet ($m_i \geqslant 2$). Beispiele sind: Farben (rot–grün–blau–gelb; $m_i = 4$), Graduierungen (schwach–mittel–stark; $m_i = 3$) oder Monatsangaben (Januar, ..., Dezember; $m_i = 12$).

Es ist üblich, die m_i Alternativen von M_i durch die Zahlzeichen „0", „1", ..., „m_i–1" zu kodieren: Wenn beim Objekt 0_k für das Merkmal M_i die Alternative „ν" eintritt ($0 \leqslant \nu \leqslant m_i - 1$), so schreibt man $x_{ki} = \nu$[2]); es ist dann

[1]) Praktisch können wegen der begrenzten Meßgenauigkeit immer nur endlich viele Werte als Meßergebnis auftreten; die Betrachtung quantitativer Daten stellt jedoch eine mathematisch zweckmäßige Idealisierung dar.

[2]) Statt dieser Zahlzeichen wären auch Buchstaben oder sonstige Markierungen denkbar (z.B. $a_1, a_2, \ldots, a_{m_1}$ für Merkmal M_1; $b_1, b_2, \ldots, b_{m_2}$ für Merkmal M_2 usw.).

$$\mathfrak{X}_i = \{0, 1, \ldots, m_i - 1\}.$$

Die Alternativen eines qualitativen Merkmals können geordnet oder ungeordnet sein (§§ 4.b, 5.a).

Ein Merkmal M_i, bei dem speziell nur $m_i = 2$ Alternativen (0 oder 1) eintreten können, heißt ein *binäres* Merkmal (Abb. 1.4). Die Fälle $x_{ki} = 1$ bzw. $x_{ki} = 0$ bezeichnen dann dichotome Angaben wie ja/nein, vorhanden/abwesend, Erfolg/Mißerfolg, männlich/weiblich usw. Der Einfachheit halber interpretieren wir $x_{ki} = 1$ ($x_{ki} = 0$) immer durch: „Merkmal M_i ist bei 0_k vorhanden (abwesend)".

Ein qualitatives Merkmal mit mehr als zwei Alternativen ($m_i > 2$) heißt auch *mehrstufig*. Allgemein sprechen wir von quantitativen (qualitativen, binären) Daten, wenn alle p Merkmale quantitativ (qualitativ, binär) sind. In der Praxis kommen quantitative und qualitative Merkmale oft gleichzeitig vor („gemischte" Merkmale).

d) Ausgangsdaten: Ähnlichkeiten, Relationen

Die Erstellung einer $N \times p$-Datenmatrix (x_{ki}) ist die gebräuchlichste Methode, um die zur Klassifikation benötigte Information zu gewinnen, und es gibt zahlreiche Verfahren, die beim Aufbau der gesuchten Gruppierung unmittelbar die Rohdaten x_{ki} verwenden.

Eine Reihe von Klassifikationsmethoden benutzt indessen die $N \times p$-Matrix (x_{ki}) nur indirekt: Da man nämlich von einer vernünftigen Klassifikation erwartet, daß „ähnliche Objekte der gleichen, unähnliche Objekte aber verschiedenen Gruppen angehören", liegt es nahe,

- zunächst in einem ersten Schritt die Ähnlichkeit aller Objektpaare 0_j, 0_k zu messen, etwa durch Zahlen s_{jk} $(1 \leqslant j, k \leqslant N)$, und dann
- in einem zweiten Schritt diese Ähnlichkeiten s_{jk} (und nur diese) zur Konstruktion der gesuchten Gruppierung zu verwenden.

Tatsächlich gehen viele Klassifikationsverfahren auf diese Weise vor. Da hier zum Aufbau der Gruppierung lediglich eine Ähnlichkeitsmatrix (s_{jk}) benutzt wird, sprechen wir von *Ähnlichkeitsmethoden.*

Es ist oft bequemer, anstelle der Ähnlichkeit zweier Objekte 0_j, 0_k deren Unähnlichkeit (Distanz) anzugeben. Wir führen hierfür die Maßzahl d_{jk} ein $(1 \leqslant j, k \leqslant N)$ und bezeichnen die $N \times N$-Matrix (d_{jk}) als die Distanzmatrix der Objekte. Klassifikationsverfahren, die von einer Ähnlichkeitsmatrix (s_{jk}) ausgehen, lassen sich meist ebensogut mit einer Distanzmatrix (d_{jk}) formu-

lieren und umgekehrt; wir geben später immer nur eine der beiden Versionen an.

