Zuerst ersch. in: Informatik-Spektrum ; 33 (2010), 6. - S. 589-600

http://dx.doi.org/10.1007/s00287-010-0481-z

Visualisierung und Analyse multidimensionaler Datensätze

Dirk J. Lehmann · Georgia Albuquerque
Martin Eisemann · Andrada Tatu
Daniel Keim · Heidrun Schumann
Marcus Magnor · Holger Theisel

Einleitung

Im Sommer 1854 brach eine der schlimmsten Choleraepidemien in London aus; nach bereits vier Ausbrüchen innerhalb von nur 23 Jahren schritt dieser so schnell voran und war so tödlich wie noch keiner zuvor. Ohne Vorankündigung starben innerhalb weniger Tage allein im Stadtteil Soho weit über 100 Menschen. Ein Gegenmittel gab es nicht. Die ÄrztInnen vermuteten, dass gefährliche Dünste, Miasmen, für die Ausbreitung der Krankheit verantwortlich seien. Woher aber diese Dünste kommen sollten, war ein Rätsel. Der einzig mögliche Schutz bestand in der Flucht aus der Stadt.

John Snow, ein Arzt im Londoner Stadtteil Soho, erkannte, dass die gängige Miasmentheorie keine Hilfe bot. Stattdessen hatte er die Vermutung, dass sich die Krankheit von nur wenigen Infektionsherden aus verbreitete. Sein Ziel war es, diese zu finden und zu eliminieren, um die Seuche einzudämmen. Doch wo sollte er mit der Suche nach hypothetischen Krankheitsquellen beginnen, zu einer Zeit, als weder die Existenz von bakteriellen Krankheitserregern noch das Konzept von Infektionswegen bekannt waren? Seine einzige Möglichkeit bestand darin, die Suche auf seine Beobachtungen zu stützen.

Er kam auf die Idee, die Wohnorte der Choleraopfer in seinem Bezirk in einer Stadtkarte einzutragen, welche berühmt wurde als „Ghost Map" (siehe Abb. 1). Durch diese Darstellung wird sichtbar, dass die Wohnorte der Choleraopfer nicht gleichmäßig über Soho verteilt waren, sondern dass es eine klare Häufung auf der Broad Street gab. Dort befand sich eine öffentliche Wasserpumpe, an der sich die Bewohner mit Trinkwasser versorgten. Ein Zufall? John Snow überzeugte die Stadtverwaltung,

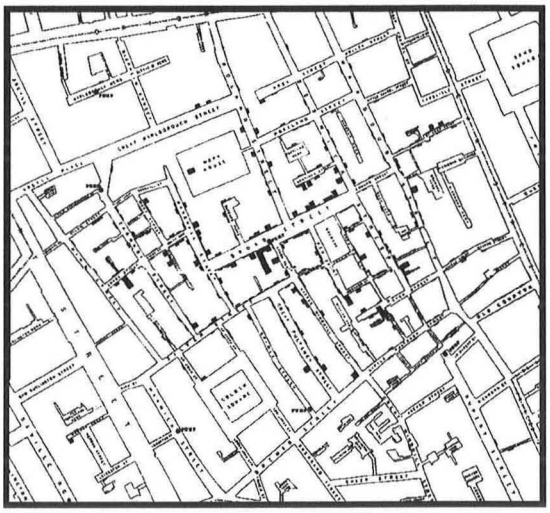


Abb. 1 In der „Ghost Map" von John Snow sind die Wohnorte der Choleraopfer eingetragen; es wird deutlich, dass sich die Todesfälle um eine konkrete Wasserpumpe herum häufen

den Schwengel der Pumpe abzumontieren, was die Bewohner zwang, ihr Wasser an anderen Pumpen zu holen. Innerhalb weniger Tage ging die Opferzahl

Dirk J. Lehmann · Holger Theisel
Universität Magdeburg, Magdeburg
E-Mail: {dirk, theisel}@isg.cs.uni-magdeburg.de

Georgia Albuquerque · Martin Eisemann · Marcus Magnor
TU Braunschweig, Braunschweig
E-Mail: {georgia, eisemann, magnor}@cg.cs.tu-bs.de

Andrada Tatu · Daniel Keim
Universität Konstanz, Konstanz
E-Mail: {tatu, keim}@dbvis.inf.uni-konstanz.de

Heidrun Schumann
Universität Rostock, Rostock
E-Mail: schumann@informatik.uni-rostock.de

Zusammenfassung

Für multidimensionale Datensätze existieren eine Reihe von automatischen Analysemethoden und Visualisierungstechniken, um ihnen innewohnende Zusammenhänge und Charakteristika aufzudecken. Die zunehmende Größe und Komplexität solcher Daten macht es notwendig, beide Ansätze miteinander zu kombinieren. In diesem Artikel stellen wir daher etablierte Methoden zur visuellen und zur automatischen Datenanalyse vor und zeigen neuere Ansätze auf, diese sinnvoll miteinander zu kombinieren. Dabei werden alle Erläuterungen anhand anschaulicher Beispiele verdeutlicht und so für den Leser nachvollziehbar.

in Soho drastisch zurück. John Snow hatte mithilfe einer Visualisierung und ihrer richtigen Analyse zahlreichen Menschen das Leben gerettet.

Das Beispiel macht deutlich, dass Datenvisualisierung und -analyse keine neue Wissenschaft ist. Das Verfahren von John Snow findet auch noch heute Anwendung, z. B. in der Kriminalistik, wenn Tatorte und Beweismittel auf Landkarten miteinander in Beziehung gesetzt werden, um versteckte Muster zu erkennen. So geht es grundsätzlich darum, aus unvollständigen Informationen allein auf Grundlage der vorhandenen (beobachteten) Daten auf nützliche, sinnhafte Zusammenhänge zu schließen. Visualisierung ist damit ein Werkzeug zum phänomenologischen Verständnis von Zusammenhängen: Anhand seiner „Ghost Map" konnte John Snow den Zusammenhang zwischen Cholera und Trinkwasser postulieren, ohne sich durch wissenschaftliche Grundlagen wie Bakteriologie oder Epidemiologie leiten lassen zu können.

Kernprinzip visueller Analyse ist es, nichtzufällig erscheinende Zusammenhänge zwischen scheinbar unabhängigen Größen aufzufinden, geleitet durch Intuition und durch unser visuelles System. Dabei unterstützt der Rechner den menschlichen Suchvorgang, indem er z. B. große Datenmengen voranalysiert, auf verdächtige Nicht-Zufälligkeiten hinweist und Daten so visuell präsentiert, dass ein Mensch etwaige Muster schnell und sicher erfassen kann.

