2 Schriften aus der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

Numerische Klassifikation (Cluster Analyse) anhand nominaler, ordinaler oder gemischter Merkmale

Theorie und Praxis mit zugehörigem Programm ORMIX auf CD

Prof. Dr. Friedrich Vogel und Dr. Rudolf Gardill

UNIVERSITY OF
BAMBERG
PRESS

Schriften aus der Fakultät Sozial- und
Wirtschaftswissenschaften der
Otto-Friedrich-Universität Bamberg 2

Schriften aus der Fakultät Sozial- und
Wirtschaftswissenschaften der
Otto-Friedrich-Universität Bamberg

Band 2

University of Bamberg Press 2010

Numerische Klassifikation (Cluster Analyse) anhand nominaler, ordinaler oder gemischter Merkmale

Theorie und Praxis mit zugehörigem Programm ORMIX auf CD

von Prof. Dr. Friedrich Vogel und Dr. Rudolf Gardill

University of Bamberg Press 2010

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.ddb.de/ abrufbar

Herstellung und Druck: docupoint GmbH Magdeburg
Umschlaggestaltung: Dezernat Kommunikation und Alumni der Otto-Friedrich-Universität Bamberg

ISSN: 1867-6197
ISBN: 978-3-923507-80-1 (Druck-Ausgabe)
eISBN: 978-3-923507-81-8 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus-2761

Numerische Klassifikation
(Cluster Analyse)
anhand nominaler, ordinaler oder gemischter Merkmale

I Theorie

II Praxis

Zusammenfassung

Numerische Klassifikation (oder Cluster Analyse) ist die Zuordnung einer Menge von Beobachtungen (Objekten) zu Teilmengen (Klassen oder Cluster), derart dass die Beobachtungen (Objekte), die einer Klasse angehören, in einem bestimmten Sinne einander ähnlich sind.

Diese Arbeit besteht aus zwei Teilen: Teil I "Theorie" und Teil II "Praxis".

Der erste Teil behandelt die theoretischen Grundlagen unseres neuen Klassifikationsprogramms ORMIX. Zunächst werden zwei Verfahren zur Bildung disjunkter Klassen erörtert: ein Austauschverfahren und ein hierarchisch-agglomeratives Verfahren. Dann werden Maße zur Messung der Güte eines Klassifikationsergebnisses im Detail diskutiert, insbesondere im Hinblick auf die Merkmalstypen: nominal, ordinal und metrisch. Die Gütefunktion für nominale und ordinale Merkmale basiert auf einem speziellen Streuungsmaß: der Entropie. Die Gütefunktion für metrische Merkmale basiert auf der Varianz. Das grundlegende Prinzip ist der Versuch der Minimierung der Streuung innerhalb der Klassen, so dass die Beobachtungen (Objekte) in derselben Klasse einander ähnlicher sind als die Beobachtungen (Objekte) verschiedener Klassen. Im Zusammenhang mit Problemen der Numerischen Klassifikation gibt es bei praktischen Anwendungen häufig gemischte Merkmale. Das heißt, die Objekte sind charakterisiert durch nominale und ordinale und metrische Merkmale. Um eine Gütefunktion für gemischte Merkmale zu konstruieren, ist zu beachten, dass die Gütefunktion für nominale und ordinale Merkmale auf der Entropie beruht, die Gütefunktion für metrische Merkmale aber auf der Varianz. Es ist nicht zulässig, diese Gütefunktionen zu addieren. Es kommt hinzu, dass die Varianz abhängt von den Skalen, auf denen die Merkmale gemessen werden. Es ist nicht möglich, metrische Merkmale derart zu skalieren, dass alle metrischen Merkmale im Prozess der Klassenbildung ein gleiches numerisches Gewicht haben; Standardisierung ist nur eine von vielen Mög-

lichkeiten, sie liefert aber keine Gleichgewichtung. Aber es ist zulässig, metrische Merkmale in ordinale Merkmale zu transformieren. Die Ordinalisierung metrischer Merkmale wird detailliert erklärt. Es wird gezeigt, dass - nach der Ordinalisierung der metrischen Merkmale - alle Merkmale im Prozess der Klassenbildung ein gleiches maximales numerisches Gewicht haben.

Der zweite Teil beschäftigt sich mit der Anwendung unseres Programms ORMIX, das nominale, ordinale, metrische Merkmale (nach Ordinalisierung) und gemischte Merkmale verarbeiten kann. Zuerst wird erklärt, wie das Programm von der CD installiert werden kann. Im Kapitel "Dateneingabe" werden die Konstruktion und das Einlesen der Datenmatrix im Detail erläutert. Dann wird gezeigt, wie Datentransformationen (beispielsweise metrische in ordinale Merkmale) durchgeführt werden können. Nach diesen Transformationen kann eine hierarchisch-agglomerative Klassifikation oder eine iterative Klassifikation durch einen linken Mausklick gestartet werden. Einige Beispieldateien finden sich auf der CD.

Die Bedienung des Programms ist einfach und meist selbsterklärend. Mit der (linken) Maustaste werden Berechnungen angestoßen und aus einer knappen Auflistung der Resultate ausführliche Detaildarstellungen ausgewählt. Ein Mausklick auf den Wert einer Gütefunktion öffnet ein Fenster mit dem Klassifikationsergebnis für die gewünschte Anzahl von Klassen und mit einer detaillierten Klassendiagnose. Die Klassifikationsergebnisse werden anschaulich in Tabellen zusammengefasst und als HTML-Seiten übersichtlich formatiert. Für die hierarchisch-agglomerative Klassifikation stehen zusätzlich Dendrogramme und ein Struktogramm zur Auswahl. Die rechte Maustaste führt zu Hilfeinformationen und ergänzenden auf den Kontext bezogenen Funktionen. Über die Zwischenablage von Windows können Daten mit anderen Programmen ausgetauscht werden.

Summary

Cluster analysis (or clustering) is the assignment of a set of observations (objects) into subsets (clusters) so that observations in the same cluster are similar in some sense.

This paper has two parts: part I "theory" and part II "practice".

The first part focusses on the theoretical foundations of our new cluster-analysis program called ORMIX. First two methods to construct disjoint clusters are discussed: a hill climbing (iterative partitioning) method and a agglomerative hierarchical clustering method. Then clustering criterions to measure the goodness of the resulting clusters are discussed in detail, in particular with respect to the type of variables: nominal, ordinal and metric. The criterion for nominal and ordinal variables is based on a special measure of dispersion: the entropy. The criterion for metric variables is based on the variance. As a basic principle it is tried to minimize the dispersion within the clusters, so that observations (objects) in the same cluster are similar in some sense. In the context with cluster problems there are in practice often mixed variables. That is the objects are characterized by nominal and ordinal and metric variables. To construct a clustering criterion for mixed variables ist must be noticed that the criterion for nominal and ordinal variables is based on the entropy, the criterion for metric variables is based on the variance. It is not admissible to summarise these criterions. Moreover the variance depends on the scales on which the variables are measured. It is not possible to scale metric variables in such a way that all metric variables have an equal numerical weight in the process of cluster building, standardisation is only one of many possibilities, it generates none equal weights for all variables. But it is allowed to transform metric variables in ordinal variables. The ordinalisation of metric variables is explained in detail. Therewith the procedure of cluster building with mixed variables and an admissible clustering criterion is explained in de-

tail. It is shown that - after ordinalisation of the metric variables - all variables in the process of cluster building have an equal maximum weight.

The second part focusses on the use of our program ORMIX which can work up with nominal, ordinal, metric variables (after ordinalisation) and mixed variables. First of all it is explained how to install the program from the CD. In chapter "data entry" the construction and the input of a data matrix, that is the basis of the clustering algorithm, into the program is treated in detail. Then it is shown how data transformations (for example metric in ordinal variables) can be realized. After these transformations a hill climbing (iterative partitioning) procedure and/or a agglomerative hierarchical clustering procedure can be performed with a left mouse click. Some examples are on the CD. The handling of the program is straightforward. It is steered with a left mouse click. A right mouse click provides many useful additional informations. The results of the analyses are clear illustrated. For the agglomerative hierarchical clustering procedure dendrograms and a struktogram can be displayed. A left mouse click on the value of an clustering criterion shows the cluster solution in the wished number of clusters and a detailed cluster diagnosis.

Numerische Klassifikation
(Cluster Analyse)
anhand nominaler, ordinaler und gemischter Merkmale

I Theorie

1 Einleitendes

Die numerische Klassifikation (oder Cluster-Analyse) - Sammelbegriff für eine Vielzahl unterschiedlichster mathematisch-statistischer und heuristischer Verfahren/Algorithmen zur Bildung "homogener" Klassen - gehört zur beschreibenden, mehrdimensionalen Statistik, zur (explorativen) Datenanalyse.

Verfahren der numerischen Klassifikation haben die Aufgabe, eine - im Allgemeinen große - Menge von n Einheiten, die durch m messbare Eigenschaften beschrieben sind, derart in eine meist kleine Anzahl von disjunkten Teilmengen zu zerlegen, dass die Einheiten, die derselben Teilmenge angehören, einander bezüglich der m Eigenschaften in einem bestimmten (numerischen) Sinne möglichst ähnlich/gleichartig sind, während gleichzeitig die Einheiten, die verschiedenen Teilmengen angehören, einander möglichst unähnlich/ungleichartig sind. Die Teilmengen heißen Klassen oder Cluster.

In der Regel sind die zu klassifizierenden Einheiten Merkmalsträger, also Personen, Haushalte, Produkte, Unternehmen, Tiere, Gemeinden, Kraftfahrzeug-Unfälle, Patienten, Länder, Berufe, Regionen, Zeitreihen, Aktien und dergleichen, die durch die Ausprägungen (mehrerer) bestimmter Merkmale wie zum Beispiel Alter, Beruf, Geschlecht, Anzahl der Kinder, Art der Krankheit, Haushaltsgröße, Haushaltseinkommen, Preise, Betriebsgröße, Anzahl der Beschäftigten, Ausgaben für Forschung und Entwicklung, Ausmaß der politischen Freiheit der Opposition eines Landes, Art der Regierungsübernahme, Anteil der Landbevölkerung, Anzahl der Fernsehgeräte im Haushalt, Proble-

me beim Treppensteigen, Bücher lesen und ähnliches beschrieben werden.[1)] Bei praktischen Anwendungen sind die Merkmale in aller Regel unterschiedlichen Typs, d.h. zur Beschreibung der Merkmalsträger werden nominale und/oder ordinale und/oder metrische Merkmale verwendet.

Bei der numerischen Klassifikation wird in der Regel von der im Allgemeinen nicht überprüfbaren Voraussetzung ausgegangen, dass in der Menge der Merkmalsträger und in Bezug auf die diese beschreibenden Merkmale eine Ordnung, eine Gliederung, ein Gefüge von Ähnlichkeiten/Unähnlichkeiten zwischen Merkmalsträgern existiert; kurz: dass mehr oder minder wohlseparierte und homogene (d.h. "natürliche") Klassen zwar vorhanden, aber nicht ohne weiteres identifizierbar sind. Es existiert eine "Abhängigkeitsstruktur", die sich (nur) durch Klassen beschreiben lässt.

Die gebildeten Klassen sollen im Hinblick auf eine bestimmte Zielsetzung brauchbar, nützlich oder zweckdienlich sein. Daher müsste eigentlich versucht werden, die Brauchbarkeit/Nützlichkeit eines Klassifikationsergebnisses zu optimieren. Nun ist aber Brauchbarkeit/Nützlichkeit nicht quantifizierbar, so dass an deren Stelle für die Steuerung der Klassenbildung numerische Ersatzkriterien verwendet werden müssen. Allerdings kann ein hinsichtlich eines solchen Ersatzkriteriums "optimales" Klassifikationsergebnis auch unbrauchbar sein.

Für die Konstruktion derartiger "Ersatzkriterien" gibt es mehrere Konzepte. Gemeinsam ist diesen Konzepten die Vorstellung, dass ein Klassifikationsergebnis um so brauchbarer/nützlicher ist, je ähnlicher/gleichartiger die Merkmalsträger innerhalb der einzelnen Klassen, d.h. je homogener die Klassen sind und - gleichzeitig - je unähnlicher die Merkmalsträger sind, die verschie-

1) Da die Klassifikation von Merkmalen, die ihrerseits durch bestimmte Eigenschaften gekennzeichnet sind, von eher untergeordneter Bedeutung ist, wird im folgenden nicht weiter darauf eingegangen.

denen Klassen angehören. Ersatzkriterium ist somit bei zahlreichen leistungsfähigen Klassifikationsverfahren die "Homogenität" der Klassen.

Da Merkmalsträger, die derselben Klasse angehören, einander bezüglich aller Klassifikationsmerkmale in einem bestimmten (numerischen) Sinne möglichst ähnlich sein sollen, ist es nahe liegend und zweckmäßig davon auszugehen, dass die Merkmalsträger einer Klasse einander dann ähnlich sind, dass die Klasse dann homogen ist, wenn die gemeinsame Streuung der m Klassifikationsmerkmale innerhalb der Klasse klein ist. Je größer die gemeinsame Streuung innerhalb der Klassen ist, desto unähnlicher sind die Merkmalsträger dieser Klassen. Ersatzkriterien zur Steuerung der Klassenbildung können somit auf der Grundlage geeigneter Streuungsmaße konstruiert werden.

2 Verfahren zur Bildung disjunkter Klassen

2.1 Vorbemerkung

Es ist nahe liegend zu versuchen, die in irgendeinem (numerischen) Sinne "optimale" Partition eines Datensatzes in K Klassen - zum Beispiel jene Partition, für die ein geeignetes Homogenitätsmaß den optimalen Wert annimmt - enumerativ zu ermitteln, indem für den gegebenen Datensatz (die Matrix $\underline{X}$) alle möglichen Partitionen in K Klassen erzeugt und die hinsichtlich der Homogenität der Klassen optimale Partition bestimmt wird. Eine solche Berechnung ist jedoch - von "kleinen" Datensätzen einmal abgesehen - nahezu unmöglich, weil die Anzahl der möglichen (und verschiedenen) Partitionen in K Klassen mit zunehmender Anzahl von Merkmalsträgern und Klassen über alle Grenzen wächst. Daraus ergibt sich für praktische Anwendungen in aller Regel die Notwendigkeit, die Anzahl der auf Optimalität zu überprüfenden Partitionen ganz erheblich zu reduzieren.

Die Reduktion der Menge der möglichen Partitionen auf eine ausreichend kleine Menge zu überprüfender Partitionen (und in dieser Hinsicht unterscheiden sich die gebräuchlichen Klassifikationsverfahren zum Teil wesentlich) erfolgt grundsätzlich in der Weise, dass

- zum einen die große Anzahl von Partitionen, von denen unterstellt werden kann, dass sie nicht optimal sind, nicht erzeugt und auf ihre Optimalität überprüft wird und
- zum anderen für die "optimale Klassenanzahl" K in Abhängigkeit vom Untersuchungsziel und der Anzahl der zu klassifizierenden Merkmalsträger N ein möglichst kleiner Bereich (beispielsweise von $K = 3$ Klassen bis $K = 8$ Klassen) vorgegeben wird.

Alle (für "größeres" n) gebräuchlichen Klassifikationsverfahren untersuchen nur einen sehr kleinen Teil der möglichen Partitionen und können daher - was mehr oder weniger bewusst in Kauf genommen wird - das (numerische) Optimum verfehlen, d.h. nur suboptimale oder lokal optimale Partitionen liefern. Anhaltspunkte für den Grad der Annäherung an das Optimum gibt es - im Allgemeinen - nicht.

Für praktische Anwendungen sind die daraus resultierenden Probleme im Allgemeinen nicht entscheidend, denn das numerische Optimum ist nur ein Ersatzkriterium für die größtmögliche Brauchbarkeit/Nützlichkeit einer Partition, und es ist keineswegs sicher, dass sich durch eine - auch ökonomisch aufwendige - Verbesserung eines lokalen Optimums eine entsprechende Zunahme an Brauchbarkeit/Nützlichkeit ergibt.

Bei praktischen Anwendungen haben sich zwei Verfahren besonders bewährt: ein so genanntes Austauschverfahren und ein hierarchisch-agglomeratives Verfahren. Beide Verfahren steuern die Konstruktion der Klassen mit Hilfe von Streuungsmaßen.

Wird die Menge der Merkmalsträger mit

$$N = \{N_1, N_2, ..., N_j, ..., N_n\}$$

bezeichnet, dann heißt jedes System

$$\{G_1, G_2, ..., G_k, ..., G_K\}$$

von nicht-leeren und paarweise verschiedenen Teilmengen (= Klassen) $G_k \subseteq N$ mit $\bigcup_{k=1}^{K} G_k = N$ eine "exhaustive" (vollständige) Klassifikation der Menge der Merkmalsträger.

2.2 Das Austauschverfahren

Sind die K Klassen paarweise disjunkt und gehört jeder Merkmalsträger genau einer Klasse an, dann heißt die disjunkte Klassifikation

$$P_K = \{G_1, G_2, ..., G_k, ..., G_K\}$$

eine Partition der Menge der Merkmalsträger in K Klassen.

Mit dem (iterativen) Austauschverfahren kann eine optimierte Partition P_K der Menge der Merkmalsträger N in eine vorgegebene Anzahl von Klassen K wie folgt gefunden werden.

Zunächst ist zur Steuerung der Klassenbildung eine der Problemstellung/Zielsetzung und dem Merkmalstyp (gegebenenfalls auch den Merkmalstypen) adäquate Gütefunktion:

$$g(\underline{X}, P_K)$$

zu bestimmen, die misst, wie gut die Partition P_K die Ähnlichkeitsstruktur der Daten $\underline{X}$ repräsentiert, oder anders formuliert, die angibt, wie homogen (im Mittel) die gebildeten Klassen sind.

Solche Gütefunktionen werden in der Regel minimiert, da sie im Allgemeinen mit Hilfe von Streuungsmaßen konstruiert werden. Im Folgenden wird nur diese Art von Gütefunktion behandelt.

Dann ist in einem zweiten Schritt eine so genannte Startpartition in K Klassen

$$P_K^0 = \left\{G_1^0, G_2^0, ..., G_K^0\right\}$$

vorzugeben oder zu erzeugen. Bei praktischen Anwendungen wird diese Startpartition (der Einfachheit halber) im Allgemeinen zufällig (mit Hilfe von Zufallszahlen) erzeugt, d.h. die Merkmalsträger werden zufällig jeweils einer und nur einer der vorgegebenen K Klassen zugeordnet.[2)] Die Art der Erzeugung der Startpartition hat keinen direkten Einfluss auf die Güte des Klassifikationsergebnisses. Das bedeutet, eine im Sinne der Gütefunktion "gute" Startpartition führt nicht zwangsläufig auch zu einem "guten" Klassifikationsergebnis.

In einem dritten Schritt wird nun versucht, die Startpartition P_K^0 zu verbessern.

Jeder Verbesserungsversuch wird mit der Gütefunktion $g(\underline{X}, P_K)$ gemessen. Dabei wird versucht, den Wert der Gütefunktion zu verringern, weil die (mittlere) Streuung innerhalb der Klassen sukzessive kleiner und somit die Partition "besser" werden. Jene Partition P_K^*, für die $g(\underline{X}, P_K^*)$ ein Extremum, im Allgemeinen das Minimum, annimmt, gilt als "optimale" Partition von N, beschrieben durch $\underline{X}$, in K Klassen und somit als Lösung des Klassifikationsproblems.

Die Verbesserung der Startpartition $P_K^0 = \left\{G_1^0, G_2^0, ..., G_K^0\right\}$, d.h. die Suche nach dem Minimum der Gütefunktion, wird wie folgt iterativ durchgeführt.

---

2) Für die Güte eines Klassifikationsergebnisses ist es nicht entscheidend, wie die Startpartition erzeugt wird. Auch eine im Sinne des gewählten Gütekriteriums "gute" Startpartition gewährleistet nicht, dass das Klassifikationsergebnis besser ist, als bei "schlechteren" Startpartitionen.

Beginnend mit $g\left(\underline{X}, P_K^0\right)$ und $j=1$ wird für jeden einzelnen Merkmalsträger ($j=1,2,...,n$) sukzessive geprüft, ob es im Hinblick auf eine Verringerung (Verbesserung) der Gütefunktion $g\left(\underline{X}, P_K\right)$ von Vorteil ist, diesen Merkmalsträger aus seiner Klasse zu entfernen und einer anderen Klasse zuzuordnen. Wenn eine Verringerung des Wertes der Gütefunktion möglich ist, wird der Merkmalsträger (der probeweise und vorläufig allen anderen Klassen zugeordnet wird) jener Klasse zugeordnet, die die größte Verkleinerung der Gütefunktion bewirkt. Dann wird für den Merkmalsträger $j=2$ überprüft, ob durch eine Verschiebung in eine andere Klasse die Gütefunktion verkleinert werden kann und so fort.

Wenn die Zuordnung aller n Merkmalsträger derart überprüft worden ist und wenn - was allerdings die Regel ist - Merkmalsträger anderen Klassen zugeordnet wurden, so gilt für diese (bessere) Partition P_K^1:

$$g\left(\underline{X}, P_K^1\right) < g\left(\underline{X}, P_K^0\right).$$

Dieses Vorgehen wird - beginnend wieder mit $j=1$ - so lange fortgesetzt, bis eine weitere Verkleinerung des Wertes der Gütefunktion - zumindest auf diese Weise - nicht mehr möglich ist, d.h. so lange, bis in einer Iterationsphase kein Merkmalsträger mehr einer anderen Klasse zugeordnet wird oder so lange bis die Verringerung des Wertes der Gütefunktion im Rahmen der Rechengenauigkeit nicht mehr berücksichtigt werden kann.

Es entsteht somit eine Folge von Partitionen:

$$P_K^0, P_K^1, ..., P_K^*$$

die im Sinne der Gütefunktion immer "besser" sind, für die also gilt:

$$g\left(\underline{X}, P_K^0\right) > g\left(\underline{X}, P_K^1\right) > \cdots > g\left(\underline{X}, P_K^*\right).$$

Die Partition

$$P_K^* = \left\{G_1^*, G_2^*, ..., G_K^*\right\}$$

gilt - relativ zur Startpartition P_K^0 - als "optimale" Partition von N beschrieben durch $\underline{X}$ in K vorgegebene Klassen.

