Informatik Forsch. Entw. (1995) 10: 91–98

Informatik
Forschung und Entwicklung

Analyse und Visualisierungshilfe für mehrdimensionale wissenschaftliche Daten

Holger Theisel

Institut für Computergraphik, Fachbereich Informatik, Universität Rostock, Postfach 999, D-18051 Rostock
(e-mail: theisel@informatik.uni-rostock.de)

Eingegangen am 8. Oktober 1993/Angenommen am 6. Oktober 1994

Zusammenfassung. Für die Visualisierung von wissenschaftlichen Daten existiert eine Vielzahl von Visualisierungstechniken, so daß es dem Nutzer oft schwerfällt, sich für eine für ihn günstige zu entscheiden. In dieser Arbeit wird ein Ansatz zur automatischen Auswahl geeigneter Techniken in Abhängigkeit vom gegebenen Datensatz vorgestellt. Der Ansatz basiert auf der Shannonschen Informationstheorie.

Schlüsselwörter: Visualisierung wissenschaftlicher Daten, Visualisierungstechnik, mehrdimensionale wissenschaftliche Daten, Informationstheorie.

Abstract. Given a multivariate data set there are a lot of different techniques to visualize it. In this paper a new approach is introduced to choose suitable techniques for given multivariate data automatically. The approach is based on Shannon's information theory.

Key words: Scientific visualization, visualization techniques, multidimensional scientific data, information theory.

CR Subject Classification: I.3, I.3.3, I.3.6

1. Einführung

Die Visualisierung mehrdimensionaler wissenschaftlicher Daten entwickelte sich in den letzten Jahren zu einem Forschungsschwerpunkt der modernen Computergraphik. Eine Vielzahl von Visualisierungstechniken wurde entwickelt, so daß es dem Nutzer schwerfällt, sich für eine für sein Problem günstige zu entscheiden.

Deshalb ist es wichtig, Systeme zu entwickeln, die dem Nutzer in Abhängigkeit von seinen Wünschen und der gegebenen Datenmenge passende Visualisierungstechniken vorschlagen. In der Literatur finden sich hierfür mehrere Ansätze. Verschiedene Herangehensweisen zur Problemlösung gehen von einer Analyse der Aufgabe, der Daten, der Interpretationsziele oder von einer Klassifizierung der Visualisierungsmethoden aus. Vorschläge zur Analyse der Aufgabe finden sich in [15] und [6]. In [6] erfolgt nach einer logischen Aufgabenbeschreibung die Ersetzung der logischen Operatoren durch elementare Wahrnehmungsoperatoren. Auf Grundlage dieser Wahrnehmungsoperatoren erfolgt die Visualisierung. In [15] wird eine Matrix gebildet, aus der für verschiedene Objektklassen und Interpretationsziele die dafür geeigneten Techniken gelesen werden können. Für die Analyse der Daten finden sich Ansätze in [3] und [5]. [4] geht von einer Analyse der Interpretations- bzw. „Visualisierungsziele" aus. Dabei wird zwischen drei großen Klassen von Visualisierungszielen unterschieden: *exploration, directed search* und *comparison.* In [11] wird eine Methodologie zur Auswahl von Visualisierungstechniken basierend auf einem natürlichen Szenenparadigma vorgestellt. [14] beschreibt eine Kopplung mit anderen, auch nichtvisuellen Formen der Analysetätigkeit. Insgesamt kann festgestellt werden, daß alle bisher existierenden Ansätze von konkreten Anwendungsfällen ausgehen. Das allgemeine Problem zur Auswahl geeigneter Techniken ist weitgehend ungelöst.

Diese Arbeit beschäftigt sich mit einer Teillösung des allgemeinen Problems. Aus der Vielzahl der möglichen Interpretationsziele wird ein spezielles ausgewählt: das Erkennen von Korrelationen zwischen den Datenwerten. Diesem Interpretationsziel kommt eine zentrale Bedeutung zu. Existieren Zusammenhänge zwischen verschiedenen Datenwerten, so hat der Nutzer in den meisten Fällen auch ein großes Interesse, diese in der Visualisierung erkennen zu können. Umgekehrt sollte bei Nichtvorhandensein von Zusammenhängen die Visualisierungstechnik so gewählt werden, daß dem Nutzer in der Visualisierung keine Zusammenhänge vorgetäuscht werden.

Gestützt auf das Interpretationsziel „Erkennen von Korrelationen zwischen den Datenwerten" soll ent-

schieden werden, welche Visualisierungstechniken für einen gegebenen Datensatz günstig sind und welche nicht. Dazu dient folgendes prinzipielles Vorgehen:

a) Analyse des zu visualisierenden Datensatzes:
Es soll automatisch ermittelt werden, ob Zusammenhänge im Datensatz existieren und wie diese aufgebaut sind. Gleichzeitig wird ermittelt, in welchen Teilen des Datensatzes viele Informationen stecken und welche Teile nur eine geringe Information enthalten.
b) Analyse aller betrachteten Visualisierungstechniken:
Für jede in Frage kommende Visualisierungstechnik muß geklärt werden, welche Art von Zusammenhängen mit ihr besonders gut visualisiert werden können und welche nicht.
c) Bewertung auf Grund der Erkenntnisse von a) und b), welche Visualisierungstechniken für den konkreten Datensatz günstig sind.

Die automatische Analyse des Datensatzes ist in Abschnitt 2 näher beschrieben. Die Analyse der betrachteten Visualisierungstechniken erfolgt für jede Technik nur einmal, wenn sie in das System aufgenommen wird. Die Ergebnisse der Analyse bilden eine Wissensbasis des Systems. Näheres dazu in Abschnitt 3. Abschnitt 4 beschreibt die Anwendung auf zwei konkrete Datensätze.