Solche Techniken lassen sich in vielen Bereichen anwenden. Ihre volle Leistungsfähigkeit entfalten sie jedoch an multidimensionalen Datensätzen, in denen sich interessante Zusammenhänge verstecken können, die ohne visuelle Analysemethoden verborgen blieben. Ein Beispiel geben Versicherungsfirmen: So kostete die Kfz-Versicherung für einen roten Wagen in den USA lange Zeit deutlich mehr als für ein baugleiches weißes Auto, weil eine Analyse der Unfallstatistiken ergeben hatte, dass rote Autos häufiger verunglücken als andersfarbige Wagen. Doch haben Automobile natürlich noch andere charakterisierende Dimensionen als nur ihre Lackierung. Eine vollständige visuelle Analyse sämtlicher zugelassener Wagen könnte z. B. ergeben, dass rote Autos häufiger von Männern gefahren werden, oder dass PS-starke Motoren nur sehr selten in weißen Autos verbaut werden, oder … Um der wahren Ursache eines phänomenologischen Zusammenhangs auf den Grund zu gehen, bedarf es daher zweierlei: der Suche nach allen scheinbaren Zusammenhängen in einem Datensatz (exhaustive search) sowie eines Menschen, der die gefundenen Zusammenhänge kausal verknüpfen und Scheinzusammenhänge auf ihre wahren Ursachen zurückführen kann. Das genannte historische Beispiel beschreibt einen einfachen Fall von multidimensionalen Daten. Ebenso einfach (und doch wirksam) ist die Wahl der Visualisierungstechnik. Die heutige Situation lässt sich dadurch beschreiben, dass die zu untersuchenden Datensätze immer größer und komplexer werden, und auf der anderen Seite eine stetig wachsende Vielzahl von automatischen und visuellen Analysemethoden zur Verfügung stehen. Eine Kombination von automatischen und visuellen Techniken ist somit notwendig und aktueller Forschungsgegenstand. In den nächsten Abschnitten geben wir eine formale Definition von multidimensionalen Daten, beschreiben existierende Standardtechniken zur automatischen und visuellen Datenanalyse und zeigen an Beispielen einige aktuelle Arbeiten zur sinnvollen Kombination solcher Techniken.

Multidimensionale Datensätze

Um imstande zu sein, ein (visualisiertes) Muster zu interpretieren bzw. um das Gesehene in einen Sinnkontext einzuordnen, ist ein allgemeines Verständnis von der Struktur zugrunde liegender multidimensionaler Datensätze unerlässlich. Aus diesem Grund werden sie in diesem Abschnitt eingeführt.

Abstract
Concerning multi-dimensional data sets there exist a lot of visual-based as well as automatical techniques to detect inherent relations and characteristics. Due to the (increasing) size and complexity of such data, it is necessary to combine both approaches. In this article, we therefore present established visual-based and automatical data analysis approaches and we reveal modern methods to combine these approaches, with the goal to enhance the data analysis process. All explanations are supported by examples to ease the reader's understanding.

Ein intuitives Beispiel eines solchen Datensatzes ist das Resultat einer Messung in einem Zimmer, in welchem eine Anzahl von physikalischen Messgrößen erfasst werden, wie beispielsweise die Temperatur und der Luftdruck. Hierbei spannt das Zimmer einen dreidimensionalen Messraum und die beiden Messgrößen einen zweidimensionalen Messgrößenraum auf: $\mathbb{R}^3 \rightarrow \mathbb{R}^2$. Die Anzahl der gemessenen Paare entspricht der Mächtigkeit der Population.

Ein beliebiger Datensatz ist somit formal charakterisiert durch eine Abbildung f von s-vielen Elementen x_i; $i = 1, \ldots, s$ eines n-dimensionalen Messraumes (spatial domain) auf s-viele Elemente ξ_j; $j = 1, \ldots, s$ eines m-dimensionalen Messgrößenraumes (data domain) und entspricht aus mathematischer Sicht einer diskreten multivariaten vektorwertigen Funktion:

$$x \rightarrow f(x) = \xi : \mathbb{R}^n \rightarrow \mathbb{R}^m.$$

Somit werden den n unabhängigen Dimensionen des Messraumes m abhängige Dimensionen des Messgrößenraumes zugeordnet, wobei die Elementanzahl s die Population der Elemente darstellt.

In der Fachliteratur wird folglich zumeist zwischen den Dimensionen beider Räume unterschieden, wenn auch nicht immer einheitlich [18]. Nicht immer ist das zielführend, weil abhängige Dimensionen auch als unabhängig betrachtet werden können und umgekehrt. Zur Charakterisierung der Dimensionalität eines Datensatzes kann es stattdessen zweckdienlich sein, die Gesamtanzahl der erfassten Dimensionen als Merkmalsraum, unter dem Begriff der *Variabeln* $k = n + m$, zu bündeln. Im Weiteren benutzen wir daher den Begriff der Variabel, wenn wir von einer Dimension des Datensatzes sprechen. Es ist unter anderem die Anzahl dieser Variabeln, welche die Komplexität des „hervorgerufenen" Visualisierungsproblems bestimmt.

Dateneigenschaften

Ein Datensatz ist jedoch nicht nur durch seine Dimensionalität charakterisiert, sondern auch durch die konkreten Eigenschaften seiner Daten selbst [21]:

- **Ordnung.** Ein Datum lässt sich als Tensor i-ter Ordnung beschreiben. Dabei entspricht ein Skalar einem Tensor nullter Ordnung, ein Vektor einem Tensor erster Ordnung, eine Matrix einem Tensor zweiter Ordnung, usw. Weil aber eine Ordnung größer als Null auch durch eine größere Variablenanzahl ausgedrückt werden kann, gehen wir zumeist von skalaren Daten aus.

- **Skaleneigenschaft**
  - **Quantitativität.** Die Daten sind Zahlen eines bestimmten Wertebereiches konkreter Zahlenmengen ($\mathbb{Q}, \mathbb{Z}, \mathbb{R}, \ldots$).
  - **Qualitativität.** Unterliegen die Daten einer Ordnungsrelation, wie z. B. größer, kleiner oder gleich, wird von Ordinalität gesprochen; sind sie andererseits textuelle Bezeichner, wie z. B. eine Farbe (rot, grün, blau) oder eine Form (rund, eckig, länglich), handelt es sich um nominelle Daten. Ein nominelles Datum wird zumeist als a priori Klassifikator der Population genutzt, um ihre Elemente eindeutig einer Klasse zuzuordnen. Vorweggreifend sei darauf hingewiesen, dass die Klassenzugehörigkeit innerhalb einer Visualisierung zumeist durch eine klassenkonsistente Farbkodierung kenntlich gemacht wird.

Zusammenfassend lässt sich ein *multidimensionaler Datensatz* als ein Datensatz mit mindestens zwei (oder mehr) skalaren Variabeln verstehen. Um jedoch eine Vorstellung von der praktischen Arbeit zu vermitteln, sei erwähnt, dass es sich dort zumeist um Datensätze mit 30, 40 oder mehr Variabeln handelt, mit einer Population, die durchaus in die Millionen gehen kann.

Letztlich ist ebenfalls auch diese Mächtigkeit der Population charakterisierend für einen Daten-

satz, weil eine Zunahme gewöhnlich mit einer sich verschlechternden Performance[1] einhergeht. Für Datensätze, die mit Standardvisualisierungsmethoden visualisierbar wären, bedeutet dies, dass sie ab einer kritischen Mächtigkeit nicht mehr (vollständig) visualisierbar sind, da der zu erwartende Nutzen der Visualisierung den zeitlichen Aufwand nicht mehr rechtfertigt. Diese Problematik ist ein noch immer aktueller Gegenstand der Forschung und in seiner Gesamtheit ungelöst. Teillösungen mit GPU-basierten Ansätzen existieren jedoch schon heute. Sie haben den Vorteil, zeitaufwendige Berechnungen parallel (gleichzeitig) auszuführen, anstatt seriell (nacheinander).