So einfach und plausibel dieser Algorithmus zur Erzeugung einer "optimalen" Partition P_K^* auch ist, so vielfältig sind die Probleme im Detail, die sich mit seiner Anwendung ergeben.

Das Austauschverfahren liefert eine im (numerischen) Sinne der Gütefunktion $g(\underline{X}, P_K)$ "beste" Partition P_K^*. Die Frage, ob diese Partition auch die hinsichtlich des Untersuchungsziels brauchbarste/nützlichste ist, ist formal (numerisch) nicht zu beantworten.

Für das Austauschverfahren (und andere partitionierende Verfahren) ist die Anzahl der zu bildenden Klassen vorzugeben. Da den Daten einerseits im Allgemeinen keine Klassenstruktur in genau K Klassen aufgedrückt werden soll, andererseits aber Informationen über die optimale (im Sinne von "wahre" oder "natürliche") Anzahl der Klassen in aller Regel fehlen, ist das Problem einer zweckdienlichen, der Datenstruktur entsprechenden Festsetzung von K nicht eindeutig lösbar. Es ist daher keineswegs auszuschließen, dass die "optimale" Klassenanzahl verfehlt wird.

Bei praktischen Anwendungen war häufig eine hierarchisch-agglomerative "Vor-Klassifikation", die nur zu dem Zweck durchgeführt wurde, Informationen über die Anzahl der zu bildenden (natürlichen) Klassen zu gewinnen, besonders hilfreich (vgl. Abbildung 1: Dendrogramm).

Gelegentlich gibt es auch theoretische Anhaltspunkte für die Festlegung der Klassenanzahl. Für die Einteilung der Länder der Erde nach ihrem "Entwicklungsstand" kann z.B. der Standpunkt vertreten werden, dass es "unterentwickelte" und "hoch entwickelte" Länder sowie zumindest eine Klasse von Län-

dern (Schwellenländer) gibt, die sich hinsichtlich des Entwicklungsstands zwischen diesen beiden Typen befinden. Somit ist $K_{min} = 3$ eine untere Grenze der Klassenanzahl. Wenn man es für möglich und wahrscheinlich hält, dass zwischen den beiden extremen Typen von Ländern mehr als ein Typ von "Schwellenländern", d.h. mehr als eine Klasse existiert, dann könnte mit $K_{max} = 8$ (also sechs Typen von Schwellenländern) auch eine obere Grenze für K angegeben werden.

Auch wenn das Problem der Bestimmung der optimalen Klassenanzahl gelöst wäre, so wäre dennoch nicht gewährleistet, dass das Austauschverfahren die im numerischen Sinne von $g(\underline{X}, P_K)$ global optimale Partition P_K^{opt} findet, da ja - in der Regel - nicht alle möglichen Partitionen von N in K Klassen erzeugt und auf ihre Optimalität hin untersucht werden können. Es ist davon auszugehen, dass das Austauschverfahren nur eine suboptimale, d.h. "lokal optimale" Partition P_K^* liefert.

Ob die Partition P_K^* (nur) lokal optimal oder global optimal ist und - gegebenenfalls - wie gut das globale Optimum approximiert wurde, lässt sich im Allgemeinen nicht feststellen. Die Partition P_K^* ist (allenfalls) auch nur bezüglich des oben beschriebenen Austauschalgorithmus "optimal". Werden in der Iterationsphase jeweils nicht nur ein, sondern zwei oder mehrere Merkmalsträger gleichzeitig und vorläufig in die anderen Klassen transferiert, so können sich andere lokal optimale Partitionen ergeben, die das globale Optimum besser, aber auch schlechter approximieren.

Im Übrigen ist die Frage nach der "global optimalen" Partition bei praktischen Anwendungen nicht von überragender Bedeutung. Wenn die gesuchten Klassen "wohlsepariert" sind, dann werden sie auch von fast jedem Verfahren gefunden (vorausgesetzt, die wahre Klassenanzahl liegt zwischen K_{min} und K_{max}). Sind sie es nicht, berühren sich die Klassen oder gibt es Merkmalsträ-

ger, die im Anziehungsbereich von zwei oder mehreren Klassen liegen, dann ist das Konzept "natürlicher" Klassen zumindest in Frage gestellt, denn zwischen den beiden Extremen "homogene" Gesamtheit einerseits und natürliche Klassenstruktur andererseits ist jede Struktur denkbar; und jedes Klassifikationsverfahren wird den Daten in der Regel eine irgendwie geartete (verfahrensspezifische) Klassenstruktur aufdrücken.[3)] Dabei ist es praktisch nur von geringer Bedeutung für die Eigenschaften der einzelnen Klassen ob einzelne Merkmalsträger, die sich im Anziehungsbereich mehrerer Klassen befinden, der einen oder der anderen Klasse zugeordnet werden.

Zu bedenken ist auch, dass das Austauschverfahren ein "globales" Verfahren in dem Sinne ist, dass eine bestimmte Gütefunktion $g(\underline{X}, P_K)$ global - über alle K Klassen gleichzeitig - optimiert wird, so dass (im Allgemeinen) homogene und heterogene Klassen in einer Partition P_K^* gleichzeitig vorkommen. "Im Mittel" werden homogene Klassen erzeugt! Die Frage, wie stark die Klassenstruktur in einer Partition P_K^* ausgeprägt ist, kann bei der Klassendiagnose beantwortet werden.

2.3 Hierarchisch-agglomerative Verfahren

Es gibt eine ganze Reihe verschiedener hierarchisch-agglomerativer Verfahren. Die meisten haben allerdings zum Teil erhebliche Mängel. Sie konstruieren eine Hierarchie von Klassen wie folgt.

Ausgangspunkt der Klassenbildung sind n Klassen vom Umfang 1, also die n einzelnen Merkmalsträger. Auf jeder der insgesamt $r = n - 1$ Fusionsstufen werden sukzessive jeweils jene beiden Klassen vereinigt (fusioniert), die in einem bestimmten, noch näher zu definierenden Sinne einander am "ähnlichsten" sind, bis auf der $(n-1)$-ten Fusionsstufe alle n Merkmalsträger einer

3) Beispielsweise eine hierarchische Struktur.

Klasse angehören. Die Ähnlichkeit zweier Klassen wird mit bestimmten Gütefunktionen gemessen.

Aus dieser Vorgehensweise folgt allerdings ein nicht unwesentlicher Nachteil dieses Verfahrenstyps.

"Optimiert" wird nämlich sukzessive jede einzelne Fusion, das bedeutet, der Aufbau einer Hierarchie, nicht jedoch die Partition der r-ten Stufe P_K in $K = n - r$ Klassen. Optimiert wird auch nur durch sukzessive - in einem bestimmten Sinne optimale - Fusionen von genau zwei Klassen, nicht jedoch durch sukzessive Fusionen von mehr als zwei Klassen, was grundsätzlich möglich ist. Hinzu kommt, und das ist wesentlich, dass eine Fusion zweier Klassen in dem Sinne irreversibel ist, dass eine auf der r-ten Fusionsstufe fusionierte Klasse nicht wieder aus der Vereinigungsklasse entfernt werden kann, auch wenn es sich durch andere Analyseverfahren (z.B. ein iteratives Verfahren) herausstellen sollte, dass genau diese Klasse im Sinne des gewählten Gütekriteriums besser Teil einer anderen Klasse wäre.

Aus diesen Überlegungen folgt, dass hierarchisch-agglomerative Verfahren sowohl die "optimale" Hierarchie wie auch die optimale Partition der r-ten Stufe verfehlen können, weil auf den unteren Stufen der Hierarchie ein "falscher" Weg eingeschlagen werden kann.

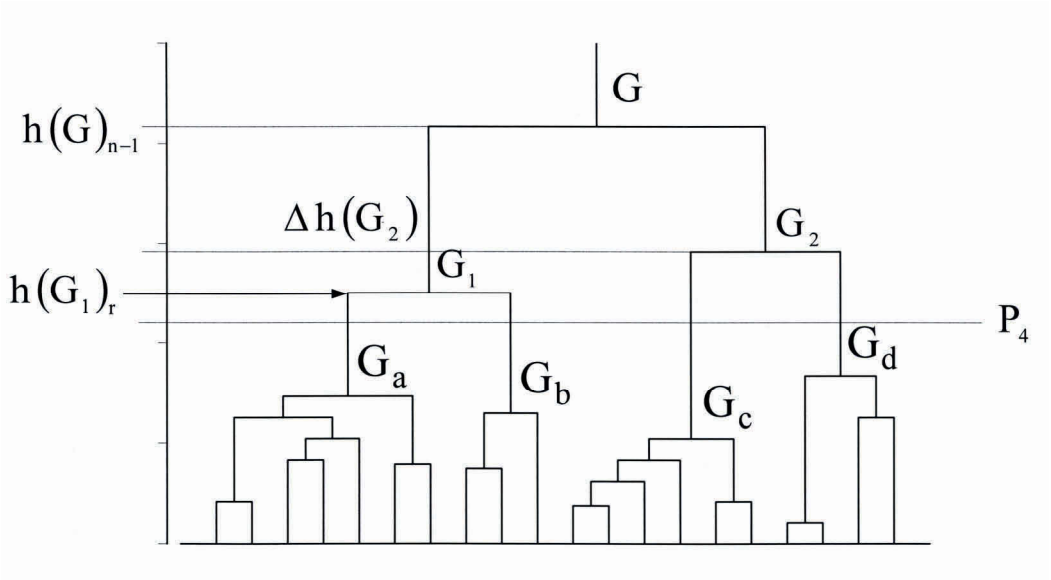
Das die so entstehende Hierarchie von Klassen repräsentierende Stammbaum-Schema, das Dendrogramm (vgl. Abbildung 1), wird im Allgemeinen durch eine Skala ergänzt, auf der

- die Homogenität oder komplementär die Heterogenität $h(G)_r$, der auf der r-ten Fusionsstufe gebildeten Klasse G oder
- in bestimmten Fällen die Homogenität der Partition der r-ten Stufe $(r = 1,2,...,n-1)$ oder

- der Heterogenitätszuwachs $\Delta h(G_\bullet) \geq 0$, der aus der Fusion zweier Klassen resultiert

abgetragen wird.

Abbildung 1: Dendrogramm



Die Klasse G_1 entsteht beispielsweise aus der Fusion der Klassen G_a und G_b. Die beiden zugehörigen Teil-Stammbäume werden auf dem Niveau $h(G_1)_r = h(G_a \cup G_b)$ verbunden.

Das Dendrogramm ist - bei nicht allzu großer Anzahl von Merkmalsträgern - ein sehr anschauliches Instrument zur Beurteilung von Klassifikationsergebnissen. Es veranschaulicht nicht nur die Hierarchie der Klassen als solche, sondern informiert auch über den Verlauf der Klassenbildung und die Daten-/Klassenstruktur.

So ist zunächst einmal abzulesen, welche Merkmalsträger paarweise einander am ähnlichsten sind, welche Merkmalsträger auf welcher Stufe einer bestimmten Klasse zugeordnet sind und natürlich auch, welche Klassen einander ähnlich und welche einander unähnlich sind. Es ist ferner abzulesen, auf

welchem "Homogenitätsniveau" die ersten "größeren" Klassen gebildet werden und wie homogen die einzelnen Klassen sind.

Auf jeder der $r = 1,2,...,n-1$ Fusionsstufen ergibt sich eine Partition P_K in $K = n - r$ Klassen, so dass die zur Datenmatrix $\underline{X}$ gehörende Hierarchie (Hi) auch durch die Folge von Partitionen

$$Hi(\underline{X}) = \{P_K\}_{K=1}^{n}$$

beschrieben werden kann.

Die Homogenität der Klassen $h(G)_r$ nimmt mit zunehmendem r, also mit abnehmender Anzahl an Klassen, im Allgemeinen ab (die Heterogenität nimmt zu), d.h. es gilt

$$h(G_\bullet) \leq h(G), \text{ falls } G_\bullet \subset G.$$

Beispielsweise ist

$$G = G_1 \cup G_2 \text{ und}$$

$$\Delta h(G_2) = h(G)_{n-1} - h(G_2)_{n-2} > 0.$$

Daraus folgt, dass aus der Homogenitätsdifferenz $\Delta h(G_\bullet) > 0$ zweier Klassen abgelesen werden kann, ob die Klasse $G_\bullet$ stabil, d.h. gut ausgeprägt und von den anderen Klassen separiert ist, dann ist $\Delta h(G_\bullet)$ groß - vgl. $\Delta h(G_2)$ - oder ob sie instabil, heterogen, noch nicht abgeschlossen ist, dann ist $\Delta h(G_\bullet)$ klein. Wenn der Datensatz gut - hierarchisch - strukturiert ist, dann gibt es mehrere stabile Klassen mit (relativ) großer Homogenitätsdifferenz.

Bei vielen praktischen Anwendungen[4)] ist die Hierarchie der Klassen selbst von geringem Interesse, da es fraglich ist, ob ein Datensatz durch eine hierar-

4) In der Biologie ist die hierarchische Klassifikation von Organismen (beispielsweise in Familie, Unterfamilie, Gattung, Art) von großer Bedeutung.

chische Klassenstruktur gekennzeichnet ist oder ob diese Struktur durch die Art der Klassenbildung nur den Daten aufgedrückt wurde.

Wird eine Partition des Datensatzes mit K disjunkten Klassen angestrebt, kann ein Dendrogramm Indizien für die optimale Klassenanzahl K liefern. Im obigen Dendrogramm könnte der Datensatz zum Beispiel in $K = 2$ oder auch $K = 4$ Klassen (vgl. P_4) zerlegt und gegebenenfalls mit dem Austauschverfahren verbessert werden.

Außerdem liefern bestimmte hierarchisch-agglomerative Verfahren nützliche Informationen über die Datenstruktur, beispielsweise über mehrdimensionale Ausreißer.

Die existierenden hierarchisch-agglomerativen Verfahren unterscheiden sich nur durch die Vorschrift, durch die Gütefunktion, nach welcher auf jeder Fusionsstufe die Un-/Ähnlichkeit zweier Klassen gemessen wird. Unterschiedliche Vorstellungen von der "Art der Homogenität" der zu bildenden Klassen werden durch diese Vorschrift realisiert.

2.4 Maße zur Messung der Güte eines Klassifikationsergebnisses

2.4.1 Vorbemerkung

Die leistungsfähigsten Klassifikationsverfahren steuern die Klassenbildung mit Hilfe von Gütefunktionen, die auf Streuungsmaßen beruhen. Dahinter steht die Vorstellung, dass die Merkmalsträger, die einer Klasse angehören, einander dann ähnlich/gleichartig sind, wenn die Streuung innerhalb der Klasse klein ist.

2.4.2 Merkmalstypen

Merkmalsträger werden durch Merkmale beschrieben. Merkmale sind messbare Eigenschaften eines Merkmalsträgers. Beim Messen der Eigenschaften werden den Merkmalsträgern unter Verwendung bestimmter Skalen - d.h. un-

ter Einhaltung bestimmter Regeln - Symbole oder Zahlen zugeordnet, die Merkmalsausprägungen heißen. Unter „Messen“ wird dabei - abweichend vom üblichen Sprachgebrauch - die Zuordnung von Symbolen (z.B. von Buchstaben) oder von Zahlen zu den Eigenschaften der Merkmalsträger verstanden.

Es ist üblich und zweckmäßig drei Merkmalstypen zu unterscheiden. Nach der der Messung zugrunde liegenden Skala werden die Merkmale in nominale, ordinale und metrische Merkmale eingeteilt.

2.4.2.1 Nominale Merkmale

Die einander ausschließenden Ausprägungen nominaler Merkmale werden auf Nominalskalen gemessen, d.h. es werden ihnen Symbole oder Nominalzahlen zugeordnet, mit denen nur die Gleichheit oder Ungleichheit von Merkmalsträgern hinsichtlich des betrachteten Merkmals festgestellt werden kann. Ein typisches Beispiel ist, z.B. für die Beschreibung von Personen, der Familienstand. Die Ausprägungen können verbal bezeichnet werden: ledig, verheiratet, verwitwet, geschieden, aber auch mit Buchstaben: a, b, c, d oder mit Nominalzahlen: 1, 2, 3, 4. Für die Numerische Klassifikation ist wesentlich, dass mit diesen Nominalzahlen nicht gerechnet werden darf. Sie dienen nur der Bezeichnung einer Eigenschaft. Nominale Merkmale sind invariant gegenüber eineindeutigen (umkehrbar eindeutigen) Transformationen (wie z.B. Umbenennungen oder Vertauschungen). Nominale Merkmale werden im Folgenden mit A oder B, deren Ausprägungen mit A_i $(i = 1,2,...,k)$ oder B_j $(j = 1,2,...,m)$ bezeichnet, dabei ist k bzw. m die Anzahl der jeweiligen Ausprägungen.

Nominale Merkmale mit nur zwei Ausprägungen heißen Alternativmerkmale, binäre Merkmale oder dichotome Merkmale. Den Ausprägungen werden - aus Gründen der Zweckmäßigkeit - häufig die Ziffern (Nominalzahlen) 0 und 1 zugeordnet.

2.4.2.2 Ordinale Merkmale

Die einander ausschließenden Ausprägungen ordinaler Merkmale werden auf Ordinalskalen gemessen, d.h. es werden ihnen Symbole oder (häufig) Ordinalzahlen zugeordnet, mit denen die Merkmalsträger in eine Rangordnung gebracht werden können. Für die Ausprägungen ist eine "größer - kleiner", "heller - dunkler", "besser - schlechter" oder ähnliche Relation, nicht jedoch ein Abstand zwischen den Ausprägungen definiert. Aus diesem Grunde darf - für den Fall, dass den Ausprägungen (Ordinal-) Zahlen zugeordnet wurden - mit diesen Zahlen auch nicht gerechnet werden. Anders ausgedrückt: mathematische Operationen mit diesen Zahlen sind nicht sinnvoll, da sie keinen numerischen Wert sondern eine Kategorie (z.B. „zufrieden“) darstellen. So ist beispielsweise eine Addition "zufrieden / unzufrieden" wenig sinnvoll.

Ein typisches Beispiel für ein ordinales Merkmal ist die (Schul-) Note. Deren Ausprägungen werden entweder verbal: sehr gut, gut, befriedigend, ausreichend, mangelhaft oder häufig mit Ordinalzahlen 1, 2, 3, 4, 5 bezeichnet. Dabei bedeutet die Note 2 keineswegs, dass die bewertete Arbeit doppelt so gut ist, wie eine Arbeit, die mit der Note 4 bewertet wurde. Die Note "2" ist nur (viel?) besser (wie viel?) als die Note "4". Ordinale Merkmale sind invariant gegenüber streng monotonen (rangerhaltenden) Transformationen, wie z.B. die Transformation der verbal bezeichneten Noten in Ordinalzahlen. Ordinale Merkmale werden im Folgenden mit U oder V, deren Ausprägungen mit U_i $(i = 1,2,...,k)$ oder V_j $(j = 1,2,...,m)$ bezeichnet, dabei ist k bzw. m die Anzahl der jeweiligen Ausprägungen.

2.4.2.3 Metrische Merkmale

Die Ausprägungen $x, y \in \Re$ metrischer Merkmale, die mit X bzw. Y bezeichnet werden, werden auf Intervall- bzw. Verhältnisskalen (sogenannte Kardinalskalen) gemessen.

Dabei werden ihnen auf der Intervallskala reelle Zahlen zugeordnet, für die (nur) Abstände (Differenzen) definiert sind. Typische Beispiele für metrische Merkmale, die auf einer Intervallskala gemessen werden, sind die Temperatur gemessen in °C oder °F und die Kalenderzeit. Das bedeutet, dass z.B. für das Merkmal Temperatur die Differenz 30°C - 10°C = 20°C sinnvoll ist, nicht jedoch das Verhältnis $30/10 = 3$. In einem Raum mit 30°C ist es nicht 3-mal so warm wie in einem Raum mit 10°C. Wird die Temperatur auf einer anderen Skala, z.B. °F, gemessen, ändert sich das Verhältnis. Das liegt daran, dass bei intervallskalierten Merkmalen der Nullpunkt auf Konvention beruht (Wasser gefriert, Christi Geburt). Intervallskalierte Merkmale sind invariant gegenüber linearen Transformationen der Art

$$Y = aX + b; \quad a \in \Re; a \neq 0, b \in \Re.$$

Den Ausprägungen metrischer Merkmale, die auf einer Verhältnisskala gemessen werden, werden positive reelle Zahlen zugeordnet, für die Abstände (Differenzen) und Verhältnisse definiert sind. Verhältnisskalierte Merkmale haben einen natürlichen Nullpunkt, daher sind, im Gegensatz zu intervallskalierten Merkmalen, auch Verhältnisse (Quotienten) definiert. Typische Beispiele für metrische Merkmale, die auf einer Verhältnisskala gemessen werden, sind die Temperatur gemessen in °K, Gewicht, Länge, Volumen u. dgl. Verhältnisskalierte Merkmale sind invariant gegenüber linearen Transformationen der Art

$$Y = aX; \qquad a \in \Re; a \neq 0.$$

Die Ausprägungen des metrischen Merkmals X bzw. Y beim j-ten Merkmalsträger werden mit x_j bzw. y_j bezeichnet.

Ein Merkmal X heißt diskret, wenn die Menge seiner Ausprägungen eine diskrete Menge ist, d.h. wenn X endlich oder abzählbar unendlich viele Ausprägungen $(x_1, x_2, x_3, ...)$ hat. Beispiel: Anzahl der Kinder einer Frau.

Ein Merkmal X heißt stetig (oder kontinuierlich), wenn die Menge seiner Ausprägungen ein Kontinuum ist, d.h. wenn X überabzählbar viele Ausprägungen hat. Beispiel: das Geburtsgewicht von Schweinen.

