G. Rohwer, C. Dudel

Wintersemester 2010/11

Methoden der Datenrepräsentation und Klassifikation

Kapitel 1: Einleitung

1 Einleitung

In diesem einleitenden Kapitel wird zunächst erläutert, was wir unter „Datenrepräsentation“ verstehen, und es werden die Beispiele angegeben, die wir zur Illustration verwenden. Dann wird im zweiten Abschnitt der formale Begriffsrahmen (statistische und relationale Variablen, Abstandsfunktionen) eingeführt. Im dritten Abschnitt erläutern wir Klassifikationsverfahren, die von Daten ausgehen und deren Darstellung dienen, und unterscheiden sie von Verfahren zur Konstruktion von Typologien.

1.1 Datenrepräsentation

1. Aufgaben der Datenrepräsentation

In einer allgemeinen Formulierung kann man sagen, dass Verfahren der Datenrepräsentation die Aufgabe haben, eine Menge gegebener Daten übersichtlich und informativ darzustellen. In einer etwas engeren Bedeutung geht es um Verfahren, mit denen man Daten „anschaulich“ machen, also bildliche Darstellungen erzeugen kann. Wir sprechen dann von *graphischer* oder *bildlicher* Datenrepräsentation.

Als Ausgangspunkt kommen sowohl statistische als auch relationale Daten in Betracht. Es gibt keine strenge Unterscheidung zwischen diesen beiden Arten von Daten; denn oft können statistische Daten durch (konstruierte) Relationen ergänzt werden, und andererseits gibt es bei relationalen Daten oft zusätzliche statistische Informationen über die erfassten Objekte. Wir werden uns deshalb in diesem Text nicht nur mit statistischen Daten im engeren Sinne, sondern auch mit relationalen Daten beschäftigen.

2. Ansätze der Datenrepräsentation

Die meisten Methoden zur Darstellung und Analyse statistischer Daten gehen von Häufigkeitsverteilungen aus. Hier können auch Methoden der Datenrepräsentation ansetzen. Einfache Methoden werden bereits in Einführungen in die Statistik vermittelt; zum Beispiel Darstellungen durch Häufigkeitsfunktionen (Dichtefunktionen) und Streuungsdiagramme.

Die einfachen Verfahren der Datenrepräsentation können bei Daten mit mehr als zwei Dimensionen nicht mehr ohne weiteres verwendet werden. Ein naheliegender Ausweg besteht darin, die Objekte, für die die Daten gegeben sind, zunächst zu klassifizieren. Sobald das geschehen ist, kann man die Daten durch Häufigkeitsverteilungen für die zuvor konstruierten Klassen repräsentieren.

Wenn eine Ähnlichkeits- oder Abstandsmatrix gegeben ist oder aus den Daten konstruiert werden kann, kann noch ein anderer Weg zur Datenrepräsentation verfolgt werden. Anstelle einer sofortigen Einteilung der Objekte in Klassen (und einer nachfolgenden Darstellung ihrer Häufigkeiten) kann man die Aufgabe darin sehen, die Ähnlichkeitsstruktur der Objekte darzustellen. Dafür gibt es hauptsächlich zwei Ansätze. Ein Ansatz besteht darin, die vorgegebenen Abstände durch sichtbare Abstände in einer bildlichen Darstellung zu repräsentieren. Diese Idee wird mit Methoden der multidimensionalen Skalierung verfolgt. Ein anderer Ansatz verfolgt die Idee, Ähnlichkeitsstrukturen durch Graphen zu repräsentieren.

In diesem Buch besprechen wir in erster Linie Methoden, die eine vorgängig konstruierte Ähnlichkeits- oder Abstandsmatrix voraussetzen. Diese Methoden haben den Vorteil, dass der Anwender die darzustellenden Relationen auf der Grundlage theoretischer Überlegungen bestimmen kann. Daneben gibt es auch Methoden, die automatisch bestimmte Abstandskonstruktionen implizieren; zum Beispiel die multidimensionale Skalierung mit Hauptkomponenten oder die Verfahren der Korrespondenzanalyse. Einige dieser Verfahren, die spezielle Methoden der linearen Algebra verwenden, besprechen wir im Anhang B.