Standardmethoden zur Visualisierung multidimensioneller Daten

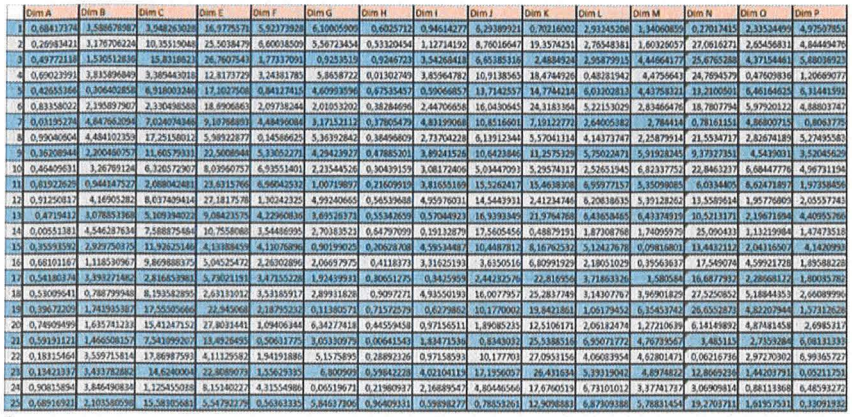
Bei den Methoden zur Datenvisualisierung wird zwischen der Visualisierung von physikalischen Daten (scientific visualization) und abstrakten Daten (information visualization) unterschieden: Dabei sind physikalische Daten insbesondere Skalar-,Vektor- und Tensorfelder, resultierend aus Messungen oder Simulationen. Abstrakte Daten können dagegen zumeist als Listen, Bäume und Graphen beschrieben werden, wie beispielsweise die Verlinkungsstruktur zwischen beliebigen Webseiten solche Daten sind. Eine klare Abgrenzung beider Teilgebiete ist nicht immer möglich oder gar notwendig. Dennoch war die historisch bedingte Unterteilung durchaus erfolgreich: Die Fokussierung auf Teilaspekte des Visualisierungsproblems führte in wenigen Jahren (und somit sehr schnell) zu großen Fortschritten, sowohl in der Theorie als auch in der Anwendung. Gegenwärtig sind jedoch Tendenzen ersichtlich, beide Teilgebiete mehr und mehr zu konsolidieren, um synergetisch weitere Fortschritte zu forcieren. Dessen ungeachtet gibt es beiderseits etablierte Methoden, multidimensionale Datensätze zu visualisieren. Im Weiteren stellen wir insbesondere drei typische Beispiele zur Visualisierung multidimensionaler Daten vor, die in Disziplinen wie der Systembiologie oder der Meteorologie Anwendung finden.

Tabellen. Eine intuitive und wenig aufwendige Möglichkeit ist die textuelle Darstellung der Daten als Tabellen bzw. Tables (oder auch spreadsheets genannt), wobei Spalten den Variabeln und Zeilen den Daten entsprechen. Die Spaltenanzahl korrespondiert mit der Anzahl der Variablen und die Zeilenanzahl mit der Mächtigkeit der Population, wie aus Abb. 2a ersichtlich ist. Obgleich diese Form der Visualisierung einen Datensatz vollständig darstellt, sind weder Zusammenhänge zwischen den Variabeln noch Häufungspunkte der Daten (cluster) ohne größeren kognitiven Aufwand erkennbar. Zusätzlich überschreitet eine große Population oder auch eine große Variabelnanzahl schnell die Darstellungsfähigkeit eines handelsüblichen Monitors.

Grafisch komprimierte Tabellen. Dem letztgenannten Nachteil begegnen grafisch komprimierte Tabellen bzw. Gaphical Compressed Tables erfolgreich, indem sie, anstatt ein Datum textuell darzustellen, eine (nur pixelbreite) qualitative Repräsentation dieser visualisieren. Dadurch kann die benötigte Monitorfläche eines Datums enorm reduziert und insgesamt die Darstellung erheblich komprimiert werden. Abbildung 2b zeigt: Im Vergleich zu den Tables sind sowohl mehr Variabeln als auch eine größere Population auf dem Monitor darstellbar. Zusätzlich werden nun auch Zusammenhänge zwischen Variabeln zumindest rudimentär erkennbar, gegebenenfalls unterstützt durch spaltenbasierte Sortierungen. Falls nötig, können kontextabhängig textuelle Daten mittels einer nutzerbasierten Selektion rekonstruiert werden (table lens, [20]), um derart die Vorteile beider tabellarischen Methoden zu kombinieren.

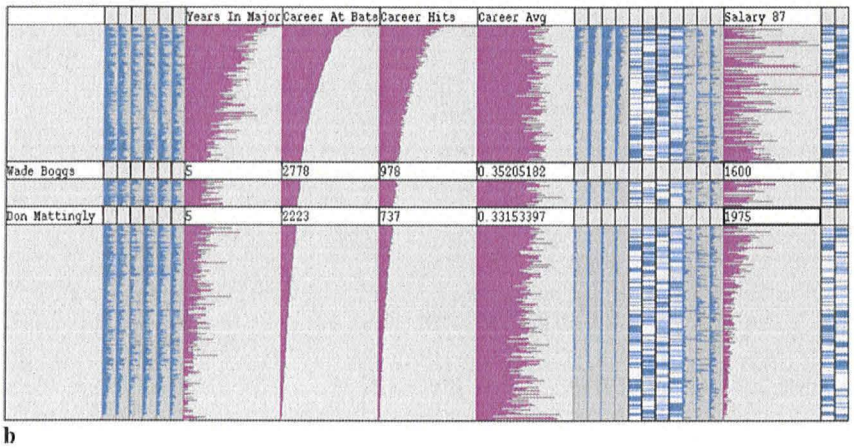
Streudiagramme. Bei den Streudiagrammen bzw. Scatterplots werden die Daten zweier Variabeln als Punkte in ein euklidisches Koordinatensystem eingetragen, deren Achsen die beiden Variabeln repräsentieren (orthogonale Projektion). Mit ihnen lassen sich bivariate Korrelationen, Cluster, Verteilungseigenschaften sowie Kompaktheit und Streuung der Daten sehr gut analysieren, wie es Abb. 3a verdeutlicht. Aussagen über multivariate Zusammenhänge (z. B. multivariate Korrelation) sind allerdings kaum möglich, zudem gehen Informationen über die Datenanzahl verloren, die auf die gleiche Position im Scatterplot abgebildet werden. Transparenzen zu verwenden, kann diesem Effekt bis zu einem gewissen Grad entgegenwirken. Für

[1] Wird von Performance gesprochen, ist je nach Kontext der zeitliche Aufwand und/oder der Speicherverbrauch eines Algorithmus bzw. einer Methode gemeint.