2.4.3 Nominale Merkmale: Streuung und Gütefunktion

Ein inzwischen gebräuchliches Streuungsmaß für nominale Merkmale ist die mittlere Entropie, deren Eigenschaften mit denen der Varianz für metrische Merkmale vergleichbar sind. Die mittlere Entropie ist für das i-te Merkmal A_i (mit L_i Ausprägungen), $i = 1,2,...,m$, wie folgt definiert:[5)]

$$H(A_i) = \operatorname{ld} n - \frac{1}{n}\sum_{l=1}^{L_i} n_{il} \operatorname{ld} n_{il}$$

$$= -\sum_{l=1}^{L_i} f_{il} \operatorname{ld} f_{il} \quad ;$$

dabei ist ld der Logarithmus zur Basis 2, n_{il} (f_{il}) ist die absolute (relative) Häufigkeit der l-ten Ausprägung und n ist die Summe der n_{il} (die Anzahl der Merkmalsträger). Man definiert: $0 \operatorname{ld} 0 = 0$. Man beachte, dass bei der Berechnung des Streuungsmaßes $H(A_i)$ mit den Ausprägungen des nominalen Merkmals nicht gerechnet wird, sondern nur mit den absoluten bzw. relativen Häufigkeiten.

Es gilt $0 \le H(A_i) \le \operatorname{ld} L_i$, mit $H(A_i) = 0$ für eine Ein-Punkt-Verteilung (Streuung minimal) und $H(A_i) = \operatorname{ld} L_i$ für eine Gleichverteilung (Streuung maximal). Somit kann $H(A_i)$ auf das Intervall [0,1] normiert werden:

$$0 \le \frac{H(A_i)}{\operatorname{ld} L_i} = H(A_i)_{\text{norm.}} \le 1.$$

5) Vgl. VOGEL, F., Beschreibende und schließende Statistik, Formeln, Definitionen, Erläuterungen, Stichwörter und Tabellen, 13. Aufl., München 2005.

Die totale Entropie

$$H_T(A_i) = n \cdot H(A_i)$$

ist ein mit der Fehlerquadratsumme (Summe der Abstandsquadrate) für metrische Merkmale vergleichbares Streuungsmaß für nominale Merkmale, das wegen

$$0 \le H_T(A_i) \le n \cdot \mathrm{ld}\, L_i$$

auf das Intervall [0,n] normiert werden kann:

$$0 \le \frac{H_T(A_i)}{\mathrm{ld}\, L_i} = H_T(A_i)_{\mathrm{norm.}} \le n$$

und dann, wie die normierte mittlere Entropie auch, unabhängig von der jeweiligen Anzahl der Ausprägungen L_i ist.

Für m nominale Merkmale, also für eine Datenmatrix $\underline{X}$ vom Typ (n,m), ist die gemeinsame normierte mittlere Entropie (Streuung) - wegen der Additionseigenschaft der Entropie - durch

$$H(\underline{X})_{\mathrm{norm.}} = \sum_{i=1}^{m} H(A_i)_{\mathrm{norm.}}$$

und die gemeinsame normierte totale Entropie durch

$$H_T(\underline{X})_{\mathrm{norm.}} = n \cdot \sum_{i=1}^{m} H(A_i)_{\mathrm{norm.}}$$

gegeben. Weil vor allem letztere zur Steuerung der Klassenbildung verwendet wird, werden - der Einfachheit wegen - die weiteren Ausführungen auf dieses Streuungsmaß beschränkt.

Die Entropie kann zerlegt werden in die Entropie innerhalb der Klassen (interne Streuung) - die bei homogenen Klassen "klein", bei heterogenen Klassen "groß" ist - und in die Entropie zwischen den Klassen (externe Streuung).

Ist eine Gesamtheit in K Klassen G_k vom Umfang $n_k, k = 1,2,...,K$, zerlegt, so ist

$$H_T(\underline{X}_k) = n_k \times \sum_{i=1}^{m} H(A_i)_{norm.,k}$$

die gemeinsame normierte totale Entropie (Streuung) der m nominalen Klassifikationsmerkmale in der k-ten Klasse ($k = 1,2,...,K$).

Mit $\underline{X}_k = \underline{X}_{n_k,m}$ wird dabei die der Klasse G_k zugeordnete Datenmatrix, mit $H(A_i)_{norm.,k}$ die normierte Entropie des i-ten Merkmals in der k-ten Klasse bezeichnet.

$H_T(\underline{X}_k)$ mißt die Homogenität der k-ten Klasse. Die k-te Klasse ist umso homogener, desto kleinere Werte dieses Maß annimmt. Im Extremfall, wenn alle m Merkmale in der k-ten Klasse empirischen Ein-Punkt-Verteilungen folgen (maximale Homogenität; die k-te Klasse besteht dann aus n_k identischen Merkmalsträgern), ist $H(\underline{X}_k) = H_T(\underline{X}_k) = 0$.

Die normierte totale interne Entropie (Streuung) für eine Partition der Gesamtheit in K Klassen (P_K) ist durch:

$$\begin{aligned} H_T(\underline{X}, P_K)_{int.,norm.} &= \sum_{k=1}^{K} n_k \sum_{i=1}^{m} H(A_i)_{norm.,k} \\ &= \sum_{k=1}^{K} H_T(\underline{X}_k) \end{aligned}$$

definiert.

Dabei ist $\underline{X}$ als Supermatrix zu interpretieren, deren Submatrizen $\underline{X}_k = \underline{X}_{n_k,m}$, $k = 1,2,...,K$, den einzelnen Klassen G_k der Partition P_K zugeordnet sind.

$H_T(\underline{X}, P_K)_{int.,norm.}$ ist ein Maß für die Homogenität einer ganzen Partition, d.h. einer Zerlegung der Menge der Merkmalsträger in genau K Klassen. Eine Par-

tition P_K gilt als umso besser, je kleiner die Werte dieses Homogenitätsmaßes sind. Das ist eine Durchschnittsbetrachtung, einzelne Klassen einer Partition können mehr oder weniger homogen sein.

Für die totale interne Entropie (die gemeinsame Streuung der m Klassifikationsmerkmale innerhalb der K Klassen) gilt:

$$0 \le H_T(\underline{X}, P_K)_{int.,norm.} \le H_T(\underline{X})_{norm.}.$$

Die Differenz

$$H_T(\underline{X})_{norm.} - H_T(\underline{X}, P_K)_{int.,norm.} = H_T(\underline{X}, P_K)_{ext.,norm.}$$

kann als normierte totale externe Entropie bezeichnet, d.h. als gemeinsame Streuung der m nominalen Klassifikationsmerkmale zwischen den K Klassen interpretiert werden.

Entscheidend für den Einsatz der Entropie zur Steuerung der Klassenbildung sind die folgenden Eigenschaften. Die Verwendung der normierten Entropien (Streuungen) innerhalb der Klassen, also von $H(A_i)_{norm.,k}$, hat den entscheidenden Vorteil, dass jedes nominale Merkmal A_i - unabhängig von der Anzahl seiner Ausprägungen L_i - mit einem numerisch gleichen maximalen Gewicht in den Klassifikationsprozeß eingeht, denn die normierte Entropie eines jeden nominalen Merkmals variiert im Intervall [0,1]. Dabei ist wesentlich, dass dies nicht nur für den gesamten Datensatz, sondern auch für die einzelnen Klassen gilt. Daher kann beim Prozeß der Klassenbildung kein nominales Merkmal ein anderes Merkmal numerisch dominieren.

$$H_T(\underline{X}, P_K)_{int.,norm.}$$

wird für den Fall, dass die Merkmalsträger nur durch m nominale Merkmale beschrieben sind, beim Austauschverfahren als - zu optimierende (zu minimierende) - Gütefunktion $g(\underline{X}, P_K)$ eingesetzt. Beim hierarchisch-agglomerativen Verfahren wird

$$H_T\left(\underline{X}, G_p \cup G_q\right)_{\text{int.,norm.}},\ p \neq q,$$

verwendet, um jene beiden Klassen zu finden, aus deren Fusion der geringste Heterogenitätszuwachs resultiert.

2.4.4 Ordinale Merkmale: Streuung und Gütefunktion

Die Streuung des i-ten ordinalen Merkmals U_i mit L_i Ausprägungen und den absoluten Häufigkeiten n_{il} kann mit dem Streuungsmaß

$$S(U_i) = \sum_{p=1}^{L_i - 1} H(B_{ip})$$

gemessen werden.[6)]

Dabei ist

$$H(B_{ip}) = \operatorname{ld} n - \frac{1}{n}\left[n_{ip}^{*} \operatorname{ld} n_{ip}^{*} + \left(n - n_{ip}^{*}\right) \operatorname{ld}\left(n - n_{ip}^{*}\right)\right] \text{ und}$$

$$n_{ip}^{*} = \sum_{l=p+1}^{L_i} n_{il},\ p = 1,2,...,L_i - 1;$$

n_{il} ist die absolute Häufigkeit der l-ten Ausprägung des i-ten ordinalen Merkmals U_i und n ist die Summe der n_{il}.

Bei dieser Art von Streuungsmessung wird das ordinale Klassifikationsmerkmal U_i so auf $L_i - 1$ selbständige binäre Merkmale B_{ip} abgebildet, dass die kumulierten Häufigkeiten n_{ip}^{*} des ordinalen Merkmals U_i mit den absoluten Häufigkeiten des entsprechenden binären Merkmals B_{ip} übereinstim-

6) F. VOGEL/ R. DOBBENER, Ein Streuungsmaß für komparative Merkmale, Jahrbücher für Nationalökonomie und Statistik, 197/2(1982), S. 145-157.
F. VOGEL, Streuungsmessung ordinalskalierter Merkmale, Jahrbücher für Nationalökonomie und Statistik, 208/3(1991), S. 299-318.
Vgl. auch F. VOGEL, Beschreibende und schließende Statistik, Formeln, Definitionen, Erläuterungen, Stichwörter und Tabellen, 13. Aufl., München 2005.

men. Dabei bleibt die Ordnungsstruktur, d.h. die Ordnungsinformation des ordinalen Merkmals erhalten.

Man beachte, dass auch bei der Berechnung des Streuungsmaßes $S(U_i)$ mit den Ausprägungen des ordinalen Merkmals (den Ordinalzahlen) nicht gerechnet wird, sondern nur mit den absoluten bzw. relativen Häufigkeiten.

Für die Klassifikation wesentlich sind insbesondere die folgenden Eigenschaften dieses Streuungsmaßes:

$S(U_i)$ ist - als Funktion der relativen Häufigkeiten - stetig, verändert sich also nur geringfügig, wenn sich die relativen Häufigkeiten nur unwesentlich ändern.

Es gilt: $0 \le S(U_i) \le L_i - 1$, mit $S(U_i) = 0$, falls $n_{il_0} = n$ für irgendein l_0 (Streuung minimal) und $S(U_i) \le L_i - 1$, falls $n_{il} = n_{iL_i} = \frac{n}{2}$ (Streuung maximal).

Daraus folgt, dass $S(U_i)$ auf das Intervall [0,1] normiert werden kann:

$$0 \le \frac{S(U_i)}{L_i - 1} = S(U_i)_{norm.} \le 1$$

und dann unabhängig ist von der jeweiligen Anzahl der Ausprägungen L_i.

$$S_T(U_i)_{norm.} = n \cdot S(U_i)_{norm.}$$

ist ein mit der Fehlerquadratsumme für metrische Merkmale vergleichbares Streuungsmaß (vgl. Entropie).

Die gemeinsame Streuung von m ordinalen Merkmalen läßt sich - analog zur Vorgehensweise bei m nominalen Merkmalen (s.o.) - durch

$$S(\underline{X}) = \sum_{i=1}^{m} S(U_i) \text{ oder durch } S(\underline{X})_{norm.} = \sum_{i=1}^{m} S(U_i)_{norm.}$$

messen.

$S(U_i)$ kann - wie die Entropie nominaler Merkmale - zerlegt werden in die Streuung innerhalb und zwischen den Klassen (interne bzw. externe Streuung).

Ist eine Gesamtheit in K Klassen G_k vom Umfang n_k, $k = 1,2,...,K$, zerlegt, so ist

$$S(\underline{X}_k) = \sum_{i=1}^{m} S(U_i)_{norm.,k}$$

die gemeinsamen normierte Streuung der m ordinalen Merkmale in der k-ten Klasse.

Mit $\underline{X}_k = \underline{X}_{m,n_k}$ wird dabei die der Klasse G_k zugeordnete Datenmatrix, mit $S(U_i)_{norm.,k}$ die normierte Streuung des i-ten Merkmals in der k-ten Klasse bezeichnet.

$S(\underline{X}_k)$ mißt die Homogenität der k-ten Klasse. Die k-te Klasse ist umso homogener, desto kleinere Werte dieses Maß annimmt. Im Extremfall, wenn alle m ordinalen Merkmale in der k-ten Klasse empirischen Ein-Punkt-Verteilungen folgen (maximale Homogenität; die Klasse besteht dann aus n_k identischen Merkmalsträgern), ist $S(\underline{X}_k) = 0$.

Die normierte interne Streuung für eine Partition der Gesamtheit in K Klassen (P_K) ist durch:

$$S(\underline{X}, P_K)_{int.,norm.} = \frac{1}{n} \sum_{k=1}^{K} n_k \sum_{i=1}^{m} S(U_i)_{norm.,k}$$

definiert.

Dabei ist $\underline{X}$ (wieder) als Supermatrix zu interpretieren, deren Submatrizen $\underline{X}_k = \underline{X}_{m,n_k}$, $k = 1,2,...,K$, den einzelnen Klassen G_k der Partition P_K zugeordnet sind.

$S(\underline{X}, P_K)_{int.,norm.}$ ist ein Maß für die Homogenität einer ganzen Partition, d.h. einer Zerlegung der Menge der Merkmalsträger in genau K Klassen. Eine Partition P_K gilt als umso besser, je kleiner die Werte dieses Homogenitätsmaßes sind. Das ist eine Durchschnittsbetrachtung, einzelne Klassen einer Partition können mehr oder weniger homogen sein.

Für die normierte interne Streuung (die gemeinsame normierte Streuung der m ordinalen Klassifikationsmerkmale innerhalb der K Klassen) gilt:

$$0 \le S(\underline{X}, P_K)_{int.,norm.} \le S(\underline{X})_{norm.}$$

und die Differenz

$$S(\underline{X})_{norm.} - S(\underline{X}, P_K)_{int.,norm.} = S(\underline{X}, P_K)_{ext.,norm.}$$

kann als normierte externe Streuung bezeichnet, d.h. als gemeinsame Streuung der m ordinalen Klassifikationsmerkmale zwischen den K Klassen interpretiert werden.

Es gilt:

$$S(\underline{X}, P_K)_{int.,norm} = 0, \text{ also } S(\underline{X}, P_K)_{ext.,norm.} = S(\underline{X})_{norm.}$$

genau dann, wenn die empirischen Verteilungen aller m Klassifikationsmerkmale in allen K Klassen Ein-Punkt-Verteilungen sind (das ist das eine Extrem).

Sind die relativen Häufigkeitsverteilungen der m Merkmale in allen K Klassen identisch, gibt es also - abgesehen von der Klassenbesetzung - keine Unterschiede zwischen den Klassen, ist

$S(\underline{X}, P_K)_{int.,norm.} = S(\underline{X})_{norm.}$, also $S(\underline{X}, P_K)_{ext.,norm.} = 0$ (das ist das andere Extrem).

Wichtig für den Einsatz dieses Streuungsmaßes zur Steuerung der Klassenbildung sind die folgenden Eigenschaften. Die Verwendung der normierten Streuungen innerhalb der Klassen, also von $S(U_i)_{norm.,k}$, hat den entscheidenden Vorteil, dass auch jedes ordinale Merkmal U_i - unabhängig von der Anzahl seiner Ausprägungen L_i - mit einem numerisch gleichen maximalen Gewicht in den Klassifikationsprozess eingeht, denn die normierte Streuung eines jeden Merkmals variiert im Intervall [0,1]. Dabei ist wesentlich, dass dies nicht nur für den gesamten Datensatz, sondern auch für die einzelnen Klassen gilt. Daher kann beim Prozeß der Klassenbildung kein ordinales Merkmal ein anderes numerisch dominieren.

$$S_T(\underline{X}, P_K)_{int.,norm.} = \sum_{k=1}^{K} n_k \sum_{i=1}^{m} S(U_i)_{norm.,k}$$

wird für den Fall, dass die Merkmalsträger nur durch m ordinale Merkmale beschrieben sind, beim Austauschverfahren als - zu optimierende (minimierende) - Gütefunktion $g(\underline{X}, P_K)$ eingesetzt. Beim hierarchisch-agglomerativen Verfahren wird

$$S_T(\underline{X}, G_p \cup G_q)_{int.,norm.},\ p \neq q,$$

verwendet, um jene beiden Klassen zu finden, aus deren Fusion der geringste Heterogenitätszuwachs resultiert.

2.4.5 Metrische Merkmale: Streuung und Gütefunktion

Ein wegen seiner Eigenschaften sehr gebräuchliches Streuungsmaß für metrische Merkmale ist die Varianz/Fehlerquadratsumme. Die Varianz ist für das i-te metrische Merkmal X_i wie folgt definiert:

$$\mathrm{Var}(X_i) = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 \geq 0 \ .$$

Es gilt: $\mathrm{Var}(X_i) = 0$, falls $x_{ij} = \bar{x}_i$ für alle $j = 1,2,...,n$.

$$n\,\mathrm{Var}(X_i)$$

heißt Fehlerquadratsumme oder Summe der Abstandsquadrate.

Die gemeinsame Streuung von m metrischen Merkmalen, also die Streuung in einer "metrischen" Datenmatrix $\underline{X}$ vom Typ m,n , kann mit

$$\mathrm{Var}(\underline{X}) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

oder mit $n\,\mathrm{Var}(\underline{X})$ gemessen werden.

Auch die Varianz/Fehlerquadratsumme kann in eine interne und externe Varianz/Fehlerquadratsumme zerlegt werden.

Ist eine Gesamtheit in K Klassen G_k vom Umfang n_k, $k = 1,2,...,K$, zerlegt, so ist

$$\mathrm{Var}(\underline{X}_k) = \frac{1}{n_k}\sum_{i=1}^{m}\sum_{j\in G_k}(x_{ij} - \bar{x}_{ik})^2$$

die gemeinsame Varianz der m metrischen Merkmale in der k-ten Klasse, $k = 1,2,...,K$.

Mit $\underline{X}_k = \underline{X}_{m,n_k}$ wird wieder die der Klasse G_k zugeordnete Datenmatrix bezeichnet.

$\mathrm{Var}(\underline{X}_k)$ mißt die Homogenität der k-ten Klasse. Die k-te Klasse ist um so homogener, je kleiner $\mathrm{Var}(\underline{X}_k)$ ist. Im Extremfall, d.h. wenn $x_{ij} = \bar{x}_{ik}$ für alle $i = 1,2,...,m$ und alle $j \in G_k$ ist, ist $\mathrm{Var}(\underline{X}_k) = 0$. Die Klasse k besteht dann aus n_k identischen Merkmalsträgern.

Die interne Varianz für eine Partition der Gesamtheit in K Klassen (P_K) ist durch

$$\mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j\in G_k}\left(x_{ij} - \bar{x}_{ik}\right)^2$$

definiert.

$\mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}}$ ist ein Homogenitätsmaß für eine Partition der Menge der Merkmalsträger in genau K Klassen. Eine Partition gilt als umso besser, je kleiner die Werte dieses Homogenitätsmaßes sind.

Es gilt:

$$0 \le \mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}} \le \mathrm{Var}(\underline{X}),$$

und die Differenz

$$\mathrm{Var}(\underline{X}) - \mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}} = \mathrm{Var}(\underline{X}, P_K)_{\mathrm{ext.}}$$

heißt externe Varianz (Varianz zwischen den Klassen).

Es ist

$$\mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}} = 0, \text{ also } \mathrm{Var}(\underline{X}, P_K)_{\mathrm{ext.}} = \mathrm{Var}(\underline{X})$$

falls $x_{ij} = \bar{x}_{ik}$ für alle Merkmale, für alle $j \in G_k$ und alle k ist, wenn also alle Klassen aus jeweils identischen Merkmalsträgern bestehen.

Andererseits ist

$$\mathrm{Var}(\underline{X}, P_K)_{\mathrm{ext.}} = 0, \text{ also } \mathrm{Var}(\underline{X}, P_K)_{\mathrm{int.}} = \mathrm{Var}(\underline{X})$$

falls $\bar{x}_{ik} = \bar{x}_i$ für alle i und k ist, wenn also die Klassenmittelwerte und der Gesamtmittelwert für alle m Merkmale übereinstimmen. Entsprechendes gilt für die Fehlerquadratsumme.

Die Verwendung der internen Varianz $Var(\underline{X}, P_K)_{int.}$ oder der Fehlerquadratsumme $n \cdot Var(\underline{X}, P_K)_{int.}$ zur Steuerung der Klassenbildung hat bei praktischen Anwendungen den entscheidenden Nachteil, dass diese Homogenitätsmaße nicht skaleninvariant sind. Der jeweilige Wert dieser Maße ist abhängig von den Maßeinheiten (der Dimension), in welchen die einzelnen Merkmale gemessen wurden. Daher haben einzelne Merkmale - nur aufgrund ihrer Maßeinheit (gemessen in kg oder g, in km oder cm) - beim Prozeß der Klassenbildung ein zahlenmäßig höheres Gewicht und können andere Merkmale numerisch dominieren.

Bei einer Klassifikation der Länder der Erde nach ihrem Entwicklungsstand[7)] haben beispielsweise Merkmale wie das Bruttoinlandsprodukt pro Kopf (Wertebereich 95 - 12.500 US $) ein derart hohes Gewicht, dass sie andere Merkmale wie z.B. die Haushaltsdichte [8)] (Wertebereich: 2,5 - 6,9 Personen) im Prozeß der Klassenbildung numerisch dominieren.

Da alle Merkmale im Prozeß der Klassenbildung numerisch gleich behandelt werden, könnte man auf den Gedanken kommen, die metrischen Merkmale vorab geeignet zu normieren, um dieses Invarianzproblem zu lösen. Die daraus entstehenden Schwierigkeiten sind überaus vielfältig, vor allem, weil die Klassifikationsergebnisse entscheidend durch die Art der Normierung determiniert werden.

Es gibt viele Möglichkeiten, metrische Merkmale zu normieren. Gebräuchlich sind die beiden folgenden Normierungen.