Distanz- und Ähnlichkeitsmaße werden ausführlich in Kap. 1 behandelt. Dort werden wir zeigen, wie solche Maße mit Hilfe der Rohdaten (x_{ki}) berechnet werden können. Das folgende Beispiel lehrt jedoch, daß sich Ähnlichkeiten und Distanzen auch direkt – d.h. ohne den Umweg über eine Datenmatrix (x_{ki}) – aus der praktisch behandelten Fragestellung ergeben können.

Beispiel 1.4: Aus verschiedenen Gründne muß jeder Militärangehörige von Zeit zu Zeit seinen Standort wechseln, wobei er an seinem neuen Arbeitsplatz i. a. eine etwas andere Spezialtätigkeit ausüben muß wie zuvor. Hierdurch entstehen lästige Einarbeitungszeiten, deren Dauer d_{jk} (bei Wechsel von Tätigkeit j zur Tätigkeit k) aus Erfahrung bekannt ist, und die man „im Mittel" möglichst klein halten will. Man wird deshalb die einzelnen Tätigkeiten (= N Objekte) in Gruppen zusammenfassen und praktisch nur noch Wechsel innerhalb der gleichen Tätigkeitsgruppe anstreben. Man hat dann das Problem, die N Tätigkeiten so zu gruppieren, daß innerhalb jeder Klasse die einzelnen Tätigkeiten möglichst ähnlich sind. Offenbar sind hier die Einarbeitungszeiten d_{jk} ein natürliches Maß für die Unähnlichkeit zweier Spezialtätigkeiten.

Zur Formulierung und Lösung eines Klassifikationsproblems ist es nicht immer nötig, die Ähnlichkeit s_{jk} zweier Objekte numerisch exakt und eindeutig auf einer Meßskala (etwa von 0 bis 1) festzulegen. Wir werden später verschiedene Gruppierungsverfahren angeben, bei denen es auf den Wert der Zahlen s_{jk} gar nicht ankommt, sondern ausschließlich auf deren Rangfolge: Man stellt dann für jedes Objektpaar $\{k, l\}$ fest, ob es ähnlicher, unähnlicher oder gleich ähnlich ist wie jedes andere Objektpaar $\{i, j\}$. Wir schreiben hierfür $\{k, l\} \prec \{i, j\}$, $\{k, l\} \succ \{i, j\}$ bzw. $\{k, l\} \doteq \{i, j\}$ und sprechen von einer *Relation* $\prec$ auf der Menge aller $\binom{N}{2}$ Objektpaare (Abb. 2.1). Relationen sind oft die einzige, zur Lösung eines Gruppierungsproblems verfügbare (bzw. verwendete) Information.

Bemerkung 1.1: Drei Personen, die sich paarweise gut vertragen (und insofern paarweise „ähnlich" sind), brauchen zu dritt keineswegs zu harmonieren. Deshalb kann die Bildung einer möglichst harmonischen Personengruppe i. a. nicht dadurch erfolgen, daß nur Ähnlichkeiten (Verträglichkeit) zwischen je zwei Personen betrachtet werden. – Dies zeigt, daß bei manchen Klassifikationsproblemen die durch eine Ähnlichkeitsmatrix (s_{jk}) vermittelte Information zu einer sinnvollen Lösung nicht ausreicht. In solchen Fällen müssen Ähnlichkeiten „höherer Ordnung" betrachtet werden, die zwischen je drei oder mehr Objekten definiert sind. Wir gehen hierauf nicht weiter ein.

e) Vorläufige Präzisierung des Klassifikationsproblems; Bemerkungen

Mit Hilfe der oben eingeführten Begriffe läßt sich die in diesem Buch behandelte Fragestellung vorläufig so formulieren:

Zu N Objekten $0_1, \ldots, 0_N$ sei eine Datenmatrix (x_{ki}), eine Ähnlichkeitsmatrix (s_{jk}), eine Distanzmatrix (d_{jk}) oder eine Relation $\prec$ bekannt; diese charakterisiert die Ähnlichkeitsstruktur der Objektmenge $S = \{0_1, \ldots, 0_N\}$.

Gesucht ist eine *Klassifikation* $\mathfrak{A} = (A_1, A_2, \ldots)$ von S, d.h. ein System $\mathfrak{A}$ von Teilmengen $A_1, A_2, \ldots$ von S (Klassen, Gruppen), welche die Ähnlichkeitsstruktur der Objekte möglichst gut wiedergibt und eine hinreichende Datenreduktion erlaubt. Diese Forderung ist verbal meist so interpretierbar, daß die Objekte in jeder Klasse A_i möglichst große Ähnlichkeit besitzen sollen (Homogenität) und daß verschiedene Klassen leicht unterscheidbar sind (Separation).