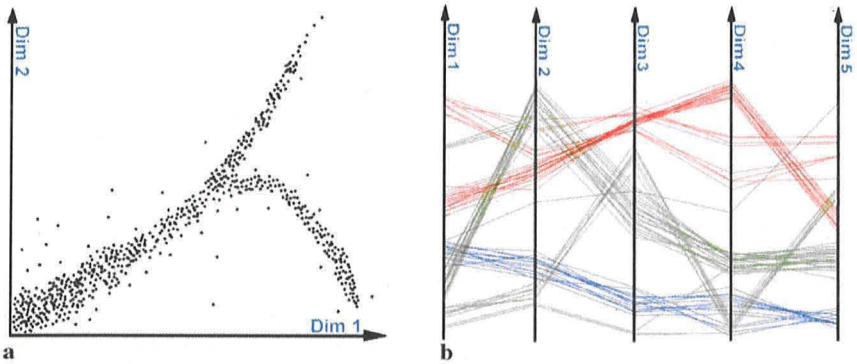
	Dim A	Dim B	Dim C	Dim E	Dim F	Dim G	Dim H	Dim I	Dim J	Dim K	Dim L	Dim M	Dim N	Dim O	Dim P
1	0,68417374	3,588678987	3,948263029	16,9775571	5,922739928	6,10005909	0,6025732	0,94614277	6,29389921	0,70216002	2,93345208	1,34060859	0,27017415	2,33534499	4,975078531
2	0,26983421	3,176706224	10,35519048	25,5038479	6,600385095	5,567234543	0,53320454	1,12714192	8,76016647	19,3574251	2,76548381	1,60326057	27,0816271	2,654568311	4,844494767
3	0,497721181	1,530512836	15,8318823	26,7607543	1,77337091	0,9253519	0,9246723	3,542684181	6,65385316	2,4884924	2,95879915	4,44664177	25,6765288	4,37154461	5,68036923
4	0,690223991	3,835896849	3,389443018	12,8173729	3,24381785	5,8658722	0,01302749	3,859564782	10,9138565	18,4744926	0,48281942	4,4756643	24,7694579	0,476098361	1,206690771
5	0,426553666	3,006402858	6,918003246	17,1027508	0,841274151	4,609935961	0,67535457	0,99066857	13,7142557	14,7744214	6,01202813	4,43758321	13,2100501	6,461646251	8,314415911
6	0,833580227	2,195897907	2,330498588	18,8906863	2,097382844	2,010532002	0,38284696	2,447066581	16,0430645	24,3183364	5,22153029	2,83466476	18,7807794	5,977920122	4,888037477
7	0,031952741	4,847663094	7,024074834	9,197868931	4,484960847	3,171521127	0,37605479	4,831990681	10,8516601	7,19122772	2,64005382	2,7844141	0,781611511	4,868007151	0,80637752
8	0,990406042	4,484102359	17,25158012	5,98922877	0,145866251	5,363928421	0,38496809	2,737042281	6,13912344	5,57041314	4,14373747	2,25879914	21,5534717	2,826741897	5,274955831
9	0,362089441	2,200460757	11,60579331	22,5008946	5,330522711	4,294239271	0,47885201	3,892415261	10,6423846	11,2575329	5,75022471	5,91928245	9,37327351	4,5439031	3,520456291
10	0,464096311	3,26769124	6,320572907	8,03960757	6,935514011	2,235445261	0,30439159	3,081724061	5,03447091	5,29574117	2,52651945	6,82337752	22,8463237	6,684477761	4,967311941
11	0,819226251	0,944147527	2,088042481	23,6315766	6,960425321	1,007198971	0,21609919	3,816551691	15,5262417	15,4638308	6,55977157	5,35098085	6,0334405	8,624718771	1,973584561
12	0,912508171	4,16905282	8,037409414	27,1817578	1,302423251	4,992406651	0,56539688	4,955976031	14,5443931	2,412347461	6,20838635	5,39128262	13,5589614	1,957768091	2,055577431
13	0,47194121	3,078851368	5,109394022	9,084235751	4,229608361	3,695263731	0,55342659	0,570449231	16,9393349	21,9764768	6,43658465	6,43374919	10,5213171	2,196716941	4,409557661
14	0,005513811	4,546287634	7,588875484	10,7558083	3,544866951	2,703835211	0,64797099	0,191328791	17,5605456	0,48879191	1,87308766	1,74095979	25,090431	1,132199841	1,474735181
15	0,355935921	2,929750375	11,92625149	4,133884591	4,120768961	0,901990251	0,20628708	4,595334871	10,4487812	8,167625321	5,12427678	0,09816801	13,4432112	2,043165071	4,14209931
16	0,681011671	1,118530967	9,869888375	5,04525472	2,263028961	2,066979751	0,4118373	3,316251931	3,6350516	6,809919291	2,18051029	0,39563637	17,549074	4,599217281	1,895882281
17	0,541803741	3,393271482	2,816653981	5,73021191	3,47155228	1,924399331	0,80651275	0,3425959	2,44232576	22,816956	3,71863326	1,580584	16,6877932	2,286681221	1,800357821
18	0,530096411	0,788799948	8,193582895	2,631310121	3,531859171	2,899311828	0,9097271	4,935501931	16,0077957	25,2837749	3,14307767	3,96901829	27,5250852	5,188443531	2,660899961
19	0,396722091	1,741955387	17,55505666	22,945068	2,187955232	0,113805711	0,71572579	0,6279862	10,1770002	19,8421861	1,06179452	6,15453742	26,6552873	4,822079441	1,573126281
20	0,749094991	1,635741223	15,41247152	27,8031441	1,094063441	6,342774181	0,44559458	0,971565111	1,89085233	12,5106171	2,06182478	1,27210639	6,14149892	4,874814581	2,6985317
21	0,591931721	1,466508157	7,541099207	13,4926495	0,506317751	3,053309791	0,00641543	1,834715361	0,8343032	25,5388516	6,95071772	4,76739567	3,485115	2,735928461	6,081313331
22	0,183154641	3,559715814	17,86987593	4,111295821	1,941918861	5,15758951	0,28892126	0,971585931	10,177703	27,0953156	4,06083954	4,62801471	0,06216736	2,972703021	6,993657271
23	0,134213371	3,433782882	14,6240004	22,3089079	1,535629351	6,8009091	0,59842228	4,021041191	17,1956057	26,431634	5,39319042	4,6974822	12,8669234	1,442037911	0,052117511
24	0,908158941	3,846490834	1,125455018	8,15140227	4,315549861	0,065196711	0,21980937	2,168895471	4,80446566	17,6760519	6,73101012	3,37741737	3,06909814	0,881133681	6,485932721
25	0,689169211	2,103580598	15,58305681	5,547922791	0,563633351	5,846377061	0,96409331	0,599382771	0,78853261	12,9098883	6,87309388	5,78881454	19,2703711	1,619575311	0,330919321

a

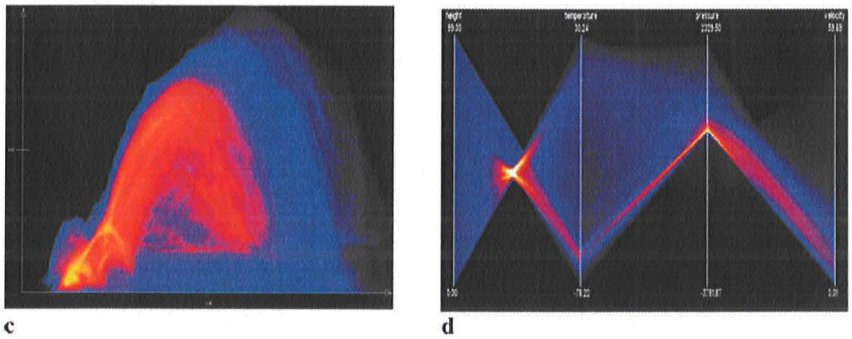


b

Abb. 2 Datenvisualisierung mittels Tables und Graphical Compressed Tables: (a) Eine Table visualisiert 15 Variabeln mit 25 Daten. (b) Eine Graphical Compressed Table gleicher Auflösung visualisiert mit 25 Variablen und über 300 Daten wesentlich mehr Informationen als die Table auf einmal; weiterhin sind ergänzend textuelle Darstellungen (als table lens) möglich [20]