- Standardisierung:

$$x^*_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}, \text{ mit } \bar{x}_i = \frac{1}{n}\sum_j x_{ij} \text{ und } s_i^2 = \frac{1}{n-1}\sum_j \left(x_{ij} - \bar{x}_i\right)^2$$

7) Vgl. VOGEL, F, Unterentwicklung - Entwicklung, Eine Studie zur Einteilung der Länder der Erde nach ihrem Entwicklungsstand, Arbeiten aus der Statistik, Bamberg 1989.
8) Die Haushaltsdichte entspricht der durchschnittlichen Haushaltsgröße.

Die standardisierten Merkmale X_i^*, $i = 1,2,...,m$, haben den Mittelwert 0 und die Varianz 1.

- [0;1]-Normierung:

$$x_{ij}^* = \frac{x_{ij} - z_i}{r_i} \text{ mit } z_i = \underbrace{\min}_{j}(x_{ij}),\ u_i = \underbrace{\max}_{j}(x_{ij}) \text{ und } r_i = u_i - z_i .$$

Dabei werden die Merkmalswerte so normiert, dass die normierten Werte x_{ij}^*, $i = 1,2,...,m$, im Intervall [0;1] liegen und die Randpunkte dieses Intervalls auch annehmen.

Es gibt allerdings noch eine ganze Reihe anderer Normierungsmöglichkeiten.[9)]

Durch eine solche Normierung soll der Einfluß ungleicher Maßeinheiten der Merkmalswerte eliminiert und die Merkmale für die Klassenbildung ein numerisch gleiches Gewicht bekommen.

Dieses Ziel kann jedoch durch eine Normierung der Merkmale nicht oder nur teilweise erreicht werden. Die Normierung der Merkmale bewirkt, dass die Achsen des m-dimensionalen Merkmalsraums - in dem die Merkmalsträger als Punkte, die Klassen als Punkthäufungen erscheinen - gestaucht oder gestreckt und somit auch Form, Ausdehnung und Lage der Punkthäufungen (Klassen) im Merkmalsraum verändert werden. Es kommt hinzu, dass die Merkmale durch die jeweilige Normierung ein jeweils anderes (welches?) numerisches Gewicht bekommen, das keineswegs das richtige Gewicht sein muss.

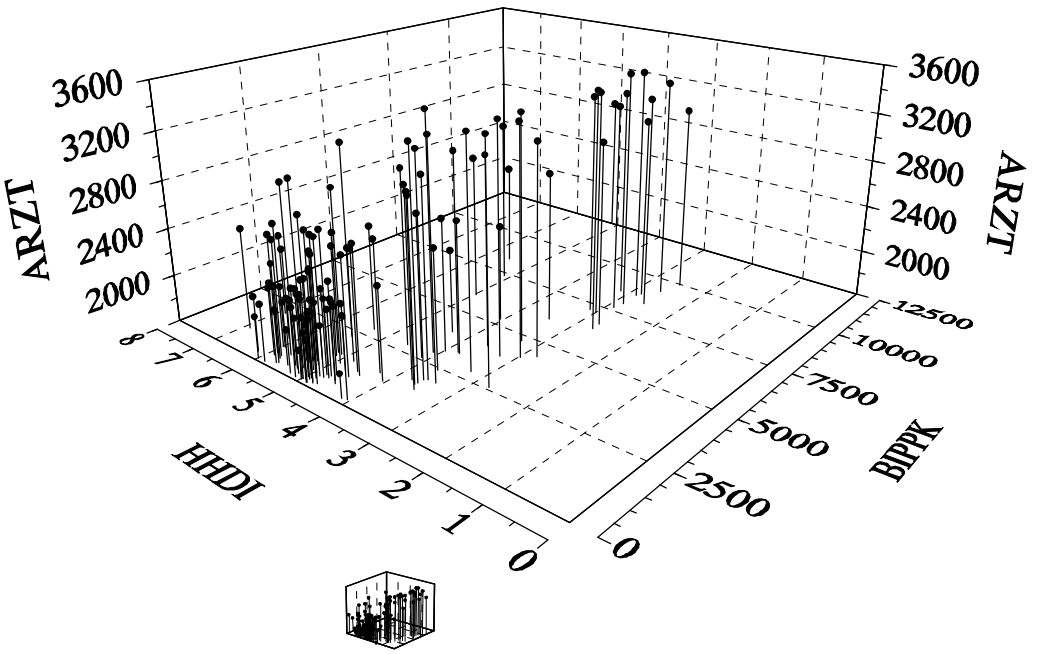
Der durch die Normierung der Merkmale entstehende (Merkmals-)Raum heißt Klassifikationsraum, weil die Klassen in diesem Raum gebildet werden.

9) Vgl. z.B. VOGEL, F., Subjektivitäten bei der Klassifikation von Einheiten, Proc. Op. Res. 7(1977), S.105-129.

Die folgende Abbildung zeigt einen dreidimensionalen Merkmalsraum mit den Merkmalen BIPPK (Bruttoinlandsprodukt pro Kopf), HHDI (Haushaltsdichte) und ARZT (Anzahl der Einwohner je Arzt) und den zugehörigen (jedoch vergrößerten) Klassifikationsraum bei einer [0;1]-Normierung der drei Merkmale.[10)]

Es ist plausibel, dass sich die Punkthäufungen (Klassen) durch die Normierung in ihrer Form, Ausdehnung und Lage sowie in ihrer Zusammensetzung verändert haben. Wenn in dem zu klassifizierenden Datensatz wohlseparierte Klassen existieren, hat eine solche Normierung keinen oder fast keinen Einfluß auf das Klassifikationsergebnis. Im Allgemeinen sind aber bei praktischen Anwendungen die Räume zwischen den Klassen auch von Merkmalsträgern besetzt, so dass diese Merkmalsträger - je nach der Art der Normierung - mal der einen mal der anderen Klasse zugeordnet werden.

Abbildung 2: Merkmalsraum



Jede Normierung liefert somit mehr oder minder verschiedene Klassifikationsergebnisse, je nach der Klassenstruktur in den empirischen Daten. Die

10) Merkmalsträger sind die Länder der Erde.

durch die jeweilige Normierung bedingte Unterschiedlichkeit der Klassifikationsergebnisse hängt auch davon ab, welches Verfahren zur Klassenbildung eingesetzt wird. Die Unterschiede zwischen den Klassifikationsergebnissen sind am größten, wenn hierarchisch-agglomerative Verfahren eingesetzt werden, weil die Verfahren dieser Gruppe den Weg zu einer Hierarchie von Klassen optimieren und der einmal eingeschlagene Weg irreversibel ist.

Da keine Normierung gegenüber den anderen ausgezeichnet ist, sind die Klassifikationsergebnisse in dieser Hinsicht subjektiv. Mit einer Normierung der Merkmale kann also das Invarianzproblem nicht gelöst werden. Es kommt hinzu, dass die gebildeten Klassen inhaltlich interpretiert werden sollen, und zwar im "Merkmalsraum" (d.h. mit den ursprünglichen, nicht normierten Merkmalen). Aber in diesem Raum existieren die im Klassifikationsraum gebildeten Klassen nicht oder nur zum Teil.

Es erscheint ratsam, Klassifikationsverfahren einzusetzen, die eine derartige Normierung metrischer Merkmale nicht voraussetzen. Außerdem sollte in dem gewählten Verfahren der Einfluß ungleicher Maßeinheiten der Merkmalswerte eliminiert werden und die Merkmale für die Klassenbildung ein numerisch gleiches Gewicht bekommen.

3 Die Verarbeitung gemischter Merkmale

3.1 Gemischte Merkmale

Bei praktischen Anwendungen werden die Merkmalsträger i.a.R. anhand nominaler, ordinaler und metrischer Merkmale, also durch gemischte Merkmale, beschrieben. Bei der bereits erwähnten Studie "Einteilung der Länder der Erde nach ihrem Entwicklungsstand" wurden neben zahlreichen metrischen Merkmalen aus den Bereichen "Wirtschaft", "Bevölkerung" sowie "Grundbedürfnisse und Soziales" auch nominale und ordinale Merkmale wie z.B. im

Bereich "Politik" das Ausmaß des Föderalismus (unitarisch, schwach föderalistisch, stark föderalistisch) und die Politische Freiheit der Opposition (Opposition unbeschränkt, etwas beschränkt, stark beschränkt, illegal)[11] verwendet.

Die Konstruktion einer Gütefunktion für solche gemischten Merkmale durch Aggregation (Addition) der merkmalsspezifischen Gütefunktionen ist aus meßtheoretischen Gründen nicht möglich, weil die Gütefunktionen für nominale und ordinale Merkmale auf der Entropie, die Gütefunktion für metrische Merkmale aber auf der Varianz beruhen.

3.2 Die Ordinalisierung metrischer Merkmale

Metrische Merkmale liefern Ordnungs- und Abstandsinformationen. Sie sind also auch ordinal, da auf ihrer Skala eine Ordnungsrelation definiert ist. Für Datenmatrizen mit gemischten Merkmalen kann die Abstandsinformation bei der numerischen Klassifikation nicht adäquat genutzt werden. Es kommt hinzu, dass die metrischen Merkmalsausprägungen/-werte sehr häufig fehlerbehaftet/un-scharf/vage sind, so dass es ohnehin nicht ratsam ist, sie wie exakte Daten zu verarbeiten. Wir haben erhebliche Zweifel, ob es in der angewandten Statistik überhaupt exakte metrische Daten gibt.

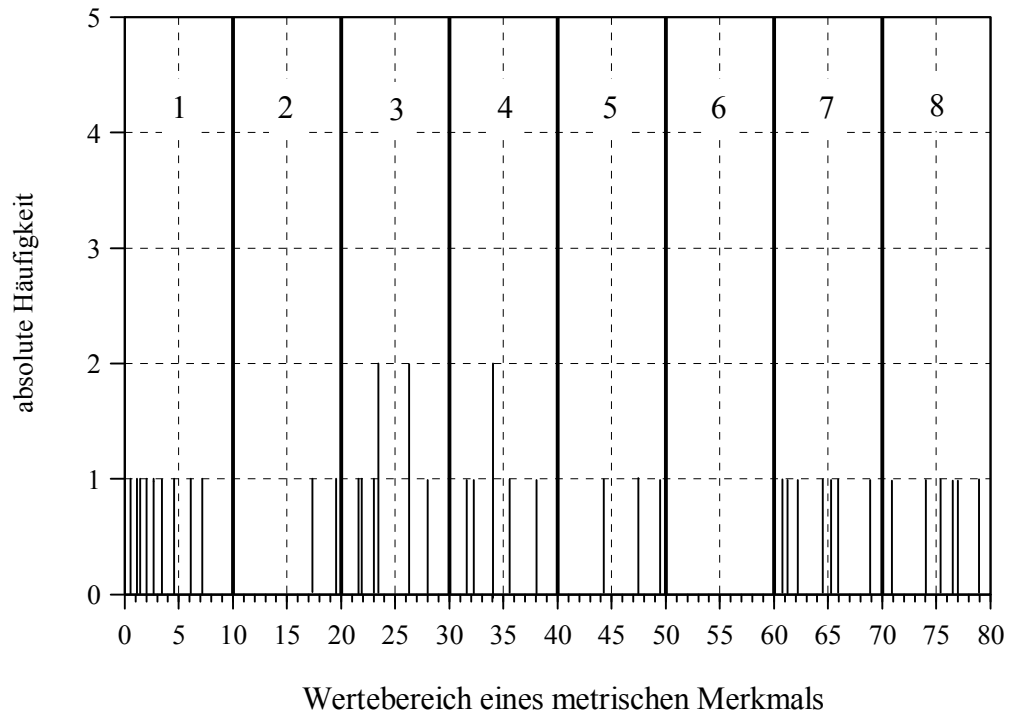
Es liegt daher nahe, metrische Merkmale in ordinale Merkmale mit endlich vielen Ausprägungen zu transformieren. Diese Transformation heißt Ordinalisierung (oder Komparatisierung). Bei der Ordinalisierung wird der Wertebereich eines jeden metrischen Merkmals in endlich viele Intervalle zerlegt. Die Intervallnummern - aufsteigend geordnet - werden als Ausprägungen ordinaler Merkmale (Ordinalzahlen) aufgefaßt und anstelle der metrischen Merkmalsausprägungen/-werte den Merkmalsträgern zugeordnet. Ein Beispiel soll zur Veranschaulichung beitragen.

11) Die Klassifikationsergebnisse zeigten, und das ist bemerkenswert, dass in der Klasse der Entwicklungsländer im engeren Sinne nahezu alle Länder Defizite im Bereich "Politik" aufwiesen.

Ein metrisches Merkmal habe Merkmalswerte zwischen 0 und 80.

Der Wertebereich wird - beispielsweise - in acht äquidistante Intervalle eingeteilt. Allen Merkmalsträgern mit Merkmalswerten zwischen 0 und 10 wird die Ordinalzahl 1, allen Merkmalsträgern mit Merkmalswerten zwischen 10 und 20 wird die Ordinalzahl 2 usf. zugeordnet. Durch diese Zuordnung entsteht ein ordinales Merkmal mit acht Ausprägungen. Neun Merkmalsträger haben nun die ordinale Ausprägung (Ordinalzahl) 1, zwei Merkmalsträger die ordinale Ausprägung 2 und sechs Merkmalsträger die ordinale Ausprägung 8.

Abbildung 3: Häufigkeitsverteilung eines metrischen Merkmals

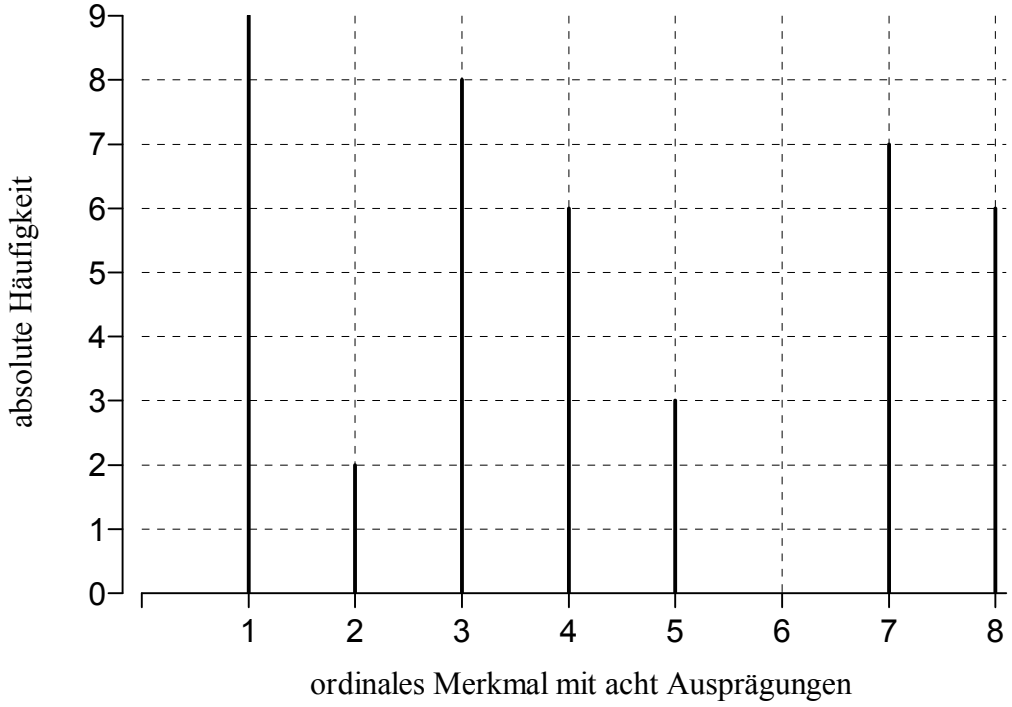


Die Streuung dieses ordinalen Merkmals kann mit $S(U)$ bzw. $S(U)_{norm.}$ gemessen werden (s.o.).

Der Wertebereich des metrischen Merkmals kann in äquidistante Klassen (s.o.), aber auch in Klassen, die mit zunehmenden Merkmalswerten breiter werden, eingeteilt werden.

Bei einer solchen Ordinalisierung geht die ohnehin nicht exakt verwertbare Abstandsinformation metrischer Merkmale verloren. Der Informationsverlust hängt ab von der Anzahl der gebildeten Klassen (für die es keine verbindlichen Regeln gibt), also von der Anzahl der Ausprägungen des ordinalen Merkmals, in das das metrische Merkmal transformiert wird. Durch die Ordinalisierung wird auch das Problem fehlerbehafteter, unscharfer oder vager Daten bei günstiger Intervallbildung weitestgehend gelöst. Untersuchungen haben gezeigt, dass Klassifikationsergebnisse hinsichtlich der Anzahl und Breite der für die Ordinalisierung ge-

Abbildung 4: Häufigkeitsverteilung des zugehörigen ordinalen Merkmals



bildeten Merkmalsklassen erstaunlich robust sind.[12)] Von Vorteil ist, dass die Ordnungsinformation metrischer Merkmale auch vollständig ausgeschöpft werden kann, indem anstelle der metrischen Merkmalsausprägungen/-werte

12) Vgl. F. VOGEL, Ein Beitrag zur Theorie 'weicher' Systeme: die Komparatisierung metrischer Merkmale, in: Beiträge zur Systemforschung, Festschrift für Adolf Adam, Springer, Wien - New York 1985, S. 314 - 333.

ihre Ordnungsnummern (Ränge) wie Ausprägungen ordinaler Merkmale behandelt werden. Das bedeutet, im Falle "ohne Bindungen" haben die aus den metrischen Merkmalen gebildeten ordinalen Merkmale n Ausprägungen. In diesem Falle bleibt allerdings das Problem fehlerbehafteter, unscharfer oder vager Daten ungelöst.

3.3 Klassifikation anhand gemischter Merkmale

3.3.1 Einführendes

Nach der Ordinalisierung der metrischen Merkmale enthält der Datensatz nur noch nominale und ordinale Merkmale mit jeweils L_i Ausprägungen $(i = 1,2,...,m)$, deren Streuungen mit Hilfe der Entropie gemessen werden können. Da die Entropie die Additionseigenschaft besitzt, kann die gemeinsame interne Streuung (Entropie) der m Merkmale durch Aggregation (Summierung) der Einzelstreuungen bestimmt werden.

3.3.2 Das Austauschverfahren

Die zu minimierende Gütefunktion

$$g(\underline{X}, P_K)$$

beruht auf Streuungsmaßen für nominale und ordinale Merkmale. Mit

$g_{nom.}(\underline{X}, P_K) = H_T(\underline{X}, P_K)_{int.,norm.}$ der Gütefunktion für nominale Merkmale und

$g_{ord.}(\underline{X}, P_K) = S_T(\underline{X}, P_K)_{int.,norm.}$ der Gütefunktion für ordinale Merkmale ist

$$g(\underline{X}, P_K) = g_{nom.}(\underline{X}_{nom.}, P_K) + g_{ord.}(\underline{X}_{ord.}, P_K)$$

die Gütefunktion für gemischte Merkmale, dabei wird davon ausgegangen dass die aus der Rohdatenmatrix abgeleitete Datenmatrix $\underline{X}$ aus zwei Sub-

matrizen mit nur nominalen bzw. nur ordinalen und - gegebenenfalls - ordinalisierten metrischen Merkmalen besteht.[13)]

$$\underline{X} = \begin{bmatrix} \underline{X}_{\text{nom.}} \\ \underline{X}_{\text{ord.}} \end{bmatrix}.$$

Mit dieser Gütefunktion wird die Menge der Merkmalsträger auf verschiedene Weise in K vorgegebene Klassen zerlegt, die iterativ verbessert werden.

3.3.3 Ein hierarchisch-agglomeratives Verfahren

Bei der hierarchisch-agglomerativen Klassifikation für nominale, ordinale und gemischte Merkmale werden auf jeder der $r = 1,2,...,n-1$ Fusionsstufen jene beiden Klassen p und q, $p \neq q$, fusioniert, aus deren Fusion der geringste Heterogenitätszuwachs

$$\Delta g\left(G_p, G_q\right) = \Delta H_T\left(G_p \cup G_q\right)_{\text{norm.}} + \Delta S_T\left(G_p \cup G_q\right)_{\text{norm.}}$$

resultiert.

Dabei ist

$$\Delta H_T\left(G_p \cup G_q\right)_{\text{norm.}} = H_T\left(\underline{X}_{p\cup q}\right)_{\text{norm.}} - H_T\left(\underline{X}_p\right)_{\text{norm.}} - H_T\left(\underline{X}_q\right)_{\text{norm.}} \text{ und}$$

$$\Delta S_T\left(G_p \cup G_q\right)_{\text{norm.}} = S_T\left(\underline{X}_{p\cup q}\right)_{\text{norm.}} - S_T\left(\underline{X}_p\right)_{\text{norm.}} - S_T\left(\underline{X}_q\right)_{\text{norm.}},$$

so dass also auch mit diesem Verfahren gemischte Merkmale verarbeitet werden können.

4 Schlussbemerkung

Die beschriebene Vorgehensweise hat einige Vorteile. Die Streuung eines jeden Merkmals, unabhängig davon, ob es nominal oder ordinal (gegebenenfalls ordinalisiert) ist, variiert zwischen 0 und 1. Folglich kann beim Prozeß der Klassenbildung kein Merkmal andere numerisch dominieren. Alle Merk-

13) Das gilt auch für die hierarchisch-agglomerative Klassifikation.

male haben ein gleiches maximales Gewicht, so dass nicht nur das Normierungs- , sondern auch das Gewichtungsproblem gelöst ist. Es kommt hinzu, dass die Klassifikationsergebnisse, die Eigenschaften der Klassen, im Merkmalsraum interpretiert werden können und dass das Problem fehlerhafter/vager/unscharfer metrischer Daten weitestgehend gelöst ist.

II Praxis

1 Einleitendes

Im Folgenden wird anhand eines sehr kleinen Beispiels gezeigt, wie mit Hilfe von ORMIX Merkmalsträger, die durch gemischte Merkmale beschrieben sind, klassifiziert werden können.