Eine mathematisch präzise Formulierung des Problems, die erst in späteren Kapiteln erfolgen kann, muß sowohl die Art der vorgegebenen Daten als auch die Struktur der gewünschten (vermuteten) Gruppierung berücksichtigen. Die Beispiele 1.1, 1.2 und 1.3 machen deutlich, daß es dabei drei wichtige, strukturell verschiedene Klassifikationsarten gibt, die wir getrennt behandeln werden:

Disjunkte Klassifikation: Die Klassen $A_1, A_2, \ldots$ von $\mathfrak{A}$ dürfen sich nicht überschneiden (Abb. 1.2)

Nichtdisjunkte Klassifikation: Die Klassen $A_1, A_2, \ldots$ von $\mathfrak{A}$ dürfen sich (beliebig oder in begrenztem Umfang) überschneiden (Abb. 1.3)

Hierarchische Klassifikation: Die Klassen $A_1, A_2, \ldots$ von $\mathfrak{A}$ sind einander nach Art eines Stammbaumes über- bzw. untergeordnet (Abb. 1.4).

Außerdem unterscheidet man *exhaustive* und *nicht exhaustive* Klassifikationen: Bei den ersteren werden alle, bei den letzteren nur die „wichtigsten" Objekte von S in die Gruppierung einbezogen (Abb. 1.2 und 1.3).

Die später beschriebenen Klassifikationsverfahren lassen sich nach diesen und auch anderen Gesichtspunkten ordnen. Wir erwähnen insbesondere die Unterscheidung zwischen stochastischen und deterministischen Modellen und gehen dabei von einer Datenmatrix $X = (x_{ki})$ aus:

– Bei einem *stochastischen* Modell werden die Zahlen x_{ki} als Realisierung von Zufallsgrößen angesehen. Das ist z. B. dann sinnvoll, wenn innerhalb jeder gesuchten Klasse von vornherein eine natürliche Variation zu erwarten ist, wenn die Zahlen x_{ki} mit Meßfehlern behaftet sind oder wenn die Objekte $0_1, \ldots, 0_N$ als Zufallsstichprobe aus einer größeren Gesamtheit herausgegriffen sind. Stochastische Modelle führen zu einer statistischen Behandlung des Klassifikationsproblems und erlauben – zumindest prinzipiell – zu testen, ob die Objektmenge S überhaupt eine Gruppenstruktur aufweist (Kap. 3).

– Bei *deterministischen* Modellen betrachtet man $\{0_1, \ldots, 0_N\}$ als eine feste Objektmenge, deren Eigenschaften x_{ki} zwar von Objekt zu Objekt variieren, die jedoch (wie etwa die Zylinderzahl eines Motors) keinen Zufallsschwankungen unterworfen sind. Wahrscheinlichkeitsaussagen (Signifikanztests o.ä.) sind dann offenbar sinnlos.

Bemerkung 1.2: Die automatische Klassifikation stellt eine Weiterentwicklung der sog. Diskriminanzanalyse dar: Bei einem Diskriminationsproblem liegen mehrere, prinzipiell bekannte Klassen (Populationen) vor, und es soll ein vorgegebenes Objekt aufgrund seiner Merkmale in eine dieser Klassen eingeordnet werden (§ 26.a). Demgegenüber ist es die Aufgabe der automatischen Klassifikation, solche Klassen zu entdecken und zu lokalisieren. Da beide Probleme offenbar eng zusammenhängen, wird der Begriff „Klassifikation" in der Literatur oft einheitlich für beide Fragestellungen verwendet.

Bemerkung 1.3: Bei Kenntnis einer $N \times p$-Datenmatrix (x_{ki}) könnte ein Laie auf recht einfache Weise eine Klassifikation erstellen, indem er die N Objekte zuerst nach dem ersten Merkmal M_1 sortiert, dann nach dem zweiten, dritten usw. Dieses Verfahren ist sinnvoll, sofern jede in S vorhandene, „natürliche" Objektklasse durch eine bestimmte Merkmalskombination eindeutig charakterisiert wäre („monothetische" Klassen). Abgesehen davon, daß zu Beginn eines Klassifikationsverfahrens die charakterisierenden Merkmalskombinationen noch nicht bekannt sind, ist die Annahme monothetischer Klassen unrealistisch: Bei den meisten praktisch vorkommenden Gruppierungen ist eine Klasse eher dadurch gekennzeichnet, daß ihre Objekte zwar „relativ viele", aber nicht alle Merkmale einer gewissen Kombination aufweisen oder daß sie zumindest mehr Merkmale gemeinsam besitzen als mit Angehörigen anderer Klassen („polythetische" Klassen). Deshalb werden bei den Verfahren der automatischen Klassifikation die p Merkmale nicht wie oben einzeln und der Reihe nach, sondern gleichzeitig und in symmetrischer Weise berücksichtigt (Ausnahme: § 41.d).