2. Die Analyse des Datensatzes

Ein gegebener mehrdimensionaler Datensatz kann aufgefaßt werden als eine Matrix M mit m Zeilen und n Spalten. m ist dabei die Anzahl der Beobachtungsfälle, n ist die Anzahl der Dimensionen. Für jeden der m Beobachtungsfälle werden also n Zustandsgrößen des Beobachtungsobjektes gemessen oder berechnet. Diese stehen dann in einer Zeile von M. Eine Spalte von M enthält die Werte einer Zustandsgröße für alle m Beobachtungsfälle.

Die Analyse von M soll mit Hilfe der Shannonschen Informationstheorie erfolgen. Dafür sollen zunächst die wichtigsten benötigten Begriffe und Zusammenhänge genannt werden.

2.1. Informationstheorie

Sei X eine diskrete Zufallsgröße, welche die Werte x_1, $x_2,\ldots,x_u$ mit den Wahrscheinlichkeiten $p(x_1)$, $p(x_2),\ldots,p(x_u)$ annimmt.

$$H(X) := \sum_{i=1}^{u} p(x_i) \cdot log_2 \frac{1}{p(x_i)}$$

sei die Information, welche aus der einmaligen Ausführung des Versuchs X entsteht. $H(X)$ wird auch Entropie (Unbestimmtheit, Ungewißheit) von X genannt: Die durch die Ausführung des Versuchs X gewonnene Information ist gleich der dadurch beseitigten Ungewißheit. Die Einheit für $H(X)$ ist das Bit.

Sei Y eine Zufallsgröße, welche die Werte $y_1, y_2,\ldots,y_v$ annimmt. (X,Y) sei die Zufallsgröße, die durch gemeinsame Beobachtung von X und Y entsteht. (X,Y) nimmt also die Werte $(x_1, y_1),(x_1, y_2),\ldots,(x_1, y_v)$, $(x_2, y_1),\ldots,$ (x_u, y_v) an. Dies ergibt

$$H(X,Y) = \sum_{i=1}^{u} \sum_{j=1}^{v} p(x_i,y_j) \cdot log_2 \frac{1}{p(x_i,y_j)}$$

Analog läßt sich die Entropie auf eine beliebige Anzahl von diskreten Zufallsgrößen erweitern.

Folgende Eigenschaften gelten für beliebige Zufallsgrößen $X_1,\ldots,X_q$, X, Y und eine beliebige Permutation π von $(1,\ldots,q)$:

- $H(X_1,\ldots,X_q) \geq 0$
- $H(X_1, X_1, X_2,\ldots,X_q) = H(X_1, X_2,\ldots,X_q)$

(speziell $H(X, X) = H(X)$)

- $H(X_1,\ldots,X_q) = H(X_{\pi(1)},\ldots,X_{\pi(q)})$
- $H(X,Y) \leq H(X) + H(Y)$
- $H(X,Y) = H(X) + H(Y) \Leftrightarrow X,Y$ unabhängig.

Die letzten beiden Eigenschaften führen zur Einführung des Begriffs der gemeinsamen Information I der Zufallsgrößen X und Y. Im allgemeinen Fall (X und Y nicht unabhängig) erhält man durch Ausführung des Versuchs X bereits einen Teil der Information über Y. Dieser Teil sei die gemeinsame Information I von X und Y. Es gilt $I(X,Y) = H(X) + H(Y) - H(X,Y)$ [8, 13].

Der Begriff der gemeinsamen Information soll hier verallgemeinert werden auf q diskrete Zufallsgrößen $X_1,\ldots,X_q$. Dies ist nötig, um bei der Analyse von Datensätzen auch Korrelationen von mehr als zwei Dimensionen ermitteln zu können.

$$\begin{aligned} I'(X_1,\ldots,X_q) := {} & H(X_1) + \ldots + H(X_q) \\ & - H(X_1, X_2) - H(X_1, X_3) - \ldots - \\ & - H(X_{q-1}, X_q) \\ & + H(X_1,X_2,X_3) + H(X_1,X_2,X_4) + \ldots + \\ & + H(X_{q-2}, X_{q-1}, X_q) \\ & \vdots \\ & \pm H(X_1,\ldots,X_q) \end{aligned}$$

$$I(X_1,\ldots,X_q) := \begin{cases} I'(X_1,\ldots,X_q), & \text{falls } I'(X_1,\ldots,X_q) > 0 \\ & \text{und } I(X_2,\ldots,X_q) > 0 \\ & \text{und } I(X_1,X_3,\ldots,X_q) > 0 \\ & \vdots \\ & \text{und } I(X_1,\ldots,X_q-1) > 0 \\ 0 & \text{sonst} \end{cases}$$

$I(X_1,\ldots,X_q)$ sei die gemeinsame Information der Zufallsgrößen $X_1,\ldots,X_q$. Folgende Eigenschaften gelten für beliebige diskrete Zufallsgrößen $X_1,\ldots,X_q$, X, Y und eine beliebige Permutation π von $(1,\ldots,q)$:

- $I(X_1,\ldots,X_q) \geq 0$
- $I(X_1, X_1,\ldots,X_q) = I(X_1,\ldots,X_q)$
- $I(X_1,\ldots,X_q) = I(X_{\pi(1)},\ldots,X_{\pi(q)})$
- $I(X_2,\ldots,X_q) \geq I(X_1, X_2,\ldots,X_q)$
- $I(X,Y) = 0 \Leftrightarrow X,Y$ unabhängig.

Mit diesen eingeführten Grundbegriffen sollen nun mehrdimensionale Datensätze analysiert werden.

2.2. Anwendung der Informationstheorie auf die Analyse von M

Um die Informationstheorie auf die Analyse von M anzuwenden, müssen folgende Voraussetzungen getroffen werden:

– Jeder Beobachtungsfall wird als zufälliges Ereignis betrachtet.
– Jede Dimension wird als Zufallsgröße betrachtet. Jeder der m Beobachtungsfälle ist also die gleichzeitige Realisierung von n Zufallsgrößen. Jede Spalte von M enthält m Realisierungen einer Zufallsgröße. Die i-te Dimension ist also eine Zufallsgröße, von der m Realisierungen in der i-ten Spalte von M stehen.
– Die Wahrscheinlichkeit eines Ereignisses sei gleich der relativen Häufigkeit dieses Ereignisses. Die Verteilung der i-ten Dimension ergibt sich also aus den Werten der i-ten Spalte von M.