Die in diesem Beitrag benutzten Daten und Beispieldateien[14)] entstammen fast alle der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS 2004). Das ALLBUS-Programm ist 1980-86, 1991 von der DFG gefördert worden. Die weiteren Erhebungen wurden von Bund und Ländern über die GESIS (Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen) finanziert. ALLBUS wird von ZUMA (Zentrum für Umfragen, Methoden und Analysen e.V., Mannheim) und Zentralarchiv für Empirische Sozialforschung (Köln) in Zusammenarbeit mit dem ALLBUS-Ausschuss realisiert. Die Daten sind beim Zentralarchiv für Empirische Sozialforschung (Köln) erhältlich. Die vorgenannten Institutionen und Personen tragen keine Verantwortung für die Verwendung der Daten in diesem Beitrag.

14) Hinsichtlich der Klassenstruktur sind die mit ORMIX gelieferten "Beispieldateien" sehr unterschiedlich. Wie leicht festzustellen ist, gibt es Beispieldateien mit ausgeprägter Klassenstruktur, solche mit einer weniger ausgeprägten Klassenstruktur und Dateien, die nahezu keine Klassenstruktur aufweisen. Tendenziell nimmt die Klassenstruktur mit zunehmender Anzahl an Merkmalsträgern und Merkmalen ab.

2 Installationsanleitung

2.1 Installationsvoraussetzungen

ORMIX nutzt die .NET-Technologie. Die Programmdateien sind zwar so klein (zusammen ca. 250 KB), dass sie leicht auf eine Diskette passen würden. Da jedoch umfangreiche Beispieldateien mit ausgeliefert werden, umfassen die Installationsdateien ca. 6-7 MByte.

Zur Installation verwendet ORMIX die von Microsoft für Softwareentwickler seit ungefähr 2005 angebotene Technik ClickOnce.

ORMIX kann unter Windows ab Version Windows 2000 installiert werden.

Voraussetzung ist .NET Framework ab Version 2.0.

Das .NET Framework ist in aktuellen Versionen von Windows enthalten oder per Update nachrüstbar. Falls .NET 2.0 nicht installiert ist, versucht das ORMIX-Setup die erforderlichen Komponenten von der Microsoft-Website herunterzuladen. Dazu ist eine Internet-Verbindung notwendig.

2.2 Zur Installation von der CD

Legen Sie die ORMIX-CD in das Laufwerk Ihres PCs ein.

Falls die Installation nicht automatisch beginnt, doppelklicken Sie im Windows-Explorer auf die Datei „setup.exe“. Eventuell benötigen sie eine Internet-Verbindung, damit .NET 2.0 installiert werden kann. Wenn Sie sich sicher sind, dass auf Ihrem PC bereits .NET 2.0 installiert ist, können Sie auch "ormix.application" ebenfalls mit Doppelklick zum Installieren verwenden.

Falls ORMIX bereits installiert ist und dazu ein anderes Laufwerk verwendet wurde, meldet Windows "Die Anwendung ORMIX kann nicht an diesem Speicherort gestartet werden, weil sie bereits an einem anderen Speicherort installiert ist". In diesem Fall ist entweder zuerst ORMIX zu deinstallieren

(siehe unten) oder die Installation ist von dem Speicherort (Laufwerk und Verzeichnis) aus durchzuführen, der das vorhergende Mal verwendet wurde und der in den Details zur Fehlermeldung angezeigt wird.

Nach einem kurzen Moment sollte ORMIX installiert sein und sofort automatisch gestartet werden.

"ORMIX" ist nun auf der Festplatte Ihres PC installiert und in der Programm-Gruppe "Dr. Rudolf Gardill" zu finden. Es kann wie jedes andere Programm gestartet werden.

2.3 ORMIX deinstallieren

Um ORMIX wieder von Ihrem PC zu entfernen wählen Sie in der Systemsteuerung die Funktion zum Verwalten von Software (bzw. in Vista: "Programme und Funktionen"). Über den Eintrag "ORMIX" mit dem Herausgeber "Dr. Rudolf Gardill" können Sie ORMIX vom Computer entfernen. Schließen Sie das Fenster der Systemsteuerung wieder.

3 Dateneingabe: Die Datenmatrix

Ausgangspunkt und Grundlage der Numerischen Klassifikation ist die Datenmatrix. Sie enthält alle Informationen über die zu klassifizierenden Merkmalsträger (Objekte).

$$\underline{X}^{*} = \left(x_{ji}^{*}\right) = \begin{bmatrix} x_{11}^{*} & x_{12}^{*} & \cdots & x_{1i}^{*} & \cdots & x_{1m}^{*} \\ x_{21}^{*} & x_{22}^{*} & \cdots & x_{2i}^{*} & \cdots & x_{2m}^{*} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ x_{j1}^{*} & x_{j2}^{*} & \cdots & x_{ji}^{*} & \cdots & x_{jm}^{*} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{n1}^{*} & x_{n2}^{*} & \cdots & x_{ni}^{*} & \cdots & x_{nm}^{*} \end{bmatrix}$$

$\underline{X}^*$ heißt Rohdatenmatrix. Ihre Elemente x_{ji}^* sind die Merkmalswerte bzw. Merkmalsausprägungen, die die n Merkmalsträger beschreiben. Die x_{ji}^* sind also, je nach Merkmalstyp, reelle Zahlen (bei metrischen Merkmalen), alphanumerische Zeichen oder Ordinalzahlen (bei ordinalen Merkmalen) oder alphanumerische Zeichen oder Nominalzahlen (bei nominalen Merkmalen).

Die Rohdatenmatrix ist stets so aufgebaut, dass die Zeilen ($j = 1,2,...,n$) den n Merkmalsträgern, die Spalten ($i = 1,2,...,m$) den m Merkmalen zugeordnet sind.

Hinweis: Die Rohdatenmatrix darf keine fehlenden Werte (missing data) enthalten.

Die j-te Zeile der Rohdatenmatrix, der Zeilenvektor

$$\underline{x}_j = [x_{j1}^*, x_{j2}^*, \ldots, x_{ji}^*, \ldots, x_{jm}^*]$$

enthält die Merkmalswerte bzw. Merkmalsausprägungen der m Merkmale für den j-ten Merkmalsträger.

Die i-te Spalte der Rohdatenmatrix, der Spaltenvektor

$$\underline{x}_i = \begin{bmatrix} x_{1i}^* \\ x_{2i}^* \\ \vdots \\ x_{ji}^* \\ \vdots \\ x_{ni}^* \end{bmatrix}$$

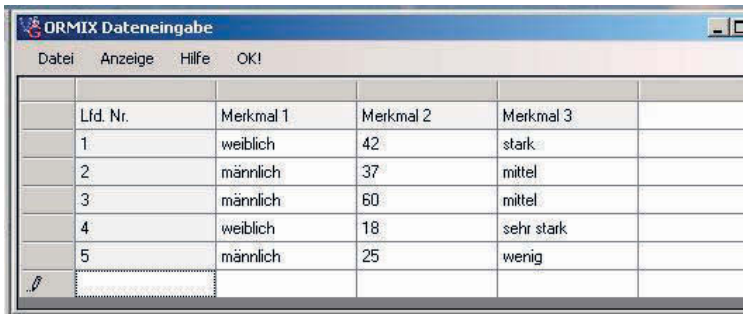
enthält die Merkmalswerte bzw. Merkmalsausprägungen des i-ten Merkmals für die n Merkmalsträger.

Für die Eingabe der Rohdatenmatrix ist diese um eine Vorspalte und eine Kopfzeile zu erweitern. In der Vorspalte stehen die alphanumerischen Ken-

nungen der n Merkmalsträger, in der Kopfzeile stehen die alphanumerischen Kennungen der m Merkmale. Es ist ratsam, die Merkmalsträger- und die Merkmalskennungen mnemotechnisch günstig festzusetzen.

	x_1	x_2	…	x_i	…	x_m
x_1	x^*_{11}	x^*_{12}	⋯	x^*_{1i}	⋯	x^*_{1m}
x_2	x^*_{21}	x^*_{22}	⋯	x^*_{2i}	⋯	x^*_{2m}
⋮	⋮	⋮	⋱	⋮	…	⋮
x_j	x^*_{j1}	x^*_{j2}	⋯	x^*_{ji}	⋯	x^*_{jm}
⋮	⋮	⋮	⋮	⋮	⋱	⋮
x_n	x^*_{n1}	x^*_{n2}	⋯	x^*_{ni}	⋯	x^*_{nm}

Bei kleiner Anzahl von Merkmalsträgern und Merkmalen kann die Rohdatenmatrix direkt im Programm (ORMIX Dateneingabe) eingegeben werden, wie das folgende Beispiel mit nur fünf Merkmalsträgern und drei Merkmalen veranschaulicht.



Die Dateneingabe wird durch "OK" bestätigt.

Die Rohdatenmatrix kann auch in einem beliebigen Editor (z.B. Notepad) erzeugt werden. Dabei darf die erste Zeile keine Leerzeile sein und die Spalten müssen durch Tabstopps getrennt sein. Sie wird z.B. mit "Matrix.txt" abgespeichert und kann unter diesem Namen in den Eingabeteil der Programms übernommen werden.

Beispiel: Erzeugung der Rohdatenmatrix im Editor. N.B.: Die Spalten sind durch Tabstopps getrennt; die erste Zeile darf keine Leerzeile sein![15)]

Lfd. Nr.	Erhebungsgebiet	Alter Befragter	Allgemeiner Schulabschluss	Berufst. Frau: herzl. Verh. z. Kind
1	Ost	30	mittlere Reife	stimme voll und ganz zu
2	Ost	82	Hauptschulabschluss	stimme voll und ganz zu
3	Ost	63	Hauptschulabschluss	stimme voll und ganz zu
4	Ost	74	Hauptschulabschluss	stimme voll und ganz zu
5	Ost	70	Hauptschulabschluss	stimme voll und ganz zu
6	Ost	29	mittlere Reife	stimme voll und ganz zu
7	Ost	78	mittlere Reife	stimme voll und ganz zu
8	West	62	Hauptschulabschluss	stimme voll und ganz zu
9	Ost	27	Abitur	stimme voll und ganz zu
10	Ost	20	mittlere Reife	stimme voll und ganz zu
11	Ost	70	Hauptschulabschluss	stimme eher zu
12	Ost	72	Hauptschulabschluss	stimme voll und ganz zu
13	Ost	20	Abitur	stimme eher zu
14	Ost	49	mittlere Reife	stimme eher zu
15	West	61	Abitur	stimme voll und ganz zu
16	Ost	40	mittlere Reife	stimme voll und ganz zu
17	Ost	64	Hauptschulabschluss	stimme voll und ganz zu
18	Ost	60	mittlere Reife	stimme voll und ganz zu
19	West	63	Hauptschulabschluss	stimme voll und ganz zu
20	West	21	Abitur	stimme voll und ganz zu
21	Ost	29	mittlere Reife	stimme voll und ganz zu
22	Ost	55	Abitur	stimme voll und ganz zu
23	West	39	Hauptschulabschluss	stimme überhaupt nicht zu
24	Ost	59	Hauptschulabschluss	weiß nicht
25	Ost	40	Hauptschulabschluss	stimme eher nicht zu

Zu Demonstrationszwecken sind die vier (willkürlich) ausgewählten Merkmale unterschiedlichen Typs: binär (Erhebungsgebiet), metrisch (Alter des Befragten) und ordinal (allgemeiner Schulabschluss sowie berufst. Frau: herzliches Verhältnis z. Kind).

Die Rohdatenmatrix kann aber auch mit MS Word oder mit MS EXCEL erzeugt werden. Bei einer Eingabe mit MS Word müssen die Spalten der Matrix ebenfalls durch Tabstopps getrennt sein. Die Übernahme in den Eingabeteil

15) Die verwendeten Daten sind ein sehr kleiner Ausschnitt aus der oben genannten ALLBUS-Umfrage.

der Programms geschieht in der Weise, dass die Matrix in MS Word bzw. in MS EXCEL markiert und kopiert, also in die Zwischenablage übertragen wird. Im Eingabeteil der Programms kann dann die Matrix mit "aus Zwischenablage einlesen" übernommen werden. Die Übernahme ist mit „OK“ zu bestätigen.

Beispiel: Erzeugung der Rohdatenmatrix mit MS-Word. N.B.: Die Spalten sind durch Tabstopps getrennt.

Lfd. Nr.	Erhebungs gebiet	Alter Befragter	Allgemeiner Schulabschluss	Berufst. Frau: herzl. Verhältn. z. Kind
1	Ost	30	mittlere Reife	stimme voll und ganz zu
2	Ost	82	Hauptschulabschluss	stimme voll und ganz zu
3	Ost	63	Hauptschulabschluss	stimme voll und ganz zu
4	Ost	74	Hauptschulabschluss	stimme voll und ganz zu
5	Ost	70	Hauptschulabschluss	stimme voll und ganz zu
6	Ost	29	mittlere Reife	stimme voll und ganz zu
7	Ost	78	mittlere Reife	stimme voll und ganz zu
8	West	62	Hauptschulabschluss	stimme voll und ganz zu
9	Ost	27	Abitur	stimme voll und ganz zu
10	Ost	20	mittlere Reife	stimme voll und ganz zu
11	Ost	70	Hauptschulabschluss	stimme eher zu
12	Ost	72	Hauptschulabschluss	stimme voll und ganz zu
13	Ost	20	Abitur	stimme eher zu
14	Ost	49	mittlere Reife	stimme eher zu
15	West	61	Abitur	stimme voll und ganz zu
16	Ost	40	mittlere Reife	stimme voll und ganz zu
17	Ost	64	Hauptschulabschluss	stimme voll und ganz zu
18	Ost	60	mittlere Reife	stimme voll und ganz zu
19	West	63	Hauptschulabschluss	stimme voll und ganz zu
20	West	21	Abitur	stimme voll und ganz zu
21	Ost	29	mittlere Reife	stimme voll und ganz zu
22	Ost	55	Abitur	stimme voll und ganz zu
23	West	39	Hauptschulabschluss	stimme überhaupt nicht zu
24	Ost	59	Hauptschulabschluss	weiß nicht
25	Ost	40	Hauptschulabschluss	stimme eher nicht zu

Beispiel: Erzeugung der Rohdatenmatrix mit MS-EXCEL.

Lfd. Nr.	Erhebungs-gebiet	Alter Befragter	Allgemeiner Schulabschluss	Berufst. Frau: herzl. Verhältnis z. Kind
1	Ost	30	mittlere Reife	stimme voll und ganz zu
2	Ost	82	Hauptschulabschluss	stimme voll und ganz zu
3	Ost	63	Hauptschulabschluss	stimme voll und ganz zu

4	Ost	74	Hauptschulabschluss	stimme voll und ganz zu
5	Ost	70	Hauptschulabschluss	stimme voll und ganz zu
6	Ost	29	mittlere Reife	stimme voll und ganz zu
7	Ost	78	mittlere Reife	stimme voll und ganz zu
8	West	62	Hauptschulabschluss	stimme voll und ganz zu
9	Ost	27	Abitur	stimme voll und ganz zu
10	Ost	20	mittlere Reife	stimme voll und ganz zu
11	Ost	70	Hauptschulabschluss	stimme eher zu
12	Ost	72	Hauptschulabschluss	stimme voll und ganz zu
13	Ost	20	Abitur	stimme eher zu
14	Ost	49	mittlere Reife	stimme eher zu
15	West	61	Abitur	stimme voll und ganz zu
16	Ost	40	mittlere Reife	stimme voll und ganz zu
17	Ost	64	Hauptschulabschluss	stimme voll und ganz zu
18	Ost	60	mittlere Reife	stimme voll und ganz zu
19	West	63	Hauptschulabschluss	stimme voll und ganz zu
20	West	21	Abitur	stimme voll und ganz zu
21	Ost	29	mittlere Reife	stimme voll und ganz zu
22	Ost	55	Abitur	stimme voll und ganz zu
23	West	39	Hauptschulabschluss	stimme überhaupt nicht zu
24	Ost	59	Hauptschulabschluss	weiß nicht
25	Ost	40	Hauptschulabschluss	stimme eher nicht zu

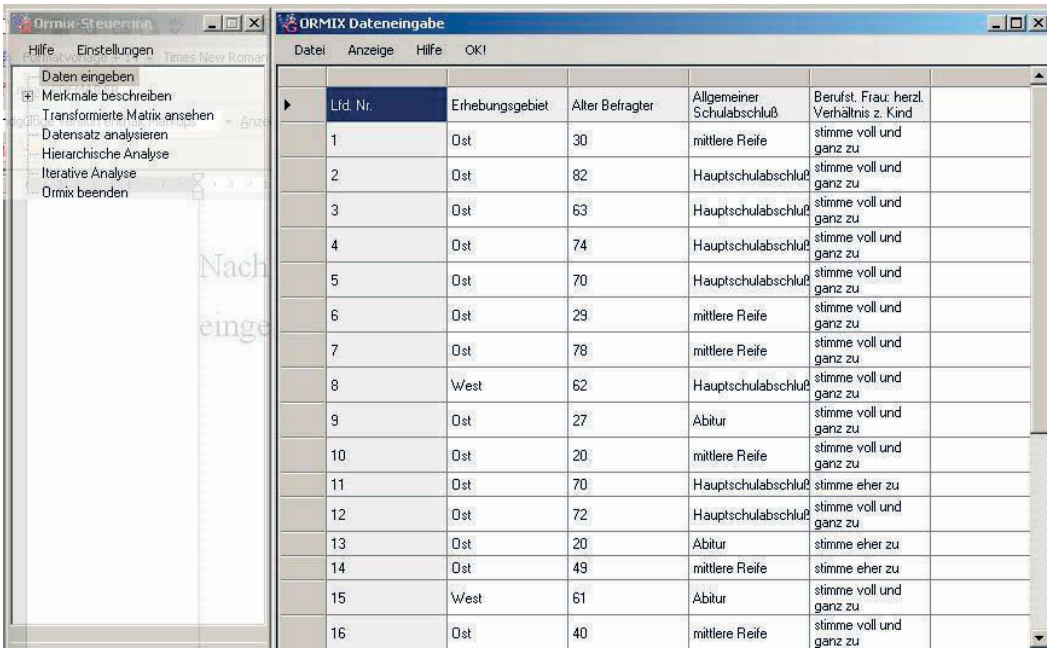
Durch Merkmalstransformationen, z.B. durch die Transformation metrischer in ordinale Merkmale, vor Beginn des Klassifikationsprozesses wird die Rohdatenmatrix in die Datenmatrix mit den Elementen x_{ji} überführt, die der Klassifikation der Merkmalsträger zugrundeliegt.

Wie diese Transformation durchgeführt wird, wird im folgenden Kapitel detailliert erläutert.

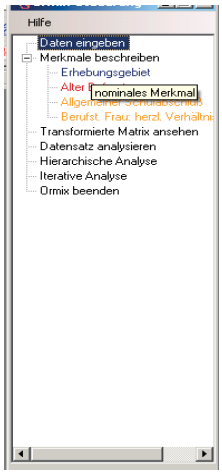
4 Die Erzeugung der Klassen

4.1 Datentransformationen

Nach dem Aufruf des Programms ORMIX wurde zunächst die Datenmatrix eingelesen oder eingegeben.



Die Dateneingabe ist mit "OK" zu bestätigen. In der ORMIX-Steuerung kann dann - mit Hilfe des Mauszeigers - die Voreinstellung für den jeweiligen Merkmalstyp abgelesen werden.



Für den jeweiligen Merkmalstyp gibt es Voreinstellungen.

Ob ein Merkmal bei der Klassifikation als nominales, ordinales oder metrisches Merkmal verarbeitet wird, kann in der Ablaufsteuerung - unabhängig von der Dateneingabe - festgelegt werden (siehe "Merkmale beschreiben").

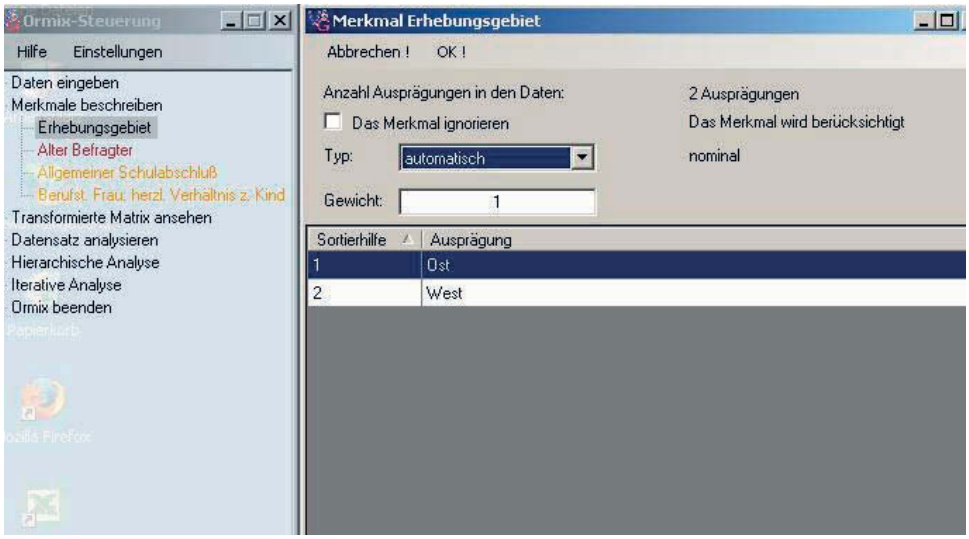
Mit einer Heuristik versucht ORMIX jedoch, einen geeigneten Merkmalstyp zu vermuten, um die Arbeit des Nutzers zu erleichtern:

- wenn alle Ausprägungen eines Merkmals numerisch und ganzzahlig geschrieben sind, vermutet ORMIX, dass es sich um ein ordinales Merkmal handelt,
- wenn alle Ausprägungen eines Merkmals numerisch sind und mindestens eine Ausprägung als nichtganzzahliger Wert geschrieben ist (zum Beispiel mit Nachkommastellen, wie in "1,0" oder in wissenschaftlicher Schreibweise "3E4"), vermutet ORMIX, dass es sich um ein metrisches Merkmal handelt,
- wenn mindestens eine Ausprägung nichtnumerisch ist, vermutet ORMIX, dass es sich um ein nominales Merkmal handelt.