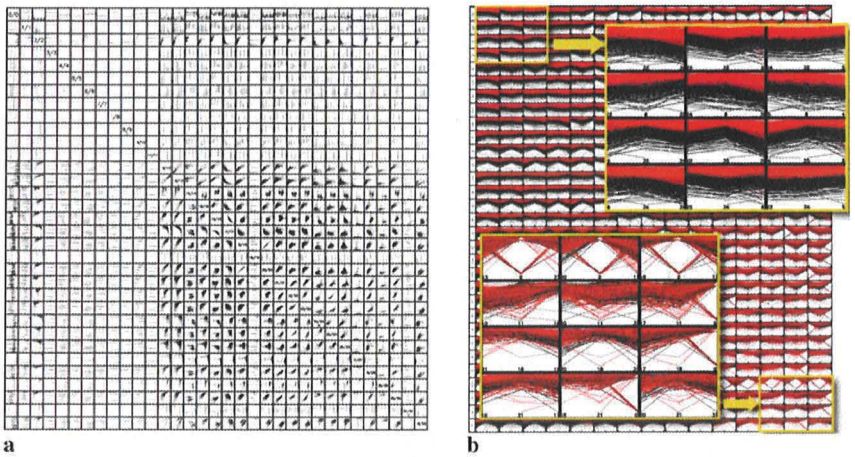


a b



c d

Abb. 3 Scatterplots und Parallele Koordinaten als diskrete (a, b) und als kontinuierliche Daten-visualisierung (c, d) [3, 9]



a b

Abb. 4 SPLOM (a) und PACOM (b) für einen Datensatz mit 32 Variabeln in der Gegenüberstellung

einen Datensatz mit k Variabeln existieren genau l verschiedene Scatterplots: $l = k\frac{(k-1)}{2}$.

Um möglichst übersichtlich verschiedene Visualisierungen, wie z. B. verschiedene Scatterplots eines Datensatzes, darzustellen, bieten sich Visualisierungsmatrizen bzw. Panelmatrizen an. Dabei handelt es sich um eine Menge von Visualisierungen, die als rechteckiges Schema angeordnet sind. Es ergibt sich somit eine vollständige Visualisierung des Datensatzes. Die Visualisierungsmatrizen unterscheiden sich untereinander in der Wahl der verwendeten Visualisierungsmethode.

Streudiagrammmatrizen. Eine Streudiagrammmatrix bzw. Scatterplotmatrix (SPLOM) [5] eines Datensatzes mit k Variabeln ist eine symmetrische $k \times k$ Visualisierungsmatrix M, bei der die i-te Spalte und die j-te Zeile $(0 \leq i, j \leq k-1)$ eindeutig mit Variabeln assoziiert sind, und bei der das Matrixelement der Position $M(i, j)$ ein Scatterplot ist, der die beiden Variabeln i und j darstellt. Derart werden alle orthogonalen Projektionen des Datensatzes in der unteren und in der oberen Dreiecksmatrix visualisiert, wie Abb. 4a aufzeigt. Der direkte Vergleich von Scatterplots unterschiedlicher Variabeln unterstützt insbesondere die Hypothesenbildung multivariater Zusammenhänge in den Daten.

Parallele Koordinaten. Ein Datum wird als Linienzug entlang vertikaler und zueinander paralleler Achsen repräsentiert. Jede Achse korrespondiert mit einer Variable; jeder Achsenschnittpunkt des Linienzuges entspricht dem Wert des Datums bezüglich dieser Variable. Somit wird der Datensatz vollständig abgebildet. Aber: Für unerfahrene Nutzer sind Parallele Koordinaten [12] nur schwer zu interpretieren. Abbildung 3b illustriert dies. Eine Achse steht immer in direkter Verbindung mit zwei anderen, wodurch es schwierig ist, Zusammenhänge zwischen nicht direkt verbundenen Achsen aufzuspüren. Folglich ist die Anordnung der Achsen bedeutend, inwieweit und ob Zusammenhänge zwischen den Variabeln interpretierbar sind (Anordnungsproblem). Wird zudem berücksichtigt, dass bei k Variabeln $k!$ solcher Reihenfolgen existieren, ist die Problematik offensichtlich genau die Parallelen Koordinaten zu finden deren Achsenanordnung eine aussagekräftige Interpretation der Daten durch den Nutzer erlaubt.

Parallele Koordinaten Matrizen. Um das Anordnungsproblem von Parallelen Koordinaten zumindest teilweise zu lösen, wurde in [1] die Parallelen-Koordinaten-Matrix (PACOM) eingeführt. Dabei handelt es sich um eine $k \times p$ Visualisierungsmatrix, bei der in jeder Zeile alle 3D-Achsenkombinationen in Parallelen Koordinaten bezüglich einer (Haupt-) Variablen d dargestellt werden. Dieses wird über die k Spalten aller Variabeln $0 \leq d \leq k-1$ fortgesetzt, wie Abb. 4b illustriert. Ein Zeile kann dabei bis zu $p := (k-1)/2$ unterschiedliche Anordnungselemente enthalten. Auch eine PACOM ist für den Laien nur schwer interpretierbar.

Sowohl für Scatterplots als auch für Parallele Koordinaten existieren zudem kontinuierliche – jedoch weniger performante – Darstellungsmethoden [3, 9] (Abb. 3c, d). Sie erlauben es, Lücken in den Daten zu „überbrücken" und werden vom Nutzer meist als intuitiver empfunden als diskrete Darstellungen.