Mit diesen Voraussetzungen kann die Analyse auf folgende Weise durchgeführt werden:

– Man ermittle alle Dimensionen, die eine hohe Entropie haben. Diese Dimensionen sind für eine Visualisierung besonders interessant und sollten hier entsprechend gut erkennbar sein.
– Man ermittle alle Tupel von Dimensionen, die eine relevante gemeinsame Information haben. Zwischen diesen Dimensionen existieren Zusammenhänge, die in der Visualisierung herausgearbeitet werden sollten.

Der Begriff der Relevanz soll sichern, daß bei verschwindend geringen gemeinsamen Informationen keine Zusammenhänge festgestellt werden. Die folgenden Ansätze für den Relevanzbegriff sind möglich:

– absolute Relevanz:
Seien $X_1,\ldots,X_q$ diskrete Zufallsgrößen. $I(X_1,\ldots,X_q)$ sei relevant bzgl. einer Grenze g, falls $I(X_1,\ldots,X_q) > g$ gilt.
g ist eine beliebige nichtnegative Zahl, die die Bitzahl angibt, ab welcher eine gemeinsame Information nicht mehr relevant ist.
– relative Relevanz:
Seien $X_1,\ldots,X_q$ diskrete Zufallsgrößen. $wb(X_i)$ sei die Anzahl der verschiedenen Werte, die X_i mit positiver Wahrscheinlichkeit annehmen kann. $I(X_1,\ldots,X_q)$ sei relevant bzgl. einer Grenze g, falls

$$\frac{I(X_1,\ldots,X_q)}{max\{log_2(wb(X_i)) : i \in \{1,\ldots,q\}\}} > g$$

gilt. Der Ausdruck im Nenner beschreibt dabei die maximal mögliche gemeinsame Information zwischen $X_1,\ldots,X_q$. Für g muß also $0 \leq g < 1$ gelten.
– Relevanz relativ zu den Teilinformationen:
$I(X_1,\ldots,X_q)$ mit $q > 1$ sei relevant bzgl. einer Grenze g, falls gilt:
1) $I(X_2,\ldots,X_q)$, $I(X_1,X_3,\ldots,X_q)$, …, $I(X_1,\ldots,X_{q-1})$ sind relevant.

2) $$\frac{I(X_1,\ldots,X_q)}{max\{I(X_2,\ldots,X_q), I(X_1,X_3,\ldots,X_q),\ldots,I(X_1,\ldots,X_{q-1})\}} > g$$

Für $q = 1$ wird explizit definiert, daß $I(X_1) = H(X_1)$ relevant ist, falls $H(X_1) > 0$ gilt. Für g gilt wieder $0 \leq g < 1$.

Diese Definition der Relevanz erweist sich als günstig, wenn von großen Datenmengen nur wenige positive Fälle interessant sind, gerade bei diesen aber weitere Zusammenhänge erkannt werden sollen.

Aus Komplexitätsgründen (für jedes Tupel von Dimensionen muß die gemeinsame Information berechnet werden) ist das bisherige Verfahren zur Analyse von M praktisch nicht ausführbar. Abhilfe schafft jedoch der folgende Algorithmus:

1) Man bestimme alle Dimensionen X_i, die eine relevante Information haben. Existieren keine solchen Dimensionen, so bricht der Algorithmus ab.
2) Man bestimme alle Paare von Dimensionen (X_i, X_j) mit $i \neq j$, die eine relevante gemeinsame Information haben. Dabei wird $I(X_i, X_j)$ nur berechnet, falls $I(X_i)$ und $I(X_j)$ relevant sind. Zur Berechnung von $I(X_i, X_j)$ kann $I(X_i) = H(X_i)$ und $I(X_j) = H(X_j)$ verwendet werden. Falls $I(X_i, X_j)$ relevant ist, werden $I(X_i, X_j)$ und $H(X_i, X_j)$ abgespeichert. Existieren keine solchen Paare, so bricht der Algorithmus ab.
3) Man bestimme alle Tripel von Dimensionen (X_i, X_j, X_k) mit i, j, k paarweise verschieden, die eine relevante gemeinsame Information haben. Dabei wird $I(X_i, X_j, X_k)$ nur berechnet, wenn $I(X_i, X_j)$, $I(X_i, X_k)$ und $I(X_j, X_k)$ relevant sind. Zur Berechnung von $I(X_i, X_j, X_k)$ kann $H(X_i)$, $H(X_j)$, $H(X_k)$, $H(X_i, X_j)$, $H(X_i, X_k)$ und $H(X_j, X_k)$ verwendet werden, da alle diese Werte bereits berechnet wurden. Ist $I(X_i, X_j, X_k)$ relevant, so werden $I(X_i, X_j, X_k)$ und $H(X_i, X_j, X_k)$ abgespeichert. Existieren keine solchen Tripel, so bricht der Algorithmus ab.
4) Man bestimme analog alle Quadrupel von Dimensionen mit relevanter gemeinsamer Information. Existieren keine solchen Quadrupel, so bricht der Algorithmus ab.
⋮
n) Man ermittle analog alle n-Tupel von Dimensionen mit relevanter gemeinsamer Information.

Dieser Algorithmus ermittelt alle Tupel von Dimensionen mit relevanter gemeinsamer Information. Die worst-case-Komplexität des Algorithmus ist zwar von exponentieller Ordnung bzgl. n, in praktischen Fällen (nicht alle Tupel von Dimensionen haben eine relevante gemeinsame Information) liegt aber die Rechenzeit in sinnvollen Bereichen (s. Abschn. 4).