3. Beispiele

Zur Illustration der Methoden verwenden wir in erster Linie Daten, die wir in der Literatur gefunden haben oder die aus Forschungszusammenhängen stammen, an denen wir teilgenommen haben. Hier ist eine kurze Übersicht:

- Ergebnisse einer Statistikklausur mit 46 Teilnehmern.

- Verteilungen von Männern und Frauen auf sechs Berufsgruppen für acht Länder.
- Nach Geschlecht und Geburtskohorten differenzierte Verteilungen von Schulabschlüssen.
- Archäologische Daten über Typen von Bronzegefäßen.
- Daten über Dominanzbeziehungen zwischen 20 Schimpansen.
- Sequenzdaten über Bestseller-Listen.
- Sequenzdaten zur beruflichen Mobilität.

1.2 Variablen, Abstände, Ähnlichkeiten

1. Statistische und relationale Variablen

Als formalen Rahmen verwenden wir statistische und relationale Variablen. Eine *statistische Variable* hat die Form

$$X : \Omega \longrightarrow \mathcal{X} \tag{1.1}$$

Ω repräsentiert eine endliche Menge von Objekten (irgendeiner Art, es kann sich auch um abstrakte Objekte oder um Situationen handeln) und wird als *Referenzmenge* der Variablen bezeichnet. Die statistische Variable X ordnet jedem Element $\omega \in \Omega$ genau einen Merkmalswert $X(\omega)$ aus dem Merkmalsraum $\mathcal{X}$ zu (der als eine Charakterisierung von ω interpretiert werden kann). Es wird angenommen, dass Merkmalswerte durch Zahlen repräsentiert werden, also $\mathcal{X}$ als eine Teilmenge der reellen Zahlen aufgefasst werden kann. – Beispielsweise steht $\Omega := \{\omega_1, \ldots, \omega_{10}\}$ für eine Menge von 10 Personen, und $\mathcal{X} := \{0, 1, \ldots, 100\}$ ist ein Merkmalsraum für das Alter (in vollendeten Jahren). Eine statistische Variable X könnte dann der Person ω_3 das Alter $X(\omega_3) = 43$ zuordnen.

Um Beziehungen zu erfassen, werden relationale Variablen verwendet. Eine (unimodale) *relationale Variable* hat die Form

$$R : \Omega \times \Omega \longrightarrow \mathcal{R} \tag{1.2}$$

Jeweils zwei Elementen der Referenzmenge Ω, etwa ω' und ω'', wird ein Wert $R(\omega', \omega'')$ in einem Merkmalsraum $\mathcal{R}$ zugeordnet, der als eine Information über das Vorhanden- oder Nichtvorhandensein einer Beziehung zwischen ω' und ω'' (und ggf. über die Art der Beziehung) interpretiert werden kann. Für den Merkmalsraum $\mathcal{R}$ wird wiederum angenommen, dass es eine numerische Repräsentation gibt. Im einfachsten Fall ist $\mathcal{R} = \{0, 1\}$, und die Werte geben an, ob eine Beziehung bestimmter Art besteht (1) oder nicht besteht (0). – Wird zum Beispiel für eine Menge von Personen $\Omega := \{\omega_1, \ldots, \omega_{10}\}$ erfasst, ob sie miteinander verwandt sind, würde $R(\omega_1, \omega_9) = 0$ bedeuten, dass ω_1 und ω_9 nicht verwandt sind.