Das ermöglicht, bereits bei der Dateneingabe die Beschreibung der Merkmale vorzubereiten und zu erleichtern.

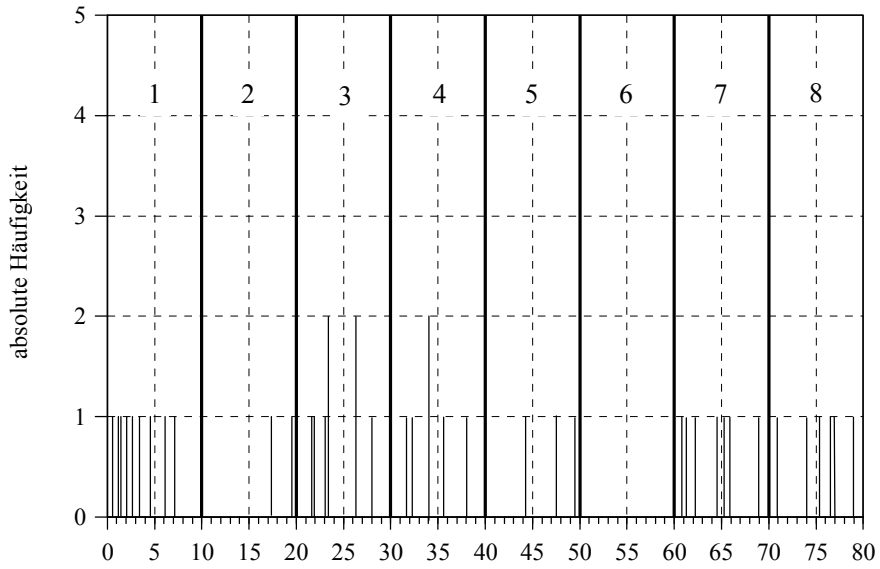
Geht man mit dem Mauszeiger in der ORMIX-Steuerung auf die einzelnen Merkmale, so wird angezeigt, ob es sich um ein nominales, ordinales oder me-trisches Merkmal handelt. Ein Anklicken mit der linken Maustaste öffnet eine Maske, in der der automatisch vorgegebene Merkmalstyp verändert werden kann. Die Transformation ist gegebenenfalls mit "OK" zu bestätigen.

Mit Hilfe der Spalte "Sortierhilfe" kann die Reihenfolge der Merkmalsausprägungen verändert werden, was vor allem bei ordinalen Merkmalen von Bedeutung ist. Die Werte in dieser Spalte können beliebig verändert werden, um dann mit Anklicken der Spaltenüberschrift „Sortierhilfe“ die Zeilen entsprechend aufsteigend bzw. absteigend zu sortieren. Beim Schließen der Maske mit OK! bleibt die Sortierfolge erhalten. Die Werte der Sortierhilfe gehen verloren: beim nächsten Öffnen der Maske sind die Zeilen beginnend bei 1 neu numeriert.



Das Alter des Befragten ist ein metrisches Merkmal. Es muss in ein ordinales Merkmal transformiert werden. Angenommen, das metrische Merkmal habe einen Wertebereich von 0 bis 80. Dann könnten z.B. wie folgt acht äquidistante Intervalle gebildet werden: 0 bis unter 10, 10 bis unter 20 usf.

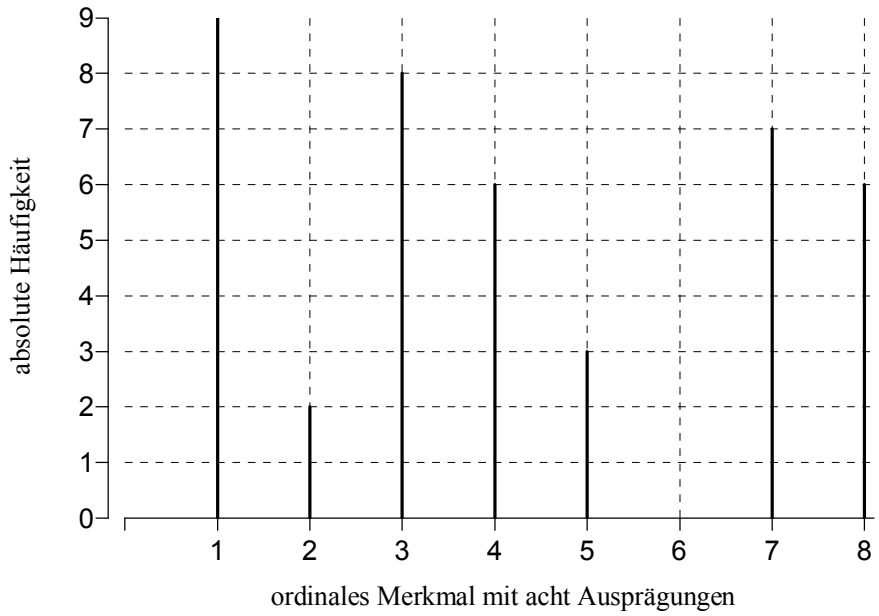
Häufigkeitsverteilung eines metrischen Merkmals



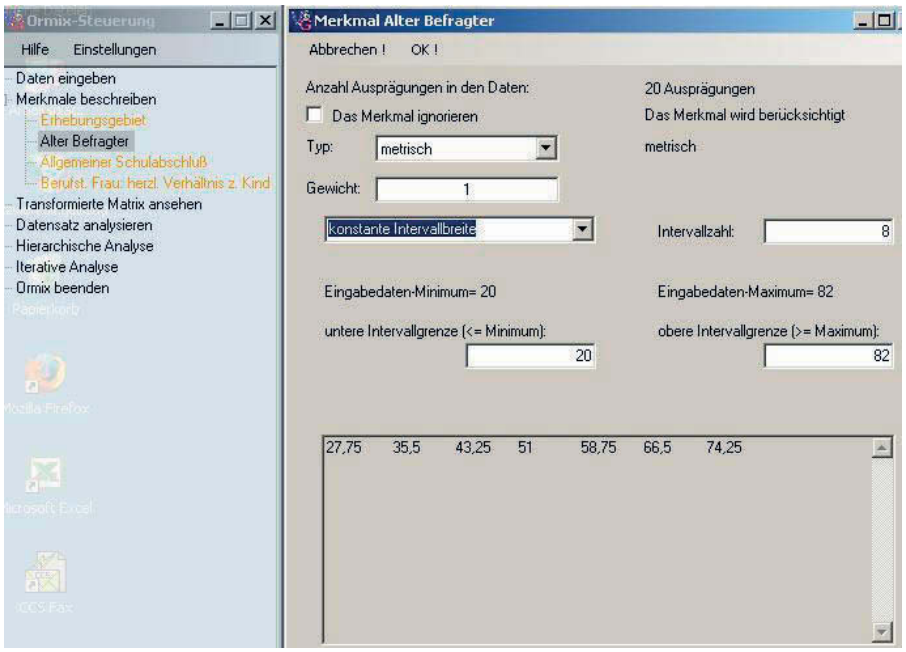
Nun wird jedem Merkmalsträger, dessen metrische Merkmalsausprägung Element des ersten Intervalls ist, die ordinale Ausprägung "1" zugeordnet. Jedem Merkmalsträger, dessen metrische Merkmalsausprägung Element des

zweiten Intervalls ist, wird die ordinale Ausprägung "2" zugeordnet usf., so dass ein ordinales Merkmal mit acht Ausprägungen entsteht.

Häufigkeitsverteilung des zugehörigen ordinalen Merkmals



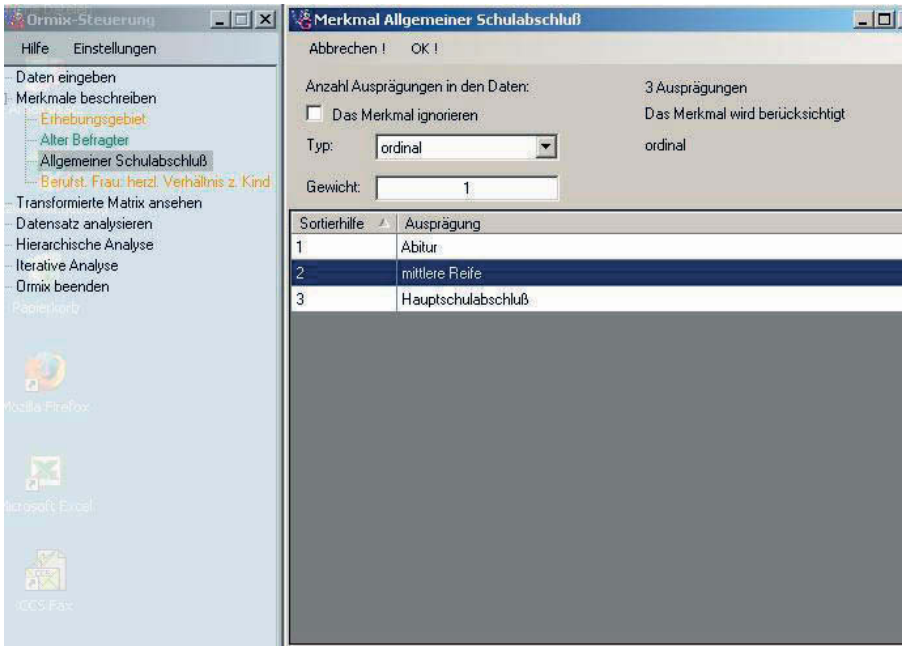
Diese Vorgehensweise wird in ORMIX folgendermaßen realisiert.



Dabei ist der Typ "metrisch" auszuwählen, falls ORMIX diesen Merkmalstyp nicht automatisch erkennt. Für metrische Merkmale können die Intervalle

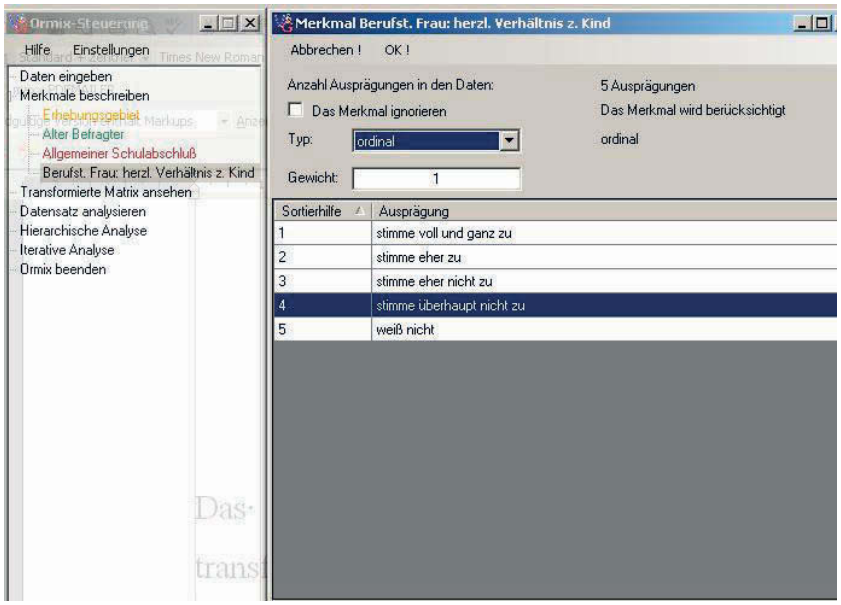
nach der Auswahl von „konstante Intervallbreite“ oder „gleichmäßig wachsende Intervallbreite“ über die Intervallzahl[16)], die untere Intervallgrenze und die obere Intervallgrenze auf einfache Weise festgelegt werden. Die resultierenden (Intervallzahl-1) Intervallgrenzen werden angezeigt. Die Auswahl „manuelle Eingabe der Intervallgrenzen“ ermöglicht, diese Liste direkt zu bearbeiten. Die Transformation ist mit "OK" zu bestätigen.

Das dritte Merkmal ist ordinal. Die Ausprägungen können mit der Sortierhilfe in die gewünschte Reihenfolge gebracht werden, z.B.: 1-Abitur, 2-Mittlere Reife, 3-Hauptschulabschluß.

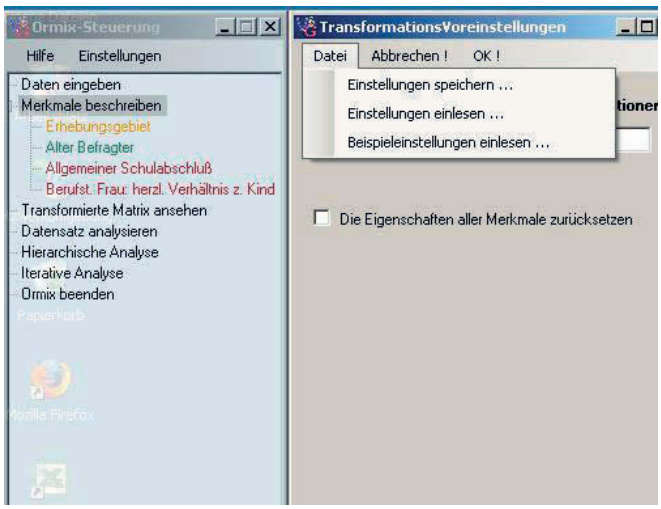


Das vierte und letzte Merkmal ist ebenfalls ordinal. Es wird wie folgt transformiert und - mit der Sortierhilfe - aufsteigend angeordnet.

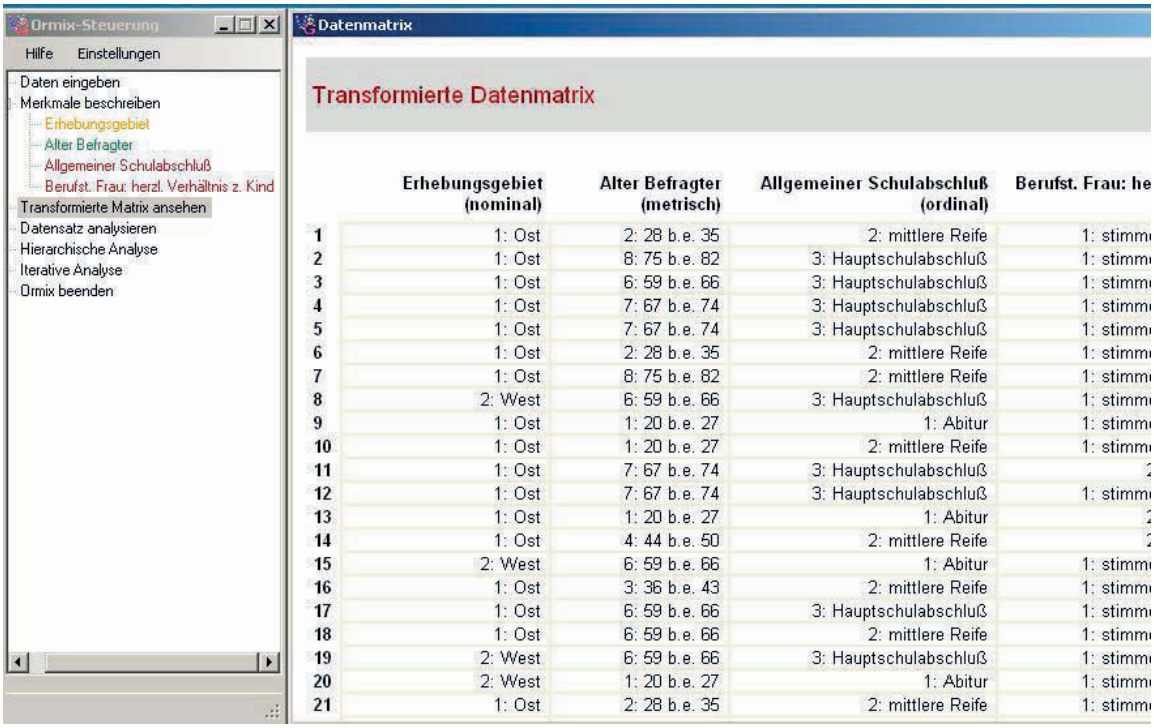
16) Wenn für die Intervallzahl 0 angegeben ist, so wird eine Voreinstellung gewählt. Diese Voreinstellung kann über ein gesondertes Fenster, das sich nach Anklicken von „Merkmale beschreiben“ in der Ablaufsteuerung öffnet, auch noch nachträglich verändert werden. Gegebenenfalls ist zuvor ein anderes offenes Einstellungsfenster durch „OK!“ oder „Abbrechen!“ zu schließene



Nach dem Eingeben der Transformationsregeln ist es ratsam diese Regeln in einer Datei zu speichern, um sie bei Bedarf wieder einlesen zu können. Mit Anklicken von "Merkmale bearbeiten" öffnet sich ein Fenster, in dessen Menüpunkt "Datei" die dazu erforderlichen Anweisungen zur Verfügung stehen.

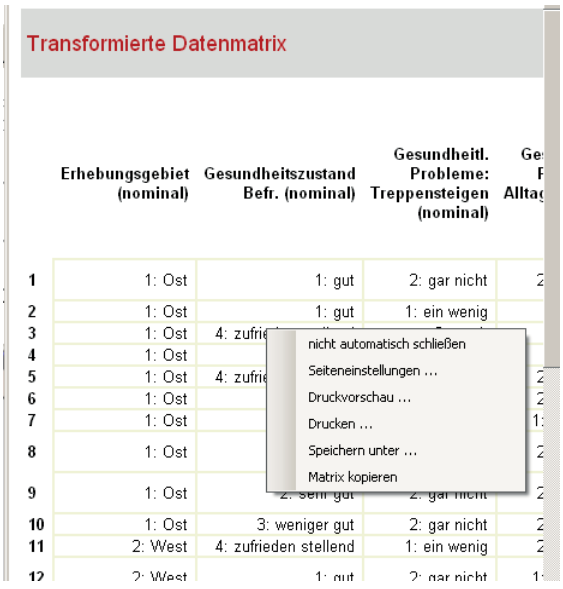


Anschließend sollte man sich die "Transformierte Datenmatrix" ansehen und gegebenenfalls den einen oder anderen der vorhergehenden Schritte wiederholen.



	Erhebungsgebiet (nominal)	Alter Befragter (metrisch)	Allgemeiner Schulabschluß (ordinal)	Berufst. Frau: he
1	1: Ost	2: 28 b.e. 35	2: mittlere Reife	1: stimm
2	1: Ost	8: 75 b.e. 82	3: Hauptschulabschluß	1: stimm
3	1: Ost	6: 59 b.e. 66	3: Hauptschulabschluß	1: stimm
4	1: Ost	7: 67 b.e. 74	3: Hauptschulabschluß	1: stimm
5	1: Ost	7: 67 b.e. 74	3: Hauptschulabschluß	1: stimm
6	1: Ost	2: 28 b.e. 35	2: mittlere Reife	1: stimm
7	1: Ost	8: 75 b.e. 82	2: mittlere Reife	1: stimm
8	2: West	6: 59 b.e. 66	3: Hauptschulabschluß	1: stimm
9	1: Ost	1: 20 b.e. 27	1: Abitur	1: stimm
10	1: Ost	1: 20 b.e. 27	2: mittlere Reife	1: stimm
11	1: Ost	7: 67 b.e. 74	3: Hauptschulabschluß	2
12	1: Ost	7: 67 b.e. 74	3: Hauptschulabschluß	1: stimm
13	1: Ost	1: 20 b.e. 27	1: Abitur	2
14	1: Ost	4: 44 b.e. 50	2: mittlere Reife	2
15	2: West	6: 59 b.e. 66	1: Abitur	1: stimm
16	1: Ost	3: 36 b.e. 43	2: mittlere Reife	1: stimm
17	1: Ost	6: 59 b.e. 66	3: Hauptschulabschluß	1: stimm
18	1: Ost	6: 59 b.e. 66	2: mittlere Reife	1: stimm
19	2: West	6: 59 b.e. 66	3: Hauptschulabschluß	1: stimm
20	2: West	1: 20 b.e. 27	1: Abitur	1: stimm
21	1: Ost	2: 28 b.e. 35	2: mittlere Reife	1: stimm

Ein Klick mit der rechten Maustaste öffnet ein kleines Fenster, welches einige Optionen für die transformierte Datenmatrix enthält (das gilt auch für andere Ergebnisse).

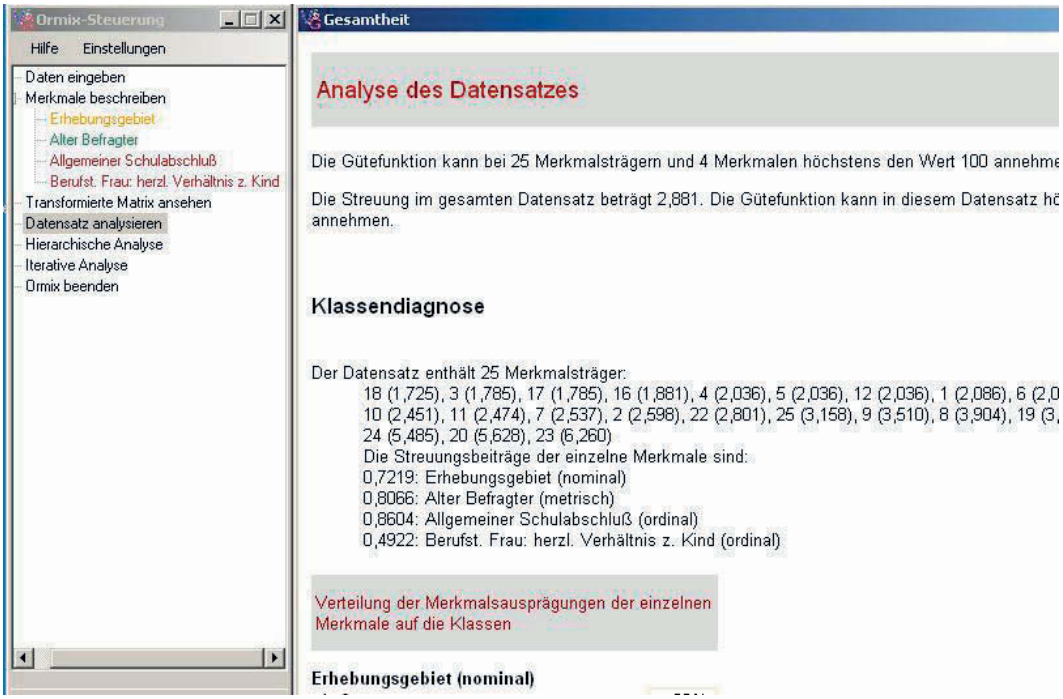


	Erhebungsgebiet (nominal)	Gesundheitszustand Befr. (nominal)	Gesundheitl. Probleme: Treppensteigen (nominal)	Ge... Alltag
1	1: Ost	1: gut	2: gar nicht	2
2	1: Ost	1: gut	1: ein wenig	
3	1: Ost	4: zufrie...		
4	1: Ost			
5	1: Ost	4: zufrie...		2
6	1: Ost			2
7	1: Ost			1:
8	1: Ost			2
9	1: Ost			2
10	1: Ost	3: weniger gut	2: gar nicht	2
11	2: West	4: zufrieden stellend	1: ein wenig	2
12	2: West	1: gut	2: gar nicht	1

Eine Besonderheit ist „nicht automatisch schließen“. Diese Einstellung entkoppelt das Ergebnisfenster von den Daten. Wenn anschließend die Daten oder Merkmalseinstellungen verändert werden, bleibt dieses Ergebnisfenster

geöffnet, obwohl die darin angezeigten Werte den geänderten Daten nicht mehr entsprechen. Ein Fenster mit den aktuellen Ergebnissen kann durch Klick auf den entsprechenden Punkt in der Ablaufsteuerung zusätzlich geöffnet werden. So können die Ergebnisse zu verschiedenen Daten gleichzeitig angezeigt und miteinander verglichen werden.