2.3. Fehlerabschätzungen

Die Analyse von M basiert auf der Annahme, daß als Wahrscheinlichkeit für das Auftreten zufälliger Ereignisse deren relative Häufigkeit verwendet wird. Dies kann besonders bei kleinen Datenmengen zu solch großen Unsicherheiten der Aussagen führen, daß durch die Analyse gefundene Zusammenhänge nicht mehr als solche betrachtet werden können.

Sei X eine Zufallsgröße, welche die Werte $1,\ldots,k$ mit den Wahrscheinlichkeiten $p_1,\ldots,p_k$ annimmt.

Die Zufallsgröße Y sei definiert als

$$Y := \begin{cases} 1, \textit{falls}\, X = 1 \\ 0 \qquad \textit{sonst} \end{cases} \tag{1}$$

$Y_1,\ldots,Y_m$ seien Realisierungen der Zufallsgröße Y. Für Erwartungswert und Varianz gilt dann:

$$E(Y_1) = \ldots = E(Y_m) = 0 \cdot (1-p) + 1 \cdot p = p \tag{2}$$

$$V(Y_1) = \ldots = V(Y_m) = [0 - E(Y_1)]^2 \cdot (1-p) + [1 - E(Y_1)]^2 \cdot p = p \cdot (1-p) \tag{3}$$

Die Zufallsgröße Z_m sei definiert als

$$Z_m := \frac{1}{m} \cdot (Y_1 + \ldots + Y_m) = h_m(X = 1) \tag{4}$$

Für Erwartungswert und Varianz gilt hier:

$$E(Z_m) = p \tag{5}$$

$$V(Z_m) = \frac{1}{m} \cdot V(Y_1) = \frac{1}{m} \cdot p \cdot (1-p) \tag{6}$$

Auf die Zufallsgröße Z_m wird nun die Tschebyscheff'sche Ungleichung angewandt:

$$p(|Z_m - E(Z_m)| \geq \varepsilon \leq \frac{V(Z_m)}{\varepsilon^2} \tag{7}$$

Durch Ersetzen ergibt sich:

$$p(|h_m(X = 1) - p(X = 1)| \geq \varepsilon) \leq \frac{p \cdot (1-p)}{m \cdot \varepsilon^2} \tag{8}$$

Der Ausdruck $|h_m(X=1) - p(X=1)|$ ist ein Maß dafür, wie weit bei m Meßwerten die relative Häufigkeit des Auftretens des Ereignisses $X=1$ von der (unbekannten) Wahrscheinlichkeit dieses Ereignisses abweicht.

Mit der Ungleichung (8) sind noch keine weiteren Aussagen möglich, da auf der rechten Seite noch $p = p(X=1)$ vorkommt. p ist jedoch unbekannt und soll durch $h_m(X=1)$ abgeschätzt werden. Für p gilt aber die Aussage

$$p \cdot (1-p) \leq \frac{1}{4} \tag{9}$$

Aus (8) und (9) ergibt sich

$$p(|h_m(X = 1) - p(X = 1)| \geq \varepsilon) \leq \frac{1}{4 \cdot m \cdot \varepsilon^2} \tag{10}$$

Um (10) anwenden zu können, muß noch geklärt werden, wie groß die Abweichung von $h_m(X=1)$ und $p(X=1)$ maximal sein darf, um $p(X=1)$ sinnvoll durch $h_m(X=1)$ abschätzen zu können. Mit anderen Worten, es muß bestimmt werden, wie groß ε gewählt wird.
Die Wahl von ε sollte abhängig sein von der Anzahl k der verschiedenen Werte, die die Zufallsgröße annehmen kann. Kann sie sehr viele verschiedene Werte annehmen, so sollte ε sehr klein gewählt werden, da sonst völlig anders strukturierte Verteilungen entstehen können. Ist k klein (z.B. $k=2$), so kann ε etwas größer gewählt werden, ohne daß Charakteristika der Verteilung verloren gehen. Es soll deshalb hier der Ansatz gewählt werden, daß ε proportional zu $\frac{1}{k}$ ist. Es gibt also einen Faktor f_1 mit $\varepsilon = \frac{f_1}{k}$. Die rechte Seite der Ungleichung (10) hat somit die Form

$$\frac{k^2}{4 \cdot f_1^2 \cdot m} \tag{11}$$

Dieser Term kann als eine Art Unsicherheitsfaktor aufgefaßt werden. Ist für eine Zufallsgröße X dieser Wert sehr klein, so wurde die tatsächliche Verteilung mit hoher Wahrscheinlichkeit richtig angenähert, auf Aussagen über $I(X)$ ist Verlaß. Wird der Wert von (11) größer, so sinkt die Verläßlichkeit von Aussagen über $I(X)$. Ist der Unsicherheitsfaktor größer als eine bestimmte Schranke f_2, so kann dem errechneten Wert von $I(X)$ keine praktische Bedeutung beigemessen werden. Als Bedingung für die Verläßlichkeit von Aussagen bleibt also

$$\frac{k^2}{4 \cdot f_1^2 \cdot m} \leq f_2 \tag{12}$$

Setzt man $f := 4 \cdot f_1^2 \cdot f_2$, so ergibt sich

$$\frac{k^2}{m} \leq f \tag{13}$$

Für die gemeinsame Information der Zufallsgrößen $X_1,\ldots,X_q$ hat die Bedingung analog (13) folgende Form:

$$\frac{(k_1 \cdot \ldots \cdot k_q)^2}{m} \leq f \tag{14}$$

k_i sind dabei die Anzahl der verschiedenen Werte, die X_i annehmen kann, m ist wieder die Anzahl der Meßwerte, f ist ein konstanter Faktor.

Der Wert für f kann nur nach praktischen Gesichtspunkten und Erfahrungswerten festgelegt werden. Auffällig an Ungleichung (14) ist jedoch, daß f im Vergleich zu q (der Anzahl der Zufallsgrößen, zwischen denen eine gemeinsame Information berechnet werden soll) kaum Einfluß hat: Der Ausdruck $\frac{(k_1 \cdot \ldots \cdot k_q)^2}{m}$ wächst exponentiell bzgl. q, so daß eine konstante Grenze f schnell erreicht wird.