2. Abstände und Ähnlichkeiten

Relationale Variablen liefern einen allgemeinen formalen Rahmen, um relationale Strukturen beliebiger Art zu repräsentieren. In diesem Text verwenden wir sie in erster Linie, um Ähnlichkeiten bzw. Unterschiede zwischen Objekten zu erfassen. Als Hilfsmittel dient der Begriff einer *Abstandsfunktion*. Eine Abstandsfunktion für eine beliebige Menge M ist eine Funktion

$$d : M \times M \longrightarrow \mathbf{R}$$

die für alle $m, m' \in M$ folgende drei Bedingungen erfüllt:

$$d(m, m') \geq 0 \tag{1.3}$$

$$d(m, m') = d(m', m) \tag{1.4}$$

$$d(m, m) = 0 \tag{1.5}$$

Das ist natürlich nur ein formaler Rahmen. In inhaltlicher Hinsicht wird angenommen, dass $d(m, m')$ als eine Größe interpretiert werden kann, die Aufschluss über den Abstand zwischen m und m' oder über das Ausmaß ihrer Unterschiedlichkeit gibt: Je größer $d(m, m')$, desto größer ist der Abstand zwischen oder die Unterschiedlichkeit oder Unähnlichkeit von m und m'. Und umgekehrt: Je kleiner $d(m, m')$, desto näher oder ähnlicher sind sich m und m'.

3. Abstandsfunktionen für Merkmalsräume

Abstandsfunktionen können für beliebige Mengen definiert werden. Wir verwenden sie in diesem Text hauptsächlich für Merkmalsräume und Objektmengen. Abstandsfunktionen für Merkmalsräume erlauben es, über Abstände zwischen Merkmalswerten zu sprechen. Zum Beispiel könnte man bei einem Merkmalsraum zur Erfassung von Einkommen eine Abstandsfunktion durch absolute Einkommensdifferenzen definieren.

Im allgemeinen gibt es unterschiedliche Möglichkeiten, um Abstandsfunktionen zu definieren. Man betrachte zur Illustration einen Merkmalsraum $\mathcal{X} = \{1, 2, 3, 4, 5\}$, durch den fünf Schulabschlüsse unterschieden werden: 1 = ohne Hauptschulabschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Eine Abstandsfunktion könnte beispielsweise durch $d(i, j) := |\, i - j \,|$ definiert werden. Der Abstand zwischen den Abschlüssen Fachhochschulreife und Abitur beträgt dann $d(4, 5) = |4 - 5| = 1$. Offenbar gibt es viele andere Möglichkeiten, um für dieses Beispiel Abstände zu definieren.

4. Abstandsfunktionen für Objektmengen

Mit Abstandsfunktionen für Objektmengen wird versucht, das Reden über Unterschiede zwischen bzw. Ähnlichkeiten von Objekten zu präzisieren. Es gibt hauptsächlich drei Ansätze.

a) Hat man eine statistische Variable $X : \Omega \longrightarrow \mathcal{X}$ und gibt es bereits eine Abstandsfunktion für den Merkmalsraum $\mathcal{X}$, etwa d, dann können durch $d(X(\omega), X(\omega'))$ Abstände zwischen den Objekten definiert werden. Wir sprechen dann von der durch d und X *induzierten Abstandsfunktion* für Ω.

b) In einigen Fällen können Werte einer Abstandsfunktion direkt (ohne Umweg über statistische Variablen) ermittelt werden. Beispielsweise kann man an Daten über soziale Netzwerke denken, durch die erfasst wird, ob und ggf. wie gut zwei Personen miteinander bekannt bzw. befreundet sind.

c) Schließlich können Werte von Abstandsfunktionen auch durch subjektive Ratingverfahren erzeugt werden. Ein Ansatz besteht darin, Menschen zu bitten, vorgegebene Objekte paarweise zu vergleichen und eine Meinung über den Grad ihrer Ähnlichkeit zu äußern.[1]

5. Abstands- bzw. Ähnlichkeitsmatrizen

Es ist oft hilfreich, die Werte einer Abstandsfunktion für eine Menge M mit n Elementen durch eine quadratische (n, n)-Matrix $\mathbf{D} = (d_{ij})$ darzustellen, wobei d_{ij} den Abstand zwischen den Elementen i und j angibt.[2] Wir sprechen dann von einer *Abstands-* oder *Ähnlichkeitsmatrix*. So lässt sich beispielsweise die in § 3 definierte Abstandsfunktion für Schulabschlüsse durch eine Abstandsmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix} \tag{1.6}$$

erfassen. Man sieht auch sogleich, wie man beliebige andere Abstandsmatrizen bilden kann. Erforderlich ist nur, dass es sich um symmetrische Matrizen mit nichtnegativen Elementen handelt und dass alle Elemente in der Hauptdiagonale gleich Null sind.