Beim Anklicken von "Datensatz analysieren" in der ORMIX-Steuerung werden alle Merkmalsträger zu einer Klasse zusammengefasst und ein Überblick über wichtige Parameter sowie die Merkmale und ihre Ausprägungen gegeben.

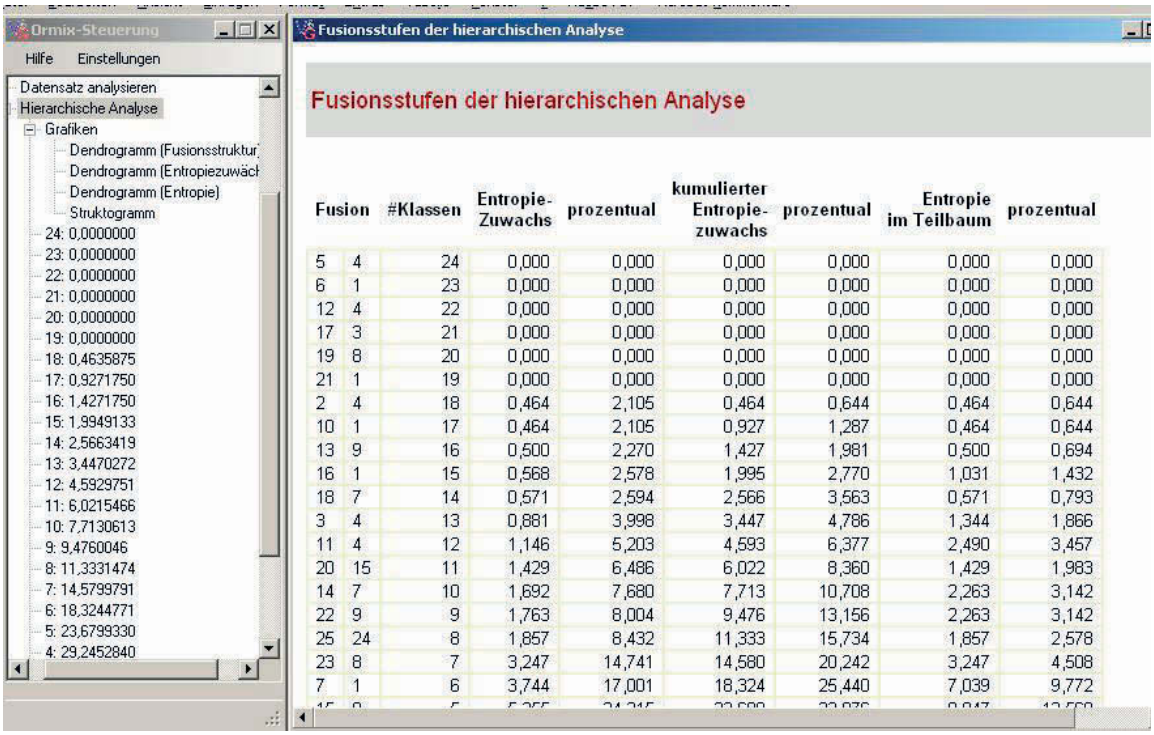


In der Klassendiagnose (das gilt auch für die einzelnen Klassen) sind die - in diesem Beispiel durchnummerierten - Merkmalsträger nach den Zahlen in Klammern aufsteigend geordnet. Je kleiner die Zahl in Klammern ist, desto „typischer/charakteristischer“ ist der zugehörige Merkmalsträger für den betrachteten Datensatz bzw. die betrachtete Klasse.[17)]

17) Genauer: Der typische Merkmalsträger einer Klasse ist dadurch definiert, dass bei seiner Entfernung aus der Klasse der Wert der gemeinsamen Streuung aller Merkmale in der so verkleinerten Klasse größer

4.2 Hierarchisch-agglomerative Klassifikation

Durch Mausklick mit der linken Maustaste auf " Hierarchische Analyse " in der ORMIX-Steuerung wird die hierarchisch-agglomerative Klassifikation gestartet.



Fusion		#Klassen	Entropie-Zuwachs	prozentual	kumulierter Entropie-zuwachs	prozentual	Entropie im Teilbaum	prozentual
5	4	24	0,000	0,000	0,000	0,000	0,000	0,000
6	1	23	0,000	0,000	0,000	0,000	0,000	0,000
12	4	22	0,000	0,000	0,000	0,000	0,000	0,000
17	3	21	0,000	0,000	0,000	0,000	0,000	0,000
19	8	20	0,000	0,000	0,000	0,000	0,000	0,000
21	1	19	0,000	0,000	0,000	0,000	0,000	0,000
2	4	18	0,464	2,105	0,464	0,644	0,464	0,644
10	1	17	0,464	2,105	0,927	1,287	0,464	0,644
13	9	16	0,500	2,270	1,427	1,981	0,500	0,694
16	1	15	0,568	2,578	1,995	2,770	1,031	1,432
18	7	14	0,571	2,594	2,566	3,563	0,571	0,793
3	4	13	0,881	3,998	3,447	4,786	1,344	1,866
11	4	12	1,146	5,203	4,593	6,377	2,490	3,457
20	15	11	1,429	6,486	6,022	8,360	1,429	1,983
14	7	10	1,692	7,680	7,713	10,708	2,263	3,142
22	9	9	1,763	8,004	9,476	13,156	2,263	3,142
25	24	8	1,857	8,432	11,333	15,734	1,857	2,578
23	8	7	3,247	14,741	14,580	20,242	3,247	4,508
7	1	6	3,744	17,001	18,324	25,440	7,039	9,772

Es wird tabellarisch gezeigt, welche Klassen auf welcher Fusionsstufe fusioniert werden und wie die Entropie (Streuung) innerhalb der Klassen mit abnehmender Anzahl an Klassen (#Klassen) zunimmt. Auf der linken Seite - in der ORMIX-Steuerung - werden verschiedene Dendrogramme zur Ansicht angeboten. Das einfachste Dendrogramm zeigt die Fusionsstruktur. Es stellt dar, welche Merkmalsträger und/oder Klassen miteinander zu neuen Klassen verschmolzen werden.

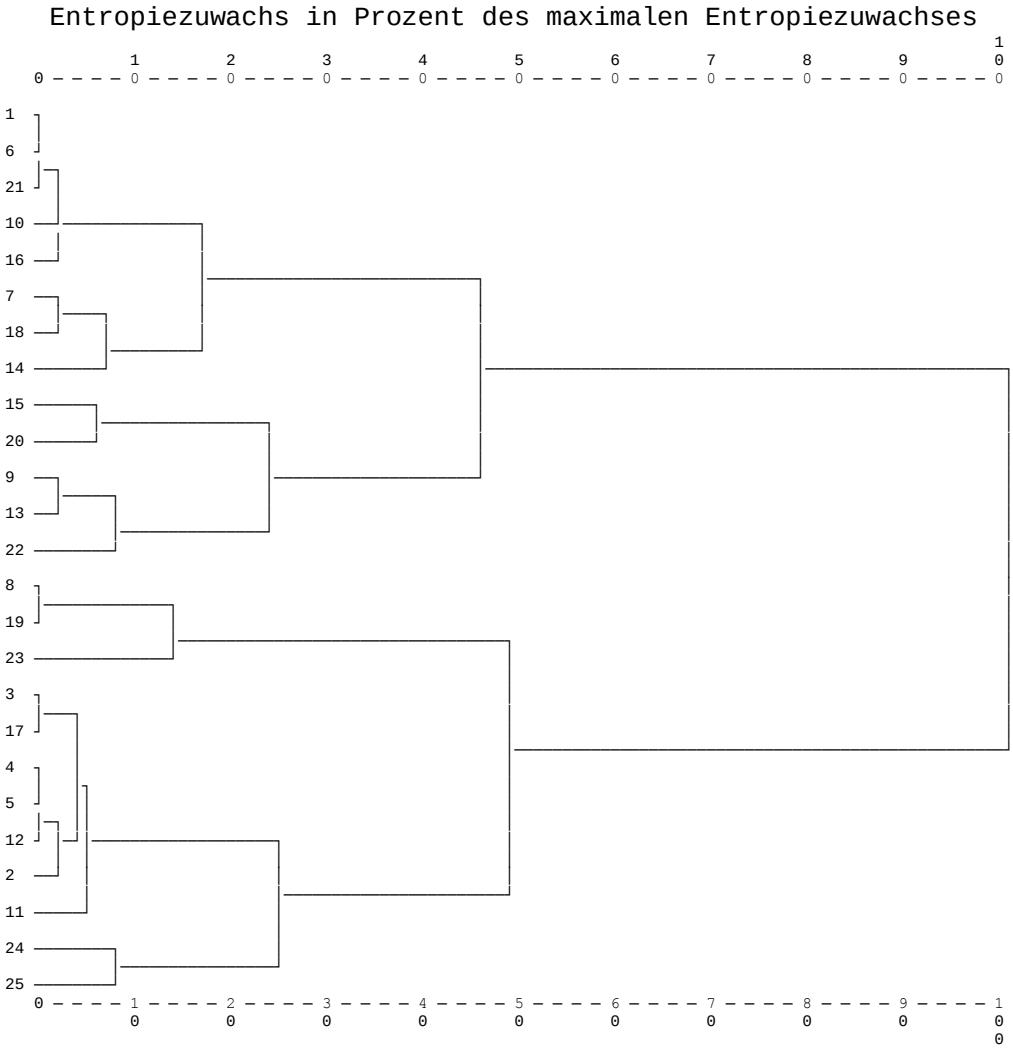
(oder höchstens gleich) ist als bei der Entfernung eines jeden anderen Merkmalsträgers. Es ist offensichtlich, dass es unter Umständen mehrere verschiedene typische Merkmalsträger in einer Klasse geben kann. Typische Merkmalsträger sind, ähnlich wie klassen-spezifische arithmetische Mittelwerte bei metrischen Merkmalen, gute Klassenrepräsentanten, die besonders dann hilfreich sind, wenn die verschiedenen Klassen einer Partition inhaltlich/realwissenschaftlich gegeneinander abgegrenzt werden sollen.

Informationen über die Entropieveränderungen kann man aus den Dendrogrammen ablesen, die die in der fünften bzw. in der letzten Tabellenspalte dargestellten prozentualen Veränderungen zeigen, aus dem „Dendrogramm (Entropiezuwächse)“ bzw. aus dem „Dendrogramm (Entropie)“.

Das „Dendrogramm (Entropiezuwächse)“ wird als Beispiel etwas ausführlicher beschrieben. Auf der oberen Skala ist der Entropiezuwachs (in Prozent des maximalen Entropiezuwachses), der aus der jeweiligen Fusion resultiert, abgetragen. Am linken Rand stehen die Merkmalsträgerkennungen.

Das Dendrogramm zeigt, dass der (kleine) Datensatz gut in zwei oder vier Klassen eingeteilt werden kann.

Dendrogramm

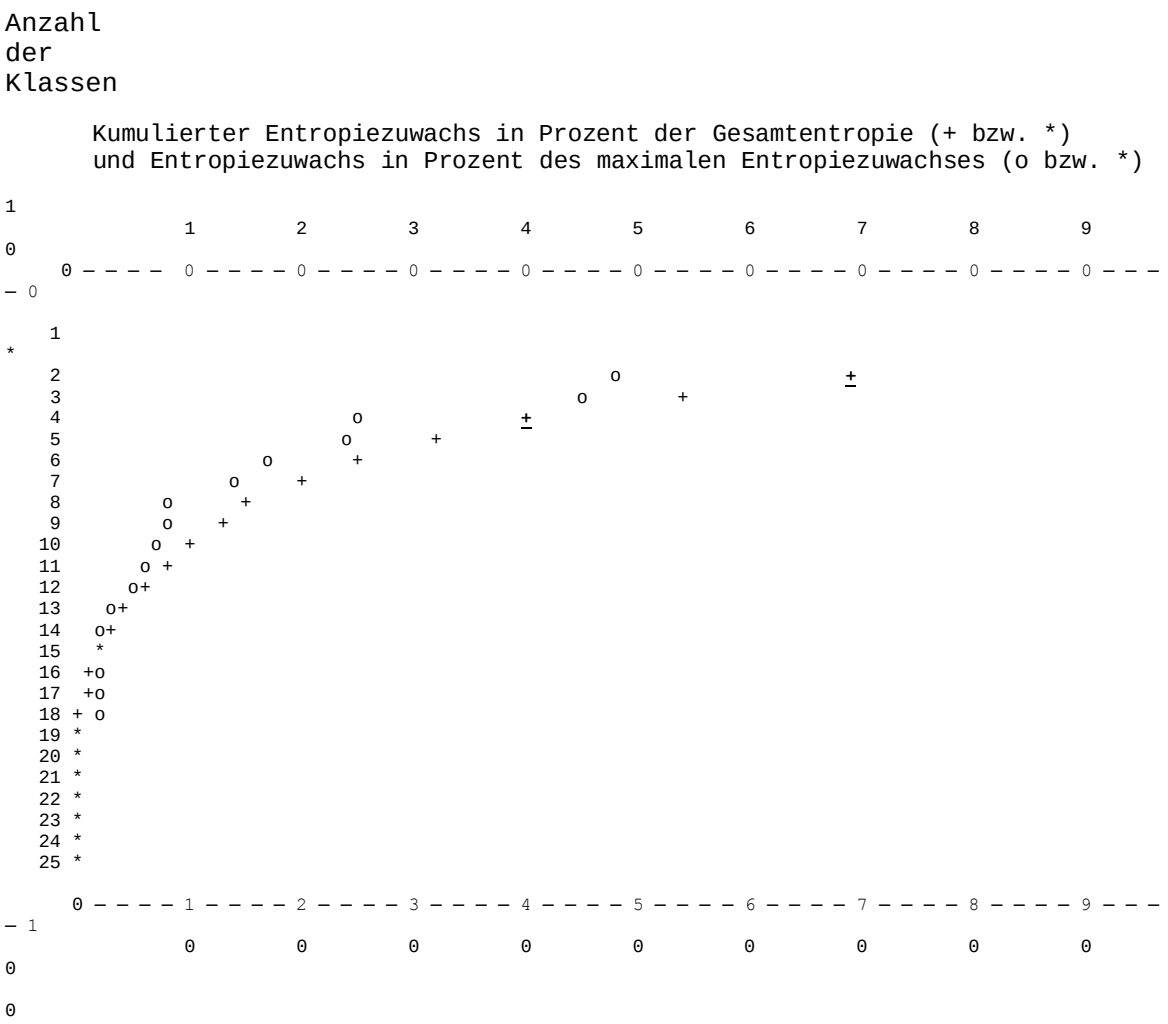


Bei großer Anzahl an Merkmalsträgern ist das zugehörige Dendrogramm nicht sehr übersichtlich. Durch Einstellen einer sehr kleinen Schriftgröße vor dem Anzeigen des Dendrogramms (über Einstellungen / Schriftarten einstellen / Schriftgröße für die Ergebnisse und hier z.b. 0,2 statt 0,8) kann dies zu Lasten der Lesbarkeit der Beschriftung ein wenig gelindert werden.

Bei großer Anzahl an Merkmalsträgern können Informationen über die Anzahl der zu bildenden Klassen im Allgemeinen besser dem Struktogramm entnommen werden. Im Unterschied zu den Dendrogrammen kann dieses den Ablauf der einzelnen Fusionsschritte und damit auch den über diese Schritte kumulierten Entropiezuwachs veranschaulichen.

Auch im Struktogramm zeigt es sich, dass der Datensatz gut in zwei Klassen oder in vier Klassen eingeteilt werden kann, da der Entropiezuwchs bei einer Klasse deutlich größer ist als bei zwei Klassen bzw. bei drei Klassen deutlich größer als bei vier Klassen. Bei der Einteilung in zwei Klassen (siehe $\pm$) sind 69,42% der Gesamtstreuung innerhalb der beiden Klassen (und folglich 30,58% der Gesamtstreuung zwischen den Klassen), bei einer Einteilung in vier Klassen (siehe $\pm$) sind 40,6% der Gesamtstreuung innerhalb der Klassen (und folglich 59,4% der Gesamtstreuung zwischen den Klassen).

Struktogramm

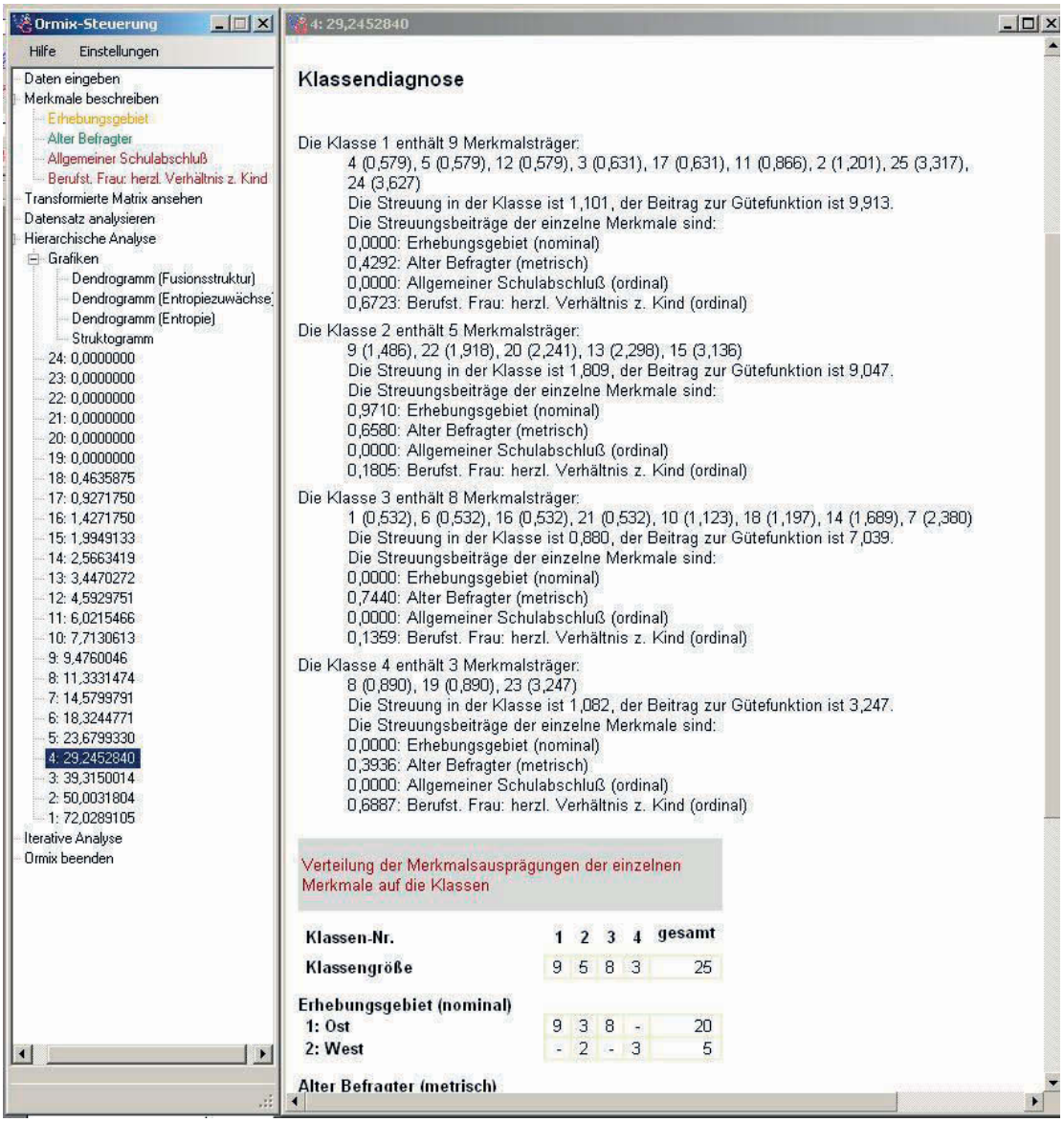


Wird in der ORMIX-Steuerung auf 4: 29,2452840 (vier Klassen) geklickt, öffnet sich ein Fenster mit dem "Klassifikationsergebnis des hierarchischen Verfahrens", das insbesondere die Klassendiagnose enthält.

Bemerkenswert erscheint unter anderem, dass von den vier Klassen drei Klassen ausschließlich von ost- bzw. westdeutschen Befragten besetzt sind. Auch beim Allgemeinen Schulabschluss zeigt sich eine deutliche Trennung der Klassen.

Hinweis: Bei vielen Merkmalsträgern und Merkmalen sind derartige Unterschiede zwischen den einzelnen Klassen eher ungewöhnlich!