Als wichtigste allgemeine Aussage aus (14) läßt sich feststellen, daß zur Bestimmung der gemeinsamen Information einer höheren Anzahl von Dimensionen recht große Datenmengen nötig sind, um praktisch brauchbare Ergebnisse zu erhalten.

2.4. Klasseneinteilungen

In vielen zu visualisierenden Datensätzen treten Zufallsgrößen auf, deren Wertebereiche eine große Mächtigkeit haben (z.B. real- oder integer-Werte). Für solche Zufallsgrößen bringt die informationstheoretische Analyse folgende Probleme mit sich:

– Die Rechenzeit steigt quadratisch zur Mächtigkeit der Wertebereiche der Zufallsgrößen.
– Der Unsicherheitsfaktor ist sehr hoch, so daß dem Ergebnis kaum praktische Bedeutung beigemessen werden kann.

Um überhaupt zu Aussagen über solche Zufallsgrößen zu kommen, muß auf dem Wertebereich der Zufallsgröße eine Klasseneinteilung vorgenommen werden, wobei die Anzahl der Klassen kleiner ist als die Mächtigkeit des Wertebereichs. In die informationstheoretische Analyse geht nicht mehr der gesamte Wertebereich der Zufallsgröße ein, sondern nur noch die Elemente der Klasseneinteilung, was zu einem wesentlich geringeren Unsicherheitsfaktor führt. Neben dieser positiven Wirkung treten jedoch auch negative Effekte auf: Ein Teil der dem Datensatz innewohnenden Information fließt nicht mit in die Analyse ein.

Die Zusammenhänge zwischen Zufallsgrößen mit großem Wertebereich können verschiedenster Art sein. So besteht z. B. zwischen den Zufallsgrößen X und Y (beide die real-Zahlen als Wertebereich) ein Zusammenhang, wenn X_i und Y_i stets an der dritten Stelle nach dem Komma übereinstimmen, an den ersten beiden Stellen jedoch verschieden sind. Wichtig für die Analyse sind jedoch nur Zusammenhänge, die auch in der Visualisierung erkennbar sein können. (Der obige konstruierte Zusammenhang ist in gebräuchlichen Visualisierungen sicher nicht zu erkennen!)

Solche bei der Visualisierung erkennbaren Zusammenhänge können Aussagen sein wie: „Wenn X große Werte annimmt, nimmt auch Y große Werte an“ oder „Wenn X mittlere Werte annimmt, ist Y meist recht klein“.

Auffällig ist, daß gerade solche Zusammenhänge durch eine Klasseneinteilung nicht verloren gehen, sondern im Gegenteil noch stärker herausgearbeitet werden.

Dies alles führt zur Erkenntnis, daß es günstig ist, eine Klasseneinteilung mit möglichst wenig Klassen durchzuführen. Bei der Analyse der bisher praktisch untersuchten Datensätze (s. Abschnitt 4) hat sich eine äquidistante Einteilung in 3 Klassen bewährt, da auch bei einer verbalen Formulierung von Zusammenhängen zwischen ordinalen Daten mit großem Wertebereich oft Formulierungen wie „niedrig“, „mittel“ und „hoch“ benutzt werden.

Die bisher getroffenen Aussagen zur Klasseneinteilung gelten nur für die Klasseneinteilung zur informationstheoretischen Analyse. Zur Visualisierung erfolgt die Klasseneinteilung (wenn überhaupt) in Abhängigkeit von der Bildschirmgröße in möglichst vielen Klassen.

2.5. Unvollständig bekannte Datensätze

Bisher wurde davon ausgegangen, daß der zu visualisierende Datensatz vollständig bekannt ist. Für praktische Anwendungen kann es aber vorkommen, daß einzelne Größen nicht gemessen oder beobachtet wurden, daß also Elemente der Matrix M unbekannt sind. Für diesen Fall wird für die Zufallsgröße eine neue Realisierung eingeführt, nämlich die Realisierung „unbekannt“. Der Wertebereich einer solchen Zufallsgröße erhöht sich also um 1. Mit dieser Erweiterung wird der informationstheoretische Ansatz völlig analog gerechnet.

Tritt als Realisierung der Zufallsgröße der Wert „unbekannt“ sehr selten auf, so ergeben sich für $I(X)$ ähnliche Werte wie beim Ansatz ohne unbekannte Elemente in M. Tritt umgekehrt bei der Realisierung von X überwiegend der Wert „unbekannt“ auf, so geht $I(X)$ gegen 0.

3. Die Analyse der Visualisierungstechniken

Für jede neu in das System aufzunehmende Visualisierungstechnik müssen folgende Fragen beantwortet werden:

– Wieviele Dimensionen können mit dieser Technik sinnvoll visualisiert werden?
– Ist die Technik für Zufallsgrößen mit großem oder mit kleinem Wertebereich geeigneter?
– Wieviele Beobachtungsfälle können sinnvoll in die Visualisierung einbezogen werden, ohne daß das Bild unübersichtlich wird?
– Ist es möglich, mit der Visualisierungstechnik Zusammenhänge zwischen zwei Dimensionen herauszuarbeiten?
– Sind mit der Visualisierungstechnik Zusammenhänge von mehr als zwei Dimensionen darstellbar?

Diese Fragen sollen nun für drei konkrete Visualisierungstechniken behandelt werden. Die dabei als Beispiel verwendeten Bilder stammen aus einem mikrobiologischen Datensatz, der sich aus Untersuchungen zur computergestützten Diagnostik und Epidemiologie von Harnwegsinfektionen ergab [1]. Der Datensatz enthält 41 Dimensionen und 343 Beobachtungsfälle. In den einzelnen Dimensionen sind codiert: Labornummer, Datum und Verfahren der Entnahme, Aufenthaltsdauer, Alter, Geschlecht, Typ der Infektion, Einsender, Resistenz-, Virulenz-, Siderophore- und weitere biochemische Merkmale. Viele dieser Merkmale sind binär skaliert. Die verwendeten Abbildungen zeigen Grauwertdarstellungen von im Original farbigen Bildern.