Es sei angemerkt, dass Abstandsmatrizen bei praktischen Anwendungen unvollständig sein können, d.h. es kann vorkommen, dass für einige Elemente der Matrix Werte fehlen. Wir verwenden die Konvention, fehlende Werte durch negative Zahlen zu kennzeichnen.

[1] Man vgl. hierzu etwa Green, Carmone und Smith (1989: 60ff.); Bortz und Döring (1995: 157f.).

[2] Wir folgen in diesem Text der Konvention, für Matrizen und Vektoren fettgedruckte Buchstaben zu verwenden. Vgl. den Anhang *Rechnen mit Matrizen* bei Rohwer und Pötter (2002a).

6. Metrische und nichtmetrische Abstände

Abstandsfunktionen für Merkmalsräume sind oft metrisch, d.h. sie genügen zusätzlich zu den Bedingungen (1.1) – (1.3) auch der sogenannten *Dreiecksungleichung*:

$$\text{Für alle } m, m', m'' \in M: \; d(m, m') + d(m', m'') \geq d(m, m'') \tag{1.7}$$

und der folgenden Eindeutigkeitsbedingung:

$$\text{Wenn } d(m, m') = 0, \text{ dann ist } m = m'. \tag{1.8}$$

Man spricht dann von einer *metrischen Abstandsfunktion* oder kurz von einer *Metrik*. Zum Beispiel ist die in § 3 definierte Abstandsfunktion für Schulabschlüsse ($d(i, j) = |\, i - j\,|$) metrisch. Eine Menge von Objekten, für die eine Metrik definiert ist, wird auch ein *metrischer Raum* genannt.

Dagegen sind Abstandsfunktionen für Objektmengen oft nicht metrisch. Werden sie durch eine Metrik für einen Merkmalsraum induziert, erfüllen sie zwar die Dreiecksungleichung, oft jedoch nicht die Eindeutigkeitsbedingung (1.8). Als Beispiel kann man sich vorstellen, dass Abstände zwischen Schulabgängern durch Abstände zwischen ihren Schulabschlüssen definiert werden; denn daraus, dass für zwei Personen $d(\omega', \omega'') = 0$ ist, kann nicht gefolgert werden, dass $\omega' = \omega''$ ist. In einigen Fällen, insbesondere wenn Abstandsfunktionen durch Ratingverfahren ermittelt werden, wird auch die Dreiecksungleichung (1.7) verletzt.[3]

1.3 Klassifikationen und Typologien

1. Klassifikationen für Objektmengen

Es sei Ω eine Objektmenge. Unter einer *Klassifikation für die Objektmenge* Ω verstehen wir eine Einteilung von Ω in zwei oder mehr paarweise disjunkte Teilmengen:

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_m \tag{1.9}$$

Eine Klassifikation für Ω kann also auch als eine Partition von Ω bezeichnet werden. Hier sind einige einfache Beispiele: Klassifikation einer Menge von Personen nach ihrem Geschlecht; Klassifikation einer Menge von Studierenden nach ihrem Studiengang; Klassifikation einer Menge von Haushalten durch eine Zuordnung zu Einkommensklassen.

[3] Wenn bei einer Abstandsmatrix $\mathbf{D} = (d_{ij})$ die Dreiecksungleichung verletzt wird, kann man jedoch eine neue Abstandsmatrix mit Koeffizienten $d_{ij} + c$ (für $i \neq j$) bilden, wobei

$$c := \max\{0, \max_{i,j,k}\{d_{ij} - d_{ik} - d_{kj}\}\}$$

ist. Dann ist für die modifizierte Abstandsmatrix die Dreiecksungleichung erfüllt. Zur praktischen Durchführung kann die TDA-Prozedur `dmet` verwendet werden.