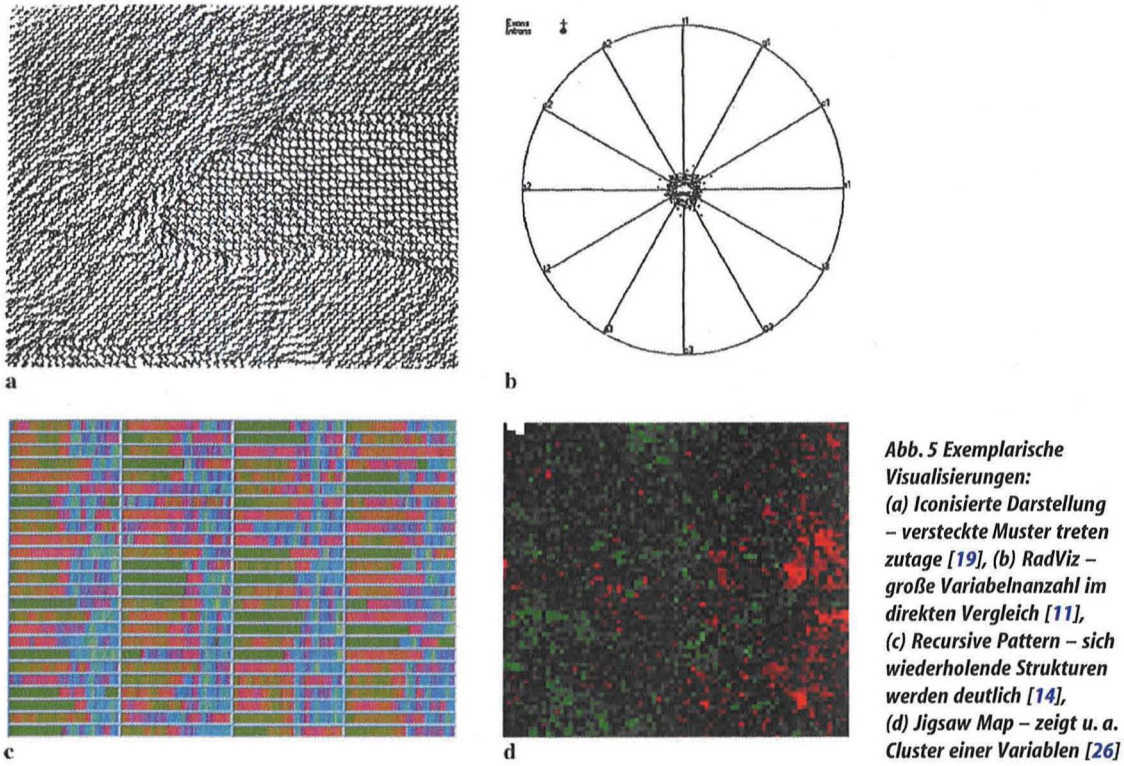


Abb. 5 Exemplarische Visualisierungen: (a) Iconisierte Darstellung – versteckte Muster treten zutage [19], (b) RadViz – große Variabelnanzahl im direkten Vergleich [11], (c) Recursive Pattern – sich wiederholende Strukturen werden deutlich [14], (d) Jigsaw Map – zeigt u. a. Cluster einer Variablen [26]

Die vorgestellten Visualisierungsmatrizen bieten einerseits eine vollständige Sicht auf den Datensatz, skalieren aber andererseits nur schlecht mit zunehmender Variablenanzahl und überfordern den Nutzer daher zunehmend: Es ist kaum mehr möglich, zwischen Visualisierungen mit interessanten und uninteressanten Mustern zu unterscheiden oder überhaupt alle sichten zu können.

Abschließend sei betont, dass es viele weitere Visualisierungsmethoden gibt, welche zumeist einen bestimmten Aspekt des Datensatzes besonders gut darstellen. Einige ausgewählte sind in Abb. 5 dargestellt. Interessierten LeserInnen sind thematisch weiterführende Werke, wie [5, 21, 27], sehr zu empfehlen.

Standardmethoden zur automatischen Datenanalyse

Bei einer Datenvisualisierung besteht auch immer das Problem, dass es zum Verlust von Informationen (visual clutter) und dadurch zu Fehlinterpretationen kommen kann: Durch die begrenzte Fläche des Monitors beispielsweise, oder wenn Strukturen, die im Merkmalsraum getrennt sind, sich in der Visualisierung überlappen. Andererseits nehmen aber auch die Anzahl der Dimensionen und das Datenvolumen beständig zu. Es besteht somit ein Bedarf, Daten automatisch zu analysieren:

Ziel ist es unter anderem, Visualisierungsmethoden zu falsifizieren oder eine multivariate statistische Datenanalyse zu erhalten. Letzteres ist eine erste Annäherung an eine vollständige explorative Analyse. Ziel ist es konsequenterweise auch, interessante Teilmengen zu finden, deren Visualisierung sich „lohnt". Eine vollständige und kontextspezifische Dateninterpretation kann aber auch das beste automatische Verfahren nicht leisten!

Methoden zur Strukturidentifikation

Im Folgenden werden zwei prominente Datensatzstrukturen näherer erläutert.

Korrelation. Eine Korrelation ist eine (mathematisch-statistische) Beziehung zwischen mehreren Variabeln. Ihre Stärke wird durch den mittleren quadratischen Fehler (mean square error, kurz MSE) zwischen einer Funktion und den Datenwerten

selbst beschrieben. Je kleiner der MSE, umso stärker ist die Korrelation, die durch diese Funktion Ausdruck findet. Da praktisch nicht ersichtlich ist, welche Funktion einen optimalen MSE liefert, wird meist für eine Anzahl (multivariater) Polynome steigenden Grades der minimale MSE (durch Wahl geeigneter Koeffizienten) berechnet. Die Funktion mit einem absoluten MSE kleiner einer bestimmten Schwelle wird als Korrelation propagiert, ansonsten gilt, dass keine Korrelation vorliegt. Dieses Vorgehen wird als Regressionsverfahren bezeichnet.

Cluster. Beim Auffinden von Clustern (clustering) werden ähnliche Objekte einer gemeinsamen Gruppe (=Cluster) zugeordnet, mithilfe von z. T. komplexen Ähnlichkeitsfunktionen. Unterschiedliche Studien haben das Verhalten von Ähnlichkeitsfunktionen im Multidimensionalen analysiert [4, 10]: Sie beschreiben, dass die Distanz des (metrisch) entferntesten Objekts – bezüglich eines Anfrageobjekts – mit steigender Dimensionalität nicht so schnell zunimmt, wie die Distanz zum (metrisch) nächsten Nachbarobjekt (Fluch der Dimensionalität bzw. curse of dimensionality):

$$\lim_{d\to\infty} \frac{\text{dist}_{\text{max}} - \text{dist}_{\text{min}}}{\text{dist}_{\text{min}}} = 0.$$

Dies bedeutet, dass die Unterscheidung zwischen dem Nächstem und dem entferntesten Objekt an Bedeutung verliert: Ergebnisse des Clusterings werden somit mit Zunahme der Variablenanzahl kontinuierlich schlechter. Weiterhin wissen wir, dass Cluster zumeist nur in Untermengen (subspaces) der Variabeln auftreten, was umso wahrscheinlicher ist, je mehr Variabeln der Datensatz hat. Daher werden Cluster zumeist nicht mehr global, sondern in lokalen Unterräumen gesucht (*subspace clustering*).

Sowohl für Korrelation als auch für Cluster gilt: Sie sind nicht immer eindeutig und die Ergebnisse variieren mit den eingesetzten Verfahren.

Transformation des Merkmalsraumes

In Disziplinen wie der Logistik oder der Bioinformatik umfassen die Daten z. T. hunderte von Variabeln. Es ist daher von großem Interesse, Merkmalsräume mit weniger Variablen zu finden, die geeignet sind, möglichst strukturerhaltend eine Transformation der Originaldaten zu ermöglichen. Üblicherweise wird dabei zwischen dimensionsreduzierenden und dimensionsselektierenden Techniken unterschieden:

Principal Component Analysis. Die Principal Component Analysis (PCA) [6] transformiert den Merkmalsraum in einen, der den größten Teil der Varianz der Daten enthält. Dabei werden Variablen (Hauptkomponenten bzw. principal components), welche den neuen Merkmalsraum aufspannen, durch die Analyse von Eigenvektoren berechnet.

Multi-dimensional Scaling. Unter multi-dimensional Scaling (MDS) [16] wird ein nichtlinearer iterativer Algorithmus verstanden, welcher Daten in ein Verhältnis zu einer Metrik setzt. Derart resultieren Ähnlichkeiten im neuen Merkmalsraum als Cluster, die wiederum mittels Clusteranalyse detektiert werden können. Entgegen der PCA wirkt das MDS bereits als Strukturfilter, gesteuert durch eine entsprechende Wahl der Metrik.

Self-organizing Maps. Eine Self-organizing Map (SOM, auch Kohonennetz) [15] ist eine unbeaufsichtigte Lernmethode, um den Merkmalsraum auf einen Raum geringerer Dimensionalität zu reduzieren. Sie ist den Methoden der neuronalen Netze zuzuordnen.

Nachteil all dieser Techniken ist, dass die neu generierten Variabeln mit ihren Originalen zumeist in nichtlinearer Weise assoziiert sind. Somit hat der von ihnen generierte Raum nicht immer eine klar erkennbare Bedeutung für den Nutzer.