Bei der *Technik der parallelen Koordinaten* werden die Koordinatenachsen parallel angeordnet und für jeden Beobachtungsfall die Punkte der benachbart liegenden Dimensionen durch Strecken verbunden. Ein Beobachtungsfall entspricht also einem Streckenzug, dessen Eckpunkte auf den Koordinatenachsen liegen. Näheres zu dieser Technik findet sich in [7] und [9].

Abbildung 1 zeigt die Visualisierung eines Ausschnittes des mikrobiologischen Datensatzes. Es wurden 5 Dimensionen (Aufenthaltsdauer, Alter, Material und zwei binär skalierte Merkmale AMP und NIF) visualisiert.

Mit dieser Technik können viele Dimensionen gleichzeitig visualisiert werden, Grenzen werden prak-

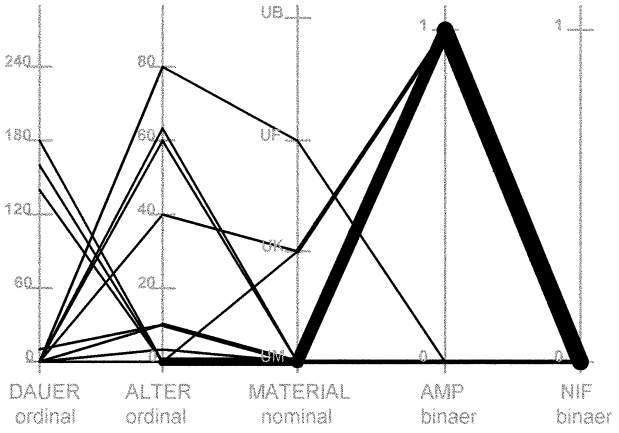

Abb. 1. Parallele Koordinaten

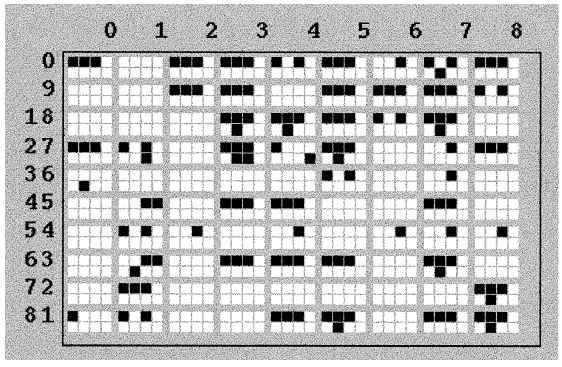

Abb. 2. Shape coding

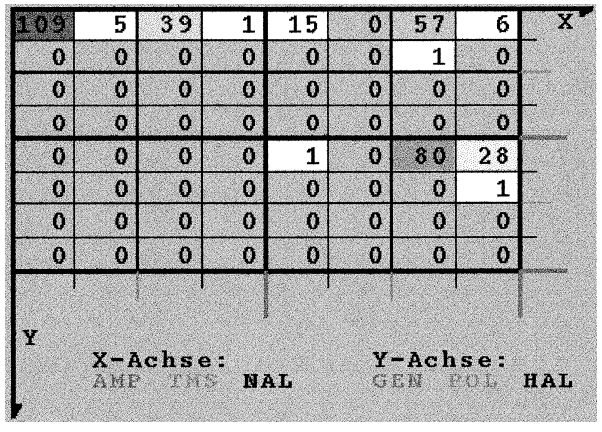

Abb. 3. Dimensional stacking

tisch nur durch die Breite des Bildschirms gesetzt. Parallele Koordinaten sind besonders geeignet für Zufallsgrößen mit großen Wertebereichen. Bei kleinen Wertebereichen (im Beispiel die Dimensionen AMP und NIF) überlagern sich Streckenzüge sehr schnell, was zu Unübersichtlichkeiten im Bild führt. Die Anzahl der zu visualisierenden Beobachtungsfälle ist bei den parallelen Koordinaten begrenzt. Mit wachsender Anzahl von Beobachtungsfällen, also Streckenzügen, kommt es schnell zu Überlagerungen. Zusammenhänge zwischen zwei Dimensionen sind sehr gut darstellbar. Hierfür muß die Reihenfolge der Koordinatenachsen so gewählt werden, daß Dimensionen, zwischen denen Zusammenhänge existieren, nebeneinander liegen. Da Zusammenhänge nur bei benachbart liegenden Koordinatenachsen erkennbar sind, sind bei der Technik der parallelen Koordinaten Zusammenhänge von mehr als zwei Dimensionen kaum erkennbar.

Bei der *shape-coding-Technik* [2] werden für jeden Beobachtungsfall die Werte der einzelnen Dimensionen in eine Ikone abgebildet und dort farbcodiert.

Abbildung 2 zeigt die Visualisierung eines Teils des mikrobiologischen Datensatzes: 90 Beobachtungsfälle und 10 Dimensionen wurden visualisiert. Die Dimensionen haben alle einen binären Wertebereich, zur Codierung eines Wertes genügen hier also die Farben Schwarz und Weiß.

Mit dieser Technik können viele Dimensionen visualisiert werden, es muß lediglich die Ikone hinreichend groß gewählt sein. Shape coding ist günstig für Zufallsgrößen mit kleinem Wertebereich, da mittels Farbcodierung nur wenig verschiedene Werte für das Auge unterscheidbar sind. Da für jeden Beobachtungsfall eine Ikone benötigt wird, ist die maximale Anzahl der Beobachtungsfälle nur durch die Bildschirmgröße begrenzt. Das Erkennen von Zusammenhängen zwischen zwei Dimensionen ist möglich, die Stärke des Verfahrens liegt aber im Erkennen von Zusammenhängen zwischen mehreren Dimensionen (bevorzugt auftretende Muster in den Ikonen).