Ergebnis des hierarchisch-agglomerativen Verfahrens



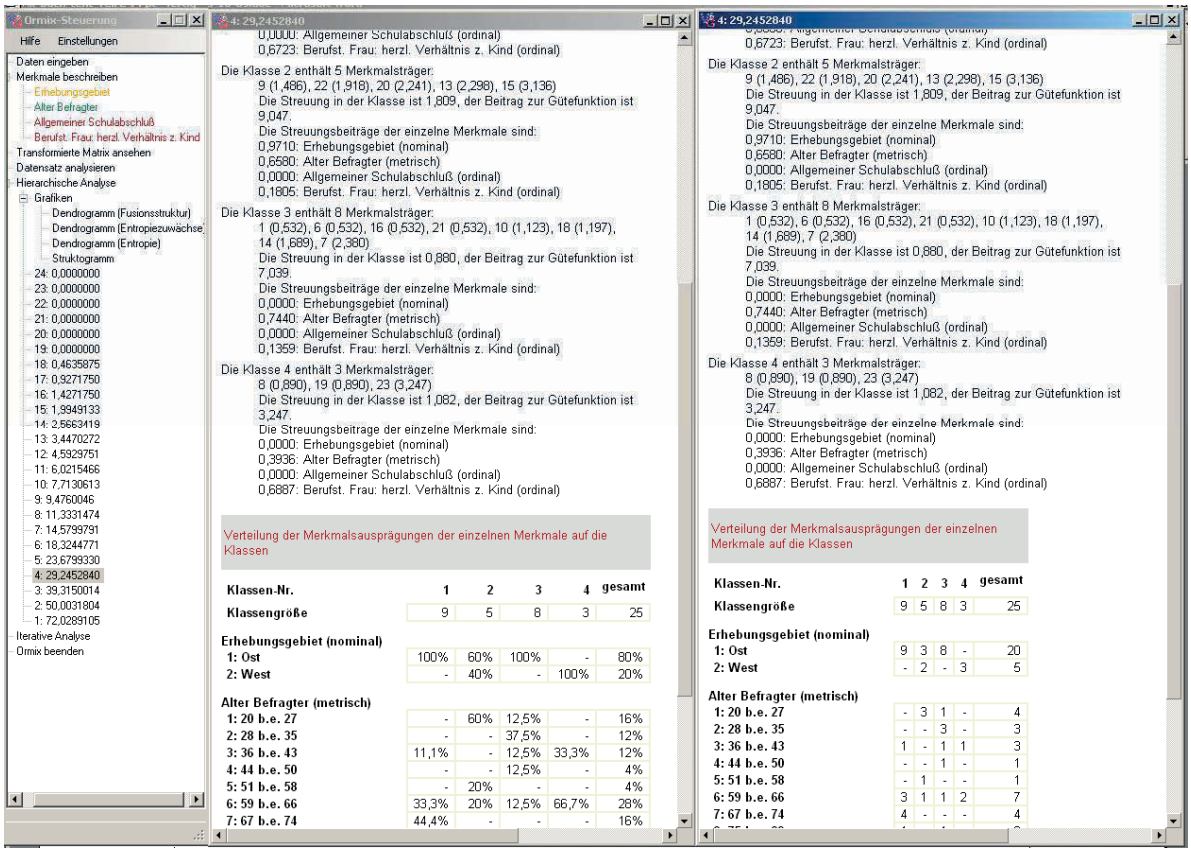
In der Klassendiagnose sind die Klassenbesetzungen aufgrund der Voreinstellung absolute Häufigkeiten. Für die Interpretation der Klassen sind aber häufig relative Häufigkeiten von Vorteil. Diese sind in ORMIX wie folgt darstellbar.

1. Rechter Mausklick in das Fenster mit der Klassendiagnose und „nicht automatisch schließen“ anklicken.
2. Fenster durch ziehen mit der linken Maustaste im blauen Bereich nach rechts verschieben.

3. In der ORMIX-Steuerung „Einstellungen“ anklicken und das Häkchen bei „Häufigkeiten absolut anzeigen“ durch anklicken entfernen.

4. Dann erneut in der ORMIX-Steuerung auf das Ergebnis „4: 28,72…“ klicken. Es erscheint die Klassendiagnose mit relativen Häufigkeiten, die - durch verschieben der Fenster - jener mit absoluten Häufigkeiten gegenübergestellt werden kann.

Klassendiagnose mit relativen und absoluten Häufigkeiten



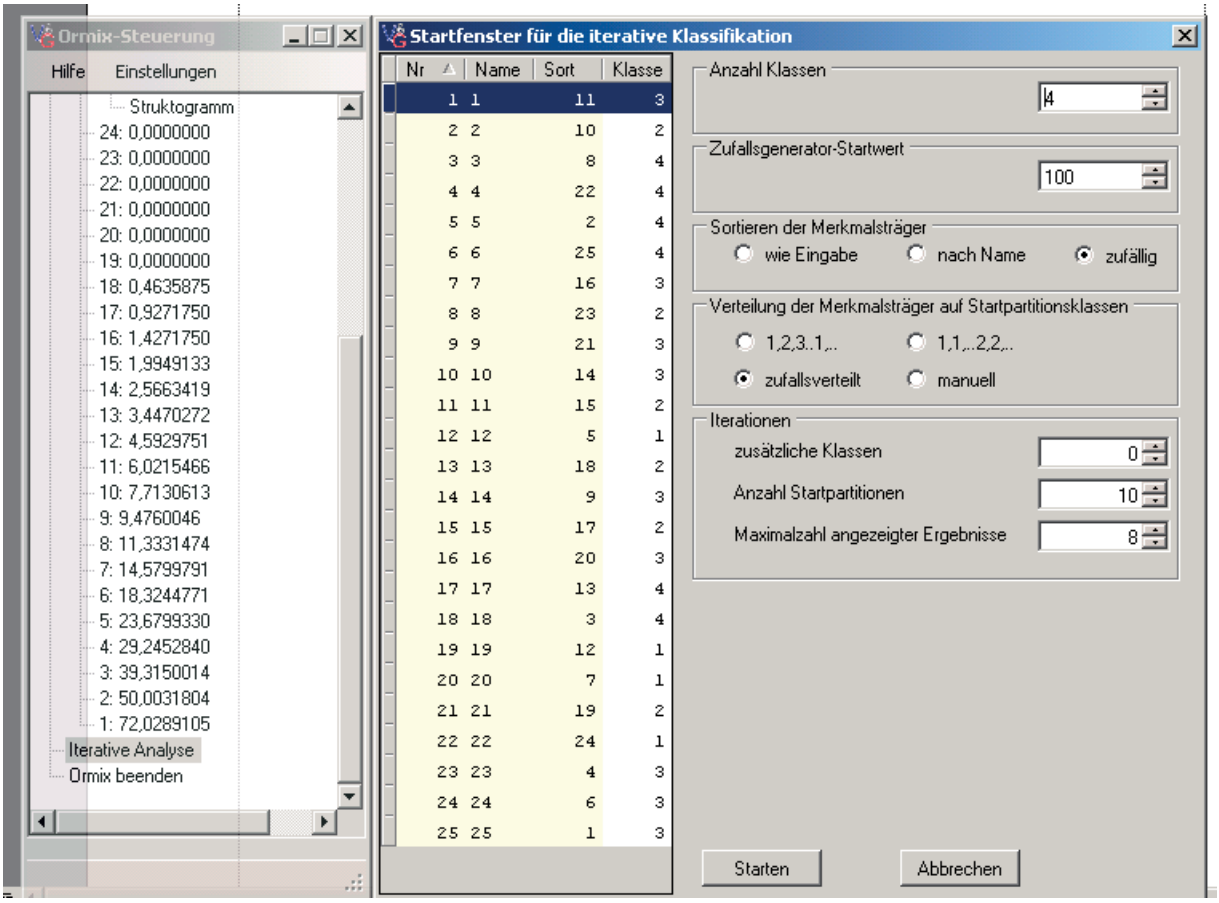
Nachdem es vernünftig erscheint, vier Klassen zu bilden (nur 40,6% der Gesamtstreuung ist innerhalb der Klassen!) ist es ratsam zu versuchen, mit dem iterativen Verfahren eine bessere Partition in vier Klassen zu erzeugen.

4.3 Iterative Klassifikation

Durch Mausklick mit der linken Maustaste auf "Iterative Analyse" in der ORMIX-Steuerung öffnet sich das folgende Fenster. Hier kann festgelegt werden, wie iterativ klassifiziert werden soll.

1. Festlegung der Anzahl der zu bildenden Klassen (die Voreinstellung: 5 ist für dieses Beispiel in 4 abzuändern),

2. Startwert des Zufallszahlengenerators (Voreinstellung: 100), dieser Startwert ist von Bedeutung, wenn ein Klassifikationsergebnis reproduziert werden soll,[18)]



3. Sortieren der Merkmalsträger: "wie Eingabe" - "nach Name" - "zufällig" (in der Regel ist zufällig die beste Lösung),

18) Der Startwert, mit dem die jeweilige Partition erzeugt wurde, kann in der Maske "Klassifikationsergebnis des iterativen Verfahrens" abgelesen werden (siehe Seite 62).

4. Verteilung der Merkmalsträger auf die Klassen der Startpartition (um Einflüsse dieser Verteilung auf die Güte des Klassifikationsergebnisses auszuschließen, ist "zufallsverteilt" im Allgemeinen die richtige Einstellung),

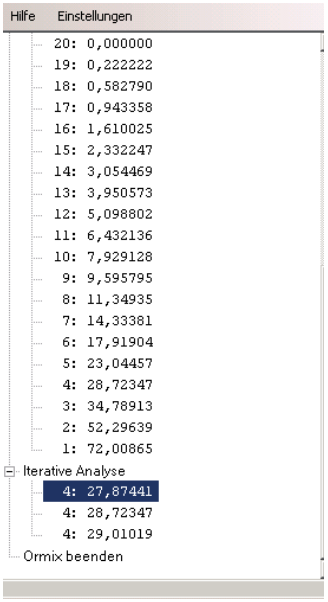
5. Iterationen: zusätzliche Klassen (Voreinstellung: 0). Wird hier eine Zahl grösser 0 eingesetzt, beispielsweise 3, dann wird der Datensatz nicht nur in vier, sondern auch in fünf, sechs und sieben Klassen eingeteilt.

6. Iterationen: Anzahl der Startpartitionen (Voreinstellung: 10). Die Anzahl der Startpartitionen kann beliebig erhöht werden. Mit zunehmender Anzahl der Startpartitionen und vor allem mit zunehmender Anzahl an Merkmals-trägern steigt die Rechenzeit erheblich. Allerdings ist die Wahrschein-lichkeit, dem "Optimum optimorum" nahezukommen, bei großer Anzahl an Startpartitionen grösser als bei einer kleinen Anzahl. In diesem kleinen Beispiel haben 10, 1.000 und 2.000 (verschiedene) Startpartitionen das gleiche Ergebnis erzeugt, was ein Zeichen dafür ist, dass diese Klasseneinteilung in vier Klassen auf diese Weise nicht mehr verbessert werden kann.

7. Iterationen: Maximalzahl angezeigter Ergebnisse. Hier kann festgelegt werden wieviele (verschiedene) Ergebnisse (mit Klassenanzahl und Wert der Gütefunktion) in der ORMIX-Steuerung angezeigt werden sollen (Voreinstellung. 8).

Mit den oben angegebenen Einstellungen wurde der Datensatz iterativ klassifiziert. Bei vier Klassen ergab sich für die Streuung innerhalb der Klassen ein Anteil von 38,71%. Gegenüber der mit Hilfe des hierarchisch-agglomerativen Verfahrens erzeugten Partition in vier Klassen mit einer Streuung innerhalb der Klassen von 39,89% ist das eine geringfügige Verbesserung, die jedoch zeigt, dass beim Aufbau der Hierarchie bei irgendeiner Stufe ein - im Sinne einer möglichst guten Partition - nicht optimaler Weg eingeschlagen wurde.

Bei der iterativen Analyse werden in der ORMIX-Steuerung nur die verschiedenen Ergebnisse angezeigt.

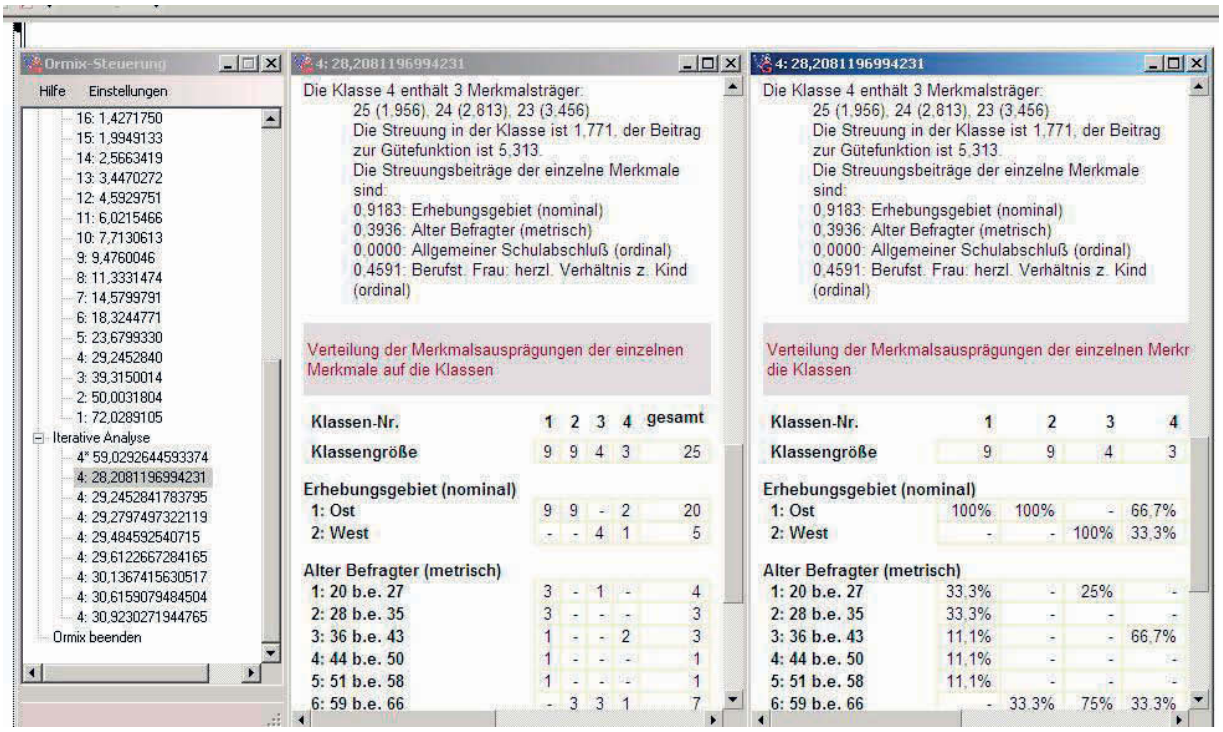


Ergebnis des iterativen Verfahrens

Die Klassendiagnose zeigt - was nach dem vorangegangenen allerdings zu erwarten war - dass auch beim iterativen Verfahren drei der vier Klassen aus-

schliesslich von ost- bzw. westdeutschen Befragten besetzt sind. In der Zusammensetzung der Klassen ergeben sich jedoch einige Unterschiede.

Klassendiagnose des iterativen Verfahrens mit absoluten und relativen Häufigkeiten



Im Zusammenhang mit dem Ergebnis des hierarchisch-agglomerativen Verfahrens (siehe oben) wurde detailliert beschrieben, wie die Klassendiagnose mit absoluten und relativen Häufigkeiten erzeugt werden kann.

Bemerkung: ORMIX wurde mit einem empirischen Datensatz (N=7.371 Merkmalsträger und M=28 Merkmalen) getestet und hat sich dabei auch für größere Datenmengen als stabil erwiesen.

Literaturhinweise

VOGEL, F., Zur Klassifikation von Einheiten anhand binärer Merkmale, Allgemeines Statistisches Archiv, 57(1973), S. 295-332.

VOGEL, F., Probleme und Verfahren der numerischen Klassifikation - Unter besonderer Berücksichtigung von Alternativmerkmalen, Vandenhoeck & Ruprecht, Göttingen 1975.

VOGEL, F., Subjektivitäten bei der Klassifikation von Einheiten, Proc. Op. Res. 7(1977), S.105-129.

DOBBENER, R., Zur Skalen- und Translationsinvarianz von Metriken, International Classification, 8 (1981), S. 64-68.

VOGEL, F./DOBBENER, R., Ein Streuungsmaß für komparative Merkmale, Jahrbücher für Nationalökonomie und Statistik, 197/2(1982), S. 145-157.

VOGEL, F., Zur Wirkungsweise und Leistungsfähigkeit eines iterativen Klassifikationsverfahrens für komparative Merkmale, Arbeiten aus der Statistik, Bamberg 1982.

DOBBENER, R., Ein heuristisches Verfahren zur Bewertung von Klassifikationsergebnissen im Hinblick auf natürliche Strukturen, Arbeiten aus der Statistik, Bamberg 1983.

DOBBENER, R., Grundlagen der Numerischen Klassifikation anhand gemischter Merkmale, Göttingen 1983.

VOGEL, F., Ein Verfahren zur Klassifikation von Merkmalsträgern anhand komparativer und gemischter Merkmale, International Classification, 10 (1983), S. 15 - 18.

VOGEL, F., Ein Beitrag zur Theorie 'weicher' Systeme: die Komparatisierung metrischer Merkmale, in: Beiträge zur Systemforschung, Festschrift für Adolf Adam, Springer, Wien - New York 1985, S. 314 - 333.

VOGEL, F., Ein Verfahren zur graphischen Darstellung von Klassifikationsergebnissen für komparative und gemischte Merkmale, in: Statistik zwischen Theorie und Praxis, Festschrift für Karl-August Schäffer, Hg. G. BUTTLER/ H. DICKMANN/ E. HELTEN/ VOGEL, F., Göttingen 1985, S. 280-289.

VOGEL, F., Unterentwicklung - Entwicklung, Eine Studie zur Einteilung der Länder der Erde nach ihrem Entwicklungsstand, Teil VI: Einteilung der Länder der Erde anhand von Merkmalen aus den Bereichen "Wirtschaft" und "Bevölkerung" und "Politik" sowie "Grundbedürfnisse und Soziales", Arbeiten aus der Statistik, Bamberg 1989.

VOGEL, F., Streuungsmessung ordinalskalierter Merkmale, Jahrbücher für Nationalökonomie und Statistik, 208/3(1991), S. 299-318.

VOGEL, F., Underdevelopment - Development, Report on a Study of Classification of the Countries of the World According to Their Stage of Development, Acta Demographica 1992, S. 237-252.

VOGEL, F., Some Remarks on a Classification of the Countries of the World According to their Stage of Development, Jahrbücher für Nationalökonomie und Statistik, 211(1993), S. 306-323.

VOGEL, F./ DOBBENER, R./ GRÜNEWALD, W., Hierarchisch-agglomerative Klassifikation von Merkmalsträgern, Programmpaket KOMIXH, Version 3 (PC-Version), Arbeiten aus der Statistik, Bamberg 1995.

VOGEL, F./ DOBBENER, R./ GRÜNEWALD, W., Iterative Klassifikation von Merkmalsträgern, Programmpaket KOMIXI, Version 6 (PC-Version), Arbeiten aus der Statistik, Bamberg 1995.

VOGEL, F., Hierarchisch-agglomerative Klassifikation von Merkmalsträgern, Programmpaket KOMIXH, Arbeiten aus der Statistik, Bamberg 2001.

VOGEL, F., Iterative Klassifikation von Merkmalsträgern, Programmpaket KOMIXI, Arbeiten aus der Statistik, Bamberg 2001.

VOGEL, F., Beschreibende und schließende Statistik, Formeln, Definitionen, Erläuterungen, Stichwörter und Tabellen, 13. Aufl., München 2005.

UNIVERSITY OF BAMBERG PRESS

Numerische Klassifikation (oder Cluster Analyse) ist die Zuordnung einer Menge von Beobachtungen (Objekten) zu Teilmengen (Klassen oder Cluster), derart dass die Beobachtungen (Objekte), die einer Klasse angehören, in einem bestimmten Sinne einander ähnlich sind.

Diese Arbeit besteht aus zwei Teilen: Der erste Teil behandelt die theoretischen Grundlagen unseres neuen Klassifikationsprogramms ORMIX. Zunächst werden zwei Verfahren zur Bildung disjunkter Klassen erörtert: ein Austauschverfahren und ein hierarchisch-agglomeratives Verfahren. Dann werden Maße zur Messung der Güte eines Klassifikationsergebnisses im Detail diskutiert, insbesondere im Hinblick auf die Merkmalstypen: nominal, ordinal und metrisch. Im Zusammenhang mit Problemen der Numerischen Klassifikation gibt es bei praktischen Anwendungen häufig gemischte Merkmale. Es wird gezeigt, wie eine Gütefunktion für gemischte Merkmale konstruiert werden kann.

Im zweiten Teil wird die Anwendung unseres Programms ORMIX beschrieben, das nominale, ordinale, metrische Merkmale und gemischte Merkmale verarbeiten kann. Die Konstruktion und das Einlesen der Datenmatrix wird im Detail erläutert. Dann wird gezeigt, wie Datentransformationen (beispielsweise metrische in ordinale Merkmale) durchgeführt werden können. Nach diesen Transformationen kann eine hierarchisch-agglomerative Klassifikation oder eine iterative Klassifikation durch einen gestartet werden. Einige Beispieldateien finden sich auf der CD.

Die Bedienung des Programms ist einfach und meist selbsterklärend. Es können Berechnungen angestoßen und aus einer knappen Auflistung der Resultate ausführliche Detaildarstellungen ausgewählt werden. Aus dem Wert einer Gütefunktion erhält man das Klassifikationsergebnis für die gewünschte Anzahl von Klassen mit einer detaillierten Klassendiagnose. Für die hierarchisch-agglomerative Klassifikation stehen zusätzlich Dendrogramme und ein Struktogramm zur Auswahl.

ISBN 978-3-923507-80-1
ISSN 1867-6197
40,- Euro