2. Scharfe und unscharfe Klassifikationen

Klassifikationen in der in § 1 gegebenen Definition können auch als *scharfe Klassifikationen* bezeichnet werden, da jedes Element der Objektmenge genau einer Teilmenge zugeordnet wird. Lässt man dagegen zu, dass sich die Teilmengen, in die man eine Objektmenge Ω einteilt, überschneiden können, gelangt man zum Begriff einer *unscharfen Klassifikation*.[4] Zum Beispiel: Einteilung einer Menge von Personen in drei Gruppen: Personen, die ein Fahrrad besitzen; Personen, die ein Auto besitzen; Personen, die weder ein Fahrrad noch ein Auto besitzen. Interessante Beispiele gibt es bei ungenau erfassten Daten, die nur durch Teilmengen eines Merkmalsraums charakterisiert werden können.[5]

3. Darstellung durch statistische Variablen

Klassifikationen können durch statistische Variablen dargestellt werden. Ist eine scharfe Klassifikation der Form (1.9) gegeben, kann man sie durch eine statistische Variable

$$K : \Omega \longrightarrow \{1, \ldots, m\} \tag{1.10}$$

darstellen, wobei $K(\omega) = k$ gdw. $\omega \in \Omega_k$.[6] Handelt es sich um eine unscharfe Klassifikation, muss stattdessen eine Abbildung in die Potenzmenge, also

$$K^* : \Omega \longrightarrow \mathcal{P}(\{1, \ldots, m\}) \tag{1.11}$$

verwendet werden, wobei $K^*(\omega)$ die Indizes derjenigen Teilmengen von Ω angibt, denen ω zugerechnet werden kann.

4. Typologien für Merkmalsräume

Klassifikationen (in der hier verwendeten Definition) beziehen sich auf Mengen von Objekten. Die Objekte können beliebiger Art sein; insofern ist der Begriff sehr allgemein anwendbar. Eine besondere Terminologie bietet sich an, wenn es sich um einen Merkmalsraum handelt. Wir sprechen dann von einer (scharfen oder unscharfen) *Typologie* (für einen Merkmalsraum).[7] Der Merkmalsraum kann ein- oder mehrdimensional sein.[8]

[4] Es sei angemerkt, dass der Ausdruck 'unscharfe Klassifikation' auch noch in einer anderen Bedeutung, nämlich als Bezeichnung für Fuzzy-Klassifikationen, verwendet wird; man vgl. etwa Höppner, Klawonn und Kruse (1997).

[5] Dazu ausführlich Rohwer und Pötter (2001: Teil V).

[6] 'gdw.' wird als Abkürzung für 'genau dann wenn' verwendet.

[7] Die Idee, Typologien und Typenbegriffe nicht auf Objekte, sondern auf Merkmalsräume zu beziehen, findet man bei zahlreichen Autoren, die sich mit methodischen Fragen der Typenbildung beschäftigt haben; man vgl. etwa Hempel und Oppenheim (1936); Lazarsfeld (1937); Ziegler (1973).

[8] In der Literatur findet man gelegentlich die Auffassung, dass von „eigentlichen Typenbegriffen“ erst gesprochen werden sollte, wenn zwei oder mehr Merkmalsarten

Folgt man dieser Definition, betreffen Typologien Fragen der Begriffsbildung.[9] Weder setzen sie voraus noch implizieren sie Klassifikationen von (nicht-begrifflichen) Objekten. Ein Zusammenhang zwischen Typologien und Klassifikationen kann jedoch durch statistische Variablen hergestellt werden. Um das zu verdeutlichen, beziehen wir uns auf eine statistische Variable

$$X : \Omega \longrightarrow \mathcal{X} \tag{1.12}$$

Ist außerdem eine Typlogie $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_m$ gegeben, kann man mit ihrer Hilfe sofort eine Klassifikation von Ω bilden, indem man Teilmengen $X^{-1}(\mathcal{X}_j)$ verwendet. Wir nennen sie die durch die Typologie *induzierte Klassifikation.* Offenbar führt eine scharfe Typologie auch zu einer scharfen Klassifikation.