Bei der *dimensional-stacking-Technik* [10] werden die Dimensionen zu Paaren geordnet und „ineinandergeschachtelt“. Für alle möglichen Wertekombinationen, welche die untersuchten Dimensionen annehmen können, wird die Häufigkeit des Auftretens farbcodiert.

Abbildung 3 zeigt die Visualisierung von 6 Dimensionen mit binärem Wertebereich. Die Zahlen in den einzelnen Kästchen geben dabei die absolute Häufigkeit des Auftretens der einzelnen Wertekombinationen an. Die relative Häufigkeit wurde zusätzlich in den Kästchen farbcodiert.

Die Anzahl der darstellbaren Dimensionen ist stark begrenzt, da eine größere Schachtelungstiefe schnell zu unübersichtlichen Bildern führt. Das Verfahren ist nur geeignet für Zufallsgrößen mit kleinem Wertebereich, da sonst der Platzbedarf schnell die Bildschirmgrenzen überschreitet. Dafür können praktisch unbegrenzt viele Beobachtungsfälle visualisiert werden, da bei dieser Technik nur die Häufigkeiten farbcodiert werden. Das Erkennen von Zusammenhängen zwischen zwei Dimensionen ist gut möglich, wenn sich beide Dimensionen in einer Stacktiefe befinden. Bei mehr als zwei Dimensionen ist ein „geschultes Auge“ seitens des Nutzers notwendig.

Aus der Kenntnis einer solchen Analyse einer Visualisierungstechnik und aus den automatisch ermittelten Charakteristika des Datensatzes (Anzahl der Dimensionen, Anzahl der Beobachtungsfälle, Mächtigkeiten der Wertebereiche der einzelnen Zufallsgrößen, Paare und höhere Tupel von Dimensionen, zwischen denen Zusammenhänge bestehen) kann entschieden werden, ob eine Visualisierungstechnik für einen Datensatz geeig-

net ist oder nicht. Die hier verbal vorgenommene Analyse der Visualisierungstechniken kann auch formalisiert werden, so daß die Entscheidung, ob eine Technik günstig ist oder nicht, auch automatisch erfolgen kann.

4. Ergebnisse

Der hier behandelte Ansatz wurde auf zwei konkrete Datensätze angewandt. Als Grenze für den Unsicherheitsfaktor wurde $f := 1$ gesetzt.
Der in [1] eingeführte mikrobiologische Datensatz (41 Dimensionen, 343 Beobachtungsfälle), der als Beispiel bereits im vorigen Abschnitt verwendet wurde, wurde der informationstheoretischen Analyse unterzogen. Bei Nutzung der relativen Relevanz mit g = 0.2 ergaben sich:

24 Paare von Dimensionen mit relevanter gemeinsamer Information,
3 Tripel von Dimensionen mit relevanter gemeinsamer Information,
keine höheren Tupel.

Die Rechenzeit dafür betrug 40 sec auf einem PC286.

Der Unsicherheitsfaktor für die Paare lag zwischen $\frac{(2\cdot 2)^2}{343}$ und $\frac{(2\cdot 3)^2}{343}$. Für die Tripel lag der Unsicherheitsfaktor zwischen $\frac{(2\cdot 2\cdot 2)^2}{343}$ und $\frac{(2\cdot 2\cdot 3)^2}{343}$. Allen gefundenen Zusammenhängen kann also eine praktische Bedeutung beigemessen werden.

Dies bedeutet für die einzelnen Visualisierungstechniken:

– *parallele Koordinaten:* Eine Visualisierung mit dieser Technik ist möglich, auch wenn nicht alle Anforderungen erfüllt sind. Günstig sind die 24 Paare von Dimensionen mit relevanter gemeinsamer Information, die drei Tripel von Dimensionen wären in der Visualisierung kaum erkennbar. Das Beispiel aus Abb. 1 kann bei hinreichend großem Bildschirm problemlos auf eine größere Anzahl von Dimensionen (also Koordinatenachsen) erweitert werden. Problematisch für die Technik ist, daß die meisten Merkmale binär skaliert sind, was zu starken Überlagerungen der Streckenzüge führt. Abhilfe schafft hier, eine zufällige Streuung um die Werte 0 und 1 auf den Koordinatenachsen einzuführen.
– *shape coding:* Eine Visualisierung mit dieser Technik ist möglich. Sowohl die 24 Paare als auch die 3 Tripel von Dimensionen mit relevanter gemeinsamer Information wären sichtbar. Da die meisten Dimensionen einen Wertebereich geringer Mächtigkeit haben, ist eine Farbcodierung der einzelnen Werte ausreichend. Für das Beispiel in Abb. 2 ist es gut vorstellbar, sowohl die Anzahl der Beobachtungsfälle (also Ikonen) als auch die Anzahl der Dimensionen (also der codierten Werte in einer Ikone) zu erhöhen.
– *dimensional stacking:* Diese Technik kann nicht empfohlen werden. Zum einen erlaubt es der Platzbedarf nicht, eine größere Anzahl von Dimensionen zu visualisieren. Ebenso schnell wächst der Platzbedarf bei Verwendung von Dimensionen mit einem Wertebereich höherer Mächtigkeit (z. B. bei den Dimensionen Einsender und Alter). Die Tripel von Dimensionen mit relevanter gemeinsamer Information wären in der Visualisierung nicht erkennbar.