Als weiterführende Literatur im thematisch näheren Umfeld seien [2, 7, 8, 17, 22] genannt. Etwas allgemeiner ist eine Vertiefung in die Disziplinen der multivariaten Statistik, des Data Mining und des Machine Learning sehr zu empfehlen.

Kombination von Methoden der Datenvisualisierung und Datenanalyse: Beispiele und Chancen

Wir haben bisher Methoden aufgezeigt, um multidimensionale Daten zu visualisieren oder automatisch zu verarbeiten bzw. zu analysieren. Immer einhergehend mit der Problematik einer großen Anzahl von Visualisierungen, die den Nutzer schlicht überfordern oder automatischen Methoden, die nicht geeignet sind, den Kontext mit zu berücksichtigen.

Eine Möglichkeit, dieses Problem aufzulösen, besteht in der zielgerichteten Kombination beider

Methoden. Wie ist das möglich? Zum Einen können automatische Methoden helfen, geeignete Visualisierungsmethoden auszuwählen; zum Anderen können sie genutzt werden, um geeignete Visualisierungen als Ausgangspunkt für eine Mustersuche zu finden. Die letztere Möglichkeit stellen wir exemplarisch in diesem Abschnitt vor.

Es ist dabei das Ziel, automatisch bestimmte Visualisierungen aus der Gesamtheit aller zu ermitteln, wie z. B. in [13]: Insbesondere solche, welche ein Visualisierungsziel (Korrelation, Cluster, Assoziation, etc.) vermeintlich am besten darstellen. Die Methoden, die im Bildraum der Visualisierungen selbst operieren, werden als *Quality Measures* bezeichnet.

Fünf Quality Measures stellen wir nun vor, die die Güte von Korrelation, Klassenseparierbarkeit und Clusterisierung bewerten, am Beispiel von Scatterplots und Parallelen Koordinaten.

Rotating Variance Measure. RVM [24] ist ein Maß, um lineare und nichtlineare Korrelationen in Scatterplots zu bewerten. Um das RVM zu berechnen, wird zunächst ein kontinuierliches Dichtefeld aus dem Scatterplot ermittelt. Für ein Pixel p der Position $x = (x, y)$ wird die maximale Distanz r zum nächsten Punkt im Scatterplot berechnet, zudem die lokale Dichte $\rho = 1/r$. Dieser Schritt ist für weitere Berechnungen essenziell und schließt Ausreißer aus der Bewertung aus. Stark korrelierende Dichtefelder zeigen in der Regel eine auffällig schmale, längliche Struktur mit hohen Dichtewerten, während sonst viele verteilte lokale Maxima im Dichtefeld zu erkennen sind. Um diese Verteilung zu messen, wird die Massenverteilung entlang verschiedener Messrichtungen um das Pixel p berechnet (Abb. 6). Der beste Wert jeder Bildspalte und Richtung wird als Referenz für das RVM verwendet (1); je größer, desto besser ist die Korrelation, wie aus Abb. 7 ersichtlich ist:

$$\mathrm{RVM} = \frac{1}{\sum_x \min_y v(x,y)}, \tag{1}$$

mit der Massenverteilung $v(x, y)$.

Hough Space Measure. HSM [24] ist ein Maß, um Parallele Koordinaten auf Cluster hin zu bewerten. Ein Cluster im Raum der Parallelen Koordinaten kann als eine Häufung von Geraden mit ähnlicher Lage definiert werden. Unter Verwendung dieser Transformation erhalten wir für jeden Nichthintergrundpixel eine sinusförmige Kurve in einer

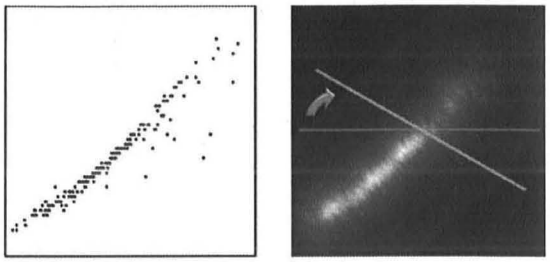


Abb. 6 Scatterplot Beispiel mit Dichtefeld: Für jedes Pixel wird die Masseverteilung entlang verschiedener Richtungen – hier als* blaue Line *dargestellt – berechnet und jeweils der minimalste Wert wird gespeichert

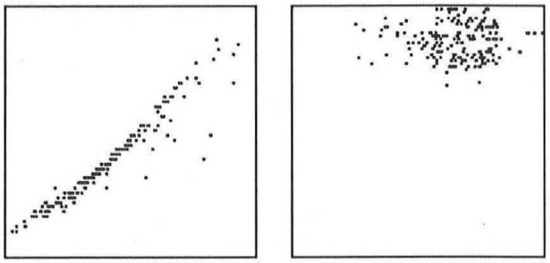


Abb. 7 Bewertung von Scatterplots bezüglich der Korrelation seiner Variabeln: Ein hoher RVM entspricht einem Scatterplot mit stark korrelierenden Variablen, ein niedriger RVM-Wert hingegen deutet auf schwach korrelierende Variablen hin

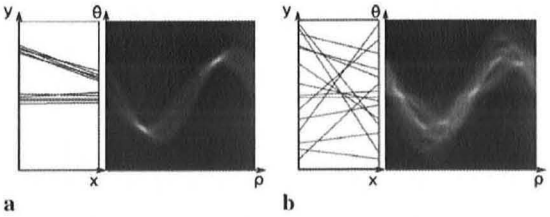


Abb. 8 Beispiele von Parallelen Koordinaten und ihren korrespondierenden Hough-Räumen: (a) enthält zwei wohldefinierte Cluster von Geraden und ist für die Clustererkennung besser geeignet als (b), die keine Cluster enthält

2D-Ebene, dem sogenannten Hough- oder Akkumulatorraum. Ein Schnitt dieser Kurven deutet darauf hin, dass die zugehörigen Pixel auf einer Geraden im Bildraum liegen. Abbildung 8 zeigt zwei Beispiele vor und nach einer Hough-Transformation. Abbildung 8a enthält zwei wohldefinierte Geraden-Cluster und ist für die Clustererkennung besser geeignet als Abb. 8b, die keine Cluster enthält. Die hellen Bereiche der Ebene stellen hier Cluster von Geraden mit ähnlichen Parametern dar. Der Akkumulatorraum ist aufgeteilt in $w \times h$ Zellen. Eine „gute" Visualisie-

rung enthält wohldefinierte Cluster, wenn es Zellen mit hohen Werten im Hough-Raum gibt. Um solche Zellen zu erkennen, berechnen wir den Median m als Schwellwert, der die Akkumulatorfunktion $h(x)$ in zwei identische Teile teilt:

$$\frac{\sum h(x)}{2} = \sum g(x), \quad \text{mit}$$

$$g(x) = \begin{cases} x & \text{wenn } x \leq m; \\ m & \text{sonst.} \end{cases}$$

Das endgültige Maß wird über die Menge der Akkumulatorzellen, die einen höheren Wert als m haben, berechnet.

$$\mathrm{HSM}_{i,j} = 1 - \frac{n_{\text{cells}}}{wh},$$

wobei i, j den Indizes der jeweiligen Variabeln entsprechen. Der errechnete HSM-Wert ist hoch für Bilder, die wohldefinierte Geraden-Cluster enthalten und niedrig für Bilder, die keine Cluster enthalten.

Class Density Measure. CDM [24] ist ein Maß, um die Klassenseparierung von Scatterplots zu messen. Klassen sind durch eine konsistente Farbkodierung kenntlich gemacht. Bei einer gegebenen Menge an Scatterplots eines Datensatzes gilt es, die Plots zu selektieren, welche die Klasse am besten separieren. Durch die Farbkodierung können die Klassen sehr leicht in individuelle Bilder aufgetrennt werden. Es werden nun, zum RVM analoge, Dichtefelder benutzt, um die gegenseitige Überlappung zwischen den Klassen zu berechnen. Die Überlappung ist die Summe der absoluten Differenz der Dichtefelder aller paarweisen Kombinationen der Klassen:

$$\mathrm{CDM} = \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \sum_{i=1}^{P} ||p_k^i - p_l^i||,$$

wobei M der Menge der Dichtefelder, p_k^i dem i-ten Pixel im k-ten Dichtefeld und P der Menge an Pixeln entsprechen. Abbildung 9 zeigt ein Beispiel mit den am besten und den am schlechtesten bewerteten Scatterplots eines Datensatzes.

Distance Consistency Measure. DSC [23]: Jeder Datenwert $x_i; i = \{1, \ldots, s\}$ eines Scatterplots erhält eine Marke, die „true" ist, wenn der Abstand zwischen x_i und seinem Klassen-Zentroiden $c_o(x_i)$ kleiner ist als der Abstand zu allen anderen Klassen-Zentroiden. Ansonsten ist die Marke „false". Ein

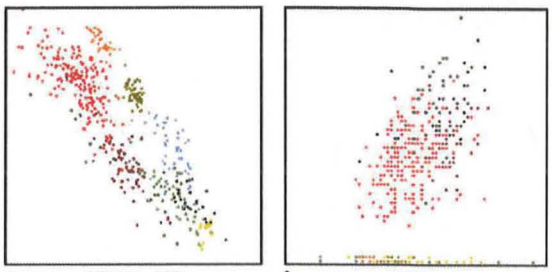


Abb. 9 Bewertung von Scatterplots mit (farbkodierten) Klassen. Ein hoher CDM entspricht einem Scatterplot mit gut separierter Klassendarstellung, ein niedriger CDM hingegen deutet auf eine starke Überlappung zwischen den Klassen hin

Klassen-Zentroid ist der Schwerpunkt aller Werte, die zu einer Klasse c gehören. Das DSC ist nun der Anteil an Marken mit der Belegung „true" bezüglich aller s Datenwerte:

$$\mathrm{DSC} = \frac{|x : \mathrm{Marke}(x, c_o(x)) = true|}{s}.$$

Je größer das DSC, desto besser sind Klassen voneinander separiert, wie Abb. 10a, b aufzeigt. Es eignet sich insbesondere für kompakte Klassen.

Distribution Consistency Measure. DC [23]: In einer ε-Umgebung jedes Datenwertes x_i wird die Anzahl $p_c(x_i)$ der Werte gleicher Klassen c gezählt. Die Entropie

$$H(x_i)_c = -\sum \frac{p_c}{\sum p_c} \log_2 \frac{p_c}{\sum p_c}$$

beschreibt nun die „Dichte" der Klasse c innerhalb dieser Umgebung. Nach dem Aufsummieren dieses Maßes nach (2) kann eine globale Aussage über das Verteilungs- und Separierungsverhalten der Klassen getroffen werden:

$$\mathrm{DC} = 100 - \frac{1}{Z} \sum_{i=1}^{s} \sum_{c} p_c \underbrace{\left(-\sum \frac{p_c}{\sum p_c} \log_2 \frac{p_c}{\sum p_c} \right)}_{H(x_i)_c} \tag{2}$$

mit der Normierung $\frac{1}{Z} = \frac{100}{\log_2(k) \sum x_i \sum_c p_c}$. Je größer das DC $\in \{0, \ldots, 100\}$, desto besser sind die Klassen separiert; wobei das Maß in diesem Fall sehr gut für nichtkonvexe Klassenverteilungen geeignet ist, wie aus Abb. 10c, d ersichtlich.

Quality Measures zeigen erstmals das Potenzial auf, welches die Kombination von Teildisziplinen

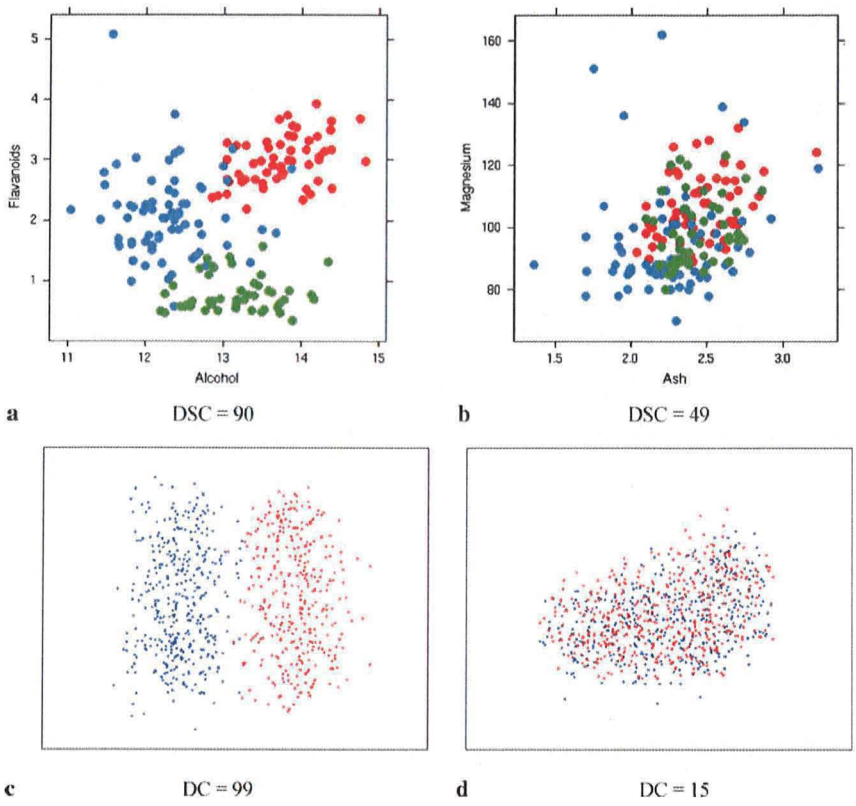


Abb. 10 Bewertung von Separation farbkodierter Klassen (rot, blau, grün) in Scatterplots nach [23]: Ein hoher DSC/DC enspricht einem Scatterplot mit gut separierter Klassendarstellung (a, c); niedrigere DSC/DC Werte hingegen deuten auf eine schlechte Separierung hin (b, d)

für die Visualisierung und Datenanalyse bietet und geben damit die Richtung zukünftiger Forschungen vor.

Ausblick

Die hier vorgestellten Ansätze beschreiben selbstverständlich nur einen kleinen Ausschnitt der Methoden zur visuellen Analyse multidimensionaler Datensätze. Nicht erwähnt wurde die Einbeziehung applikationsspezifischen Wissens, welche zu spezialisierten Ansätzen z. B. für medizinische oder biologische Daten führt (siehe weitere Artikel in diesem Heft). Auch nicht diskutiert werden konnten Fragen der Performance, speziell die Frage, welche Möglichkeiten die rasante Entwicklung der Grafikhardware bietet. Ebenso ergeben sich