Ist umgekehrt eine Klassifikation einer Objektmenge Ω gegeben, etwa $\Omega = \Omega_1 \cup \cdots \cup \Omega_m$, kann man durch sie auch eine Typologie für den Merkmalsraum $\mathcal{X}$ definieren, indem man die Teilmengen $X(\Omega_j)$ verwendet. In diesem Fall erhält man jedoch durch eine scharfe Klassifikation nicht unbedingt eine scharfe Typologie und die Typologie wird von den jeweils verwendeten Objekten abhängig.

Es sei betont, dass in diesem Text mit dem Wort 'Typologie' stets eine (scharfe oder unscharfe) Klassifikation eines Merkmalsraums gemeint ist. Liegt beispielsweise nur eine Ordnungsrelation für einen Merkmalsraum vor, wird diese nicht als eine Typologie bezeichnet.[10]

5. Verfahren zur Bildung von Klassifikationen

Es gibt sehr viele Verfahren zur Erzeugung von Klassifikationen.[11] Zwei Vorgehensweisen können grundsätzlich unterschieden werden:

- Man kann von einer Typologie ausgehen und die jeweils gegebenen Objekte ihr entsprechend in (ggf. sich überlappende) Klassen einteilen.
- Man kann Objekte klassifizieren, ohne eine Typologie vorauszusetzen. Die hierfür verwendeten Verfahren beruhen meistens darauf, dass zunächst eine Abstandsfunktion für die Objekte angenommen oder konstruiert wird; dann wird die Idee verfolgt, ähnliche Objekte zu Klassen zusammenzufassen.[12]

kombiniert werden; man vgl. etwa Hempel und Oppenheim (1936: 65f.), Lazarsfeld (1937: 120), Stinchcombe (1968: 43).

[9] Dies betont auch Bailey (1994: 4f., 66), um die Konstruktion von Typologien von der Klassifikation von Objekten zu unterscheiden.

[10] Eine Ausdehnung des Typologiebegriffs, die auch Ordnungsrelationen einschließt, wurde von Hempel und Oppenheim (1936) vorgeschlagen.

[11] Aus der umfangreichen Literatur seien hier genannt: Anderberg (1973), Sneath und Sokal (1973); Bock (1974), Späth (1975), Lorr (1983), Jain und Dubes (1988), Kaufman und Rousseeuw (1990), Everitt (1993), Bacher (1994), Mirkin (1996).

In späteren Kapiteln beschäftigen wir uns hauptsächlich mit Klassifikationsverfahren, die von Abstandsfunktionen ausgehen.

6. Verwendungen des Typenbegriffs

Das Wort 'Typ' wird in unterschiedlichen Bedeutungen verwendet. Orientiert man sich an der oben gegebenen Definition von Typologien, kann man *eine* Bedeutung fixieren: *Typen* sind Elemente von Typologien. Bei dieser Definition sind Typen Begriffe (Konstellationen begrifflicher Merkmale) und müssen von Objekten (womit hier stets nicht-sprachliche Objekte gemeint sind) unterschieden werden. Objekte können durch Typen *charakterisiert* werden; und andererseits können Typen durch Objekte *illustriert* (oder *exemplifiziert*) werden.[13]

Von diesem Typbegriff, bei dem man im engeren Sinne auch von *klassifizierenden Typen* sprechen könnte, unterscheiden wir *Idealtypen*, worunter wir in diesem Text explizit festgelegte Werte in einem Merkmalsraum verstehen, die dem Zweck dienen, gewissermaßen als Ankerpunkte für eine relationale Ordnung des Merkmalsraums zu dienen.

[12] Manchmal wird diese Idee bereits verwendet, um „klassifizieren" zu erläutern; beispielsweise von Gordon (1987: 119): „Classification can be described as the activity of dividing a set of objects into a smaller number of classes in such a way that objects in the same class are similar to one another and dissimilar to objects in other classes."

[13] Anders als beispielsweise Sodeur (1974: 9) unterscheiden wir also auch Typen (als begriffliche Konstruktionen) und Cluster (als irgendwie abgegrenzte) Mengen von Objekten.