Ein ornithologischer Datensatz, der das Auftreten verschiedener Vogelarten in bestimmten Gebieten Deutschlands in den letzten 30 Jahren beschreibt, wurde der informationstheoretischen Analyse unterzogen. Dieser Datensatz enthielt 1005 Beobachtungsfälle und 212 Dimensionen. Charakteristisch war die hohe Anzahl von unbekannten Elementen in der Matrix: Für jeden Beobachtungsfall wurde nur eine geringe Anzahl von Vogelarten beobachtet. Dies bedeutet, daß die zu den einzelnen Dimensionen gehörenden Informationen gering sind. Um überhaupt zu Aussagen über Zusammenhänge zwischen den Dimensionen zu kommen, wurde die Relevanz relativ zu den Teilinformationen und g = 0.5 gewählt. Die informationstheoretische Analyse ergab:

18 Paare von Dimensionen mit relevanter gemeinsamer Information
2 Tripel von Dimensionen mit relevanter gemeinsamer Information
keine höheren Tupel.

Die Rechenzeit für diese Analyse betrug ca. 3 min auf einer DEC5000 Workstation. Für die Paare lag der Unsicherheitsfaktor zwischen $\frac{(2\cdot 2)^2}{1005}$ und $\frac{(2\cdot 3)^2}{1005}$, für die Tripel lag er bei $\frac{(2\cdot 2\cdot 3)^2}{1005}$. Auch diesen gefundenen Tupeln kann also eine praktische Bedeutung beigemessen werden.

Keine der drei behandelten Visualisierungstechniken ist in der Lage, diesen Datensatz vollständig zu visualisieren. Für parallele Koordinaten und shape-coding sind es zu viele Beobachtungsfälle, das dimensional-stacking-Verfahren scheitert an der zu großen Anzahl von Dimensionen. In diesem Fall wären zwei Auswege möglich:

– Beschränkung auf einen Teil des Datensatzes, der seinerseits wieder der informationstheoretischen Analyse unterzogen wird.
– Suche nach neuen Visualisierungstechniken, die alle Anforderungen des Datensatzes erfüllen können.

5. Ausblick

Weiterentwicklungen des hier beschriebenen Ansatzes sind in folgende Richtungen möglich:

– Ist mit keiner dem System bekannten Technik die Visualisierung des gegebenen Datensatzes möglich, so könnten sich aus der informationstheoretischen Analyse Vorschläge ableiten lassen, wie der Datensatz zu teilen und somit zu verkleinern ist.
– Es sind Verallgemeinerungen dieses Ansatzes zu suchen, die sich nicht nur auf ein wichtiges Interpretationsziel, sondern auf einen ganzen Komplex von Interpretationszielen stützen. Notwendig ist hierfür jedoch die formale Beherrschung des Begriffs „Interpretationsziel“.

Danksagung. Der Autor möchte sich hiermit bei Frau Prof. Schumann (Universität Rostock, Fachbereich Informatik) für die wertvollen Hinweise bei der Erstellung dieser Arbeit herzlich bedanken. Ebenfalls bedanken möchte sich der Autor bei Frau Susanne Arndt für ihre Hilfe bei der Bereitstellung der Abbildungen.

Literatur

1. Arndt, S.: Visualisierung multivariater Daten am Beispiel eines mikrobiologischen Datensatzes. Rostocker Informatik-Berichte *14,* 4–19 (1993)
2. Beddow, J.: Shape coding of multidimensional data on a microcomputer display. Proceedings IEEE Visualization '90, San Francisco, 1990
3. Bergeron, R. D., Grienstein, G. G.: A reference model for the visualization of multidimensional data. Proceedings Eurographics '89, pp. 393–399. Amsterdam, London, New York, Tokyo: North Holland 1989
4. Beshers, C., Feiner, S.: Automated design of virtual worlds for visualizing multivariate relations. Proceedings IEEE Visualization '92, Boston, pp. 283–289 (1992)
5. Brodlie, K. W., Carpenter, L. A., Earnshaw, R. A., Gallop, J. R., Hubbold, R. J., Mumford, A. M., Osland, C. D., Quarendon, P.: Scientific visualization. Berlin, Heidelberg, New York: Springer 1992
6. Casner, S.: A task-analytic approach to the automated design of graphic presentations. ACM Trans. Graphics *10* (No. 2) 111–151 (1991)
7. Finsterwalder, R.: Entscheidungsunterstützende Visualisierung bei der mehrzieligen Entwurfsoptimierung. Inf. Forsch. Entw. *7* (3), 138–144 (1992)
8. Heise, W., Quattrocchi, P.: Informations- und Codierungstheorie. Berlin, Heidelberg, New York: Springer 1989
9. Inselberger, A.: The plane with parallel coordinates. Visual Comput. *1,* 69–91 (1985)
10. LeBlanc, J., Ward, M. O., Wittels, N.: Exploring n-dimensional databases. Proceedings IEEE Visualization '90, San Francisco, 1990
11. Robertson, P. K.: A methodology for scientific visualization: choosing representations based on an natural scene paradigm. Proceedings IEEE Visualization '90, San Francisco, pp. 114–123 (1990)
12. Schumann, H.: Entscheidungshilfen zur Visualisierung wissenschaftlicher Daten. Rostocker Informatik-Berichte *13,* 62–69 (1992)
13. Shannon, C. E., Weaver, W.: Mathematische Grundlagen der Informationstheorie. München: Oldenbourg 1976
14. Springmeyer, R. R., Blattner, M. M., Max, N. L.: A characterization of the scientific data analysis process. Proceedings IEEE Visualization '92, Boston, pp. 235–242 (1992)
15. Wehrend, S., Lewis, C.: A problemoriented classification for visualization techniques. Proceedings IEEE Visualization '90, San Francisco, pp. 139–142 (1990)

Holger Theisel studierte von 1989 bis 1994 Informatik an der Universität Rostock. Während dieser Zeit arbeitete er in der Arbeitsgruppe von Frau Prof. Schumann auf verschiedenen Gebieten der Computergraphik, wie der realistischen Bilddarstellung, der Visualisierung höherdimensionaler Fraktale und der Visualisierung wissenschaftlicher Daten. Von Oktober 1994 bis August 1995 weilt er zu einem Forschungsaufenthalt an der Arizona State University in Tempe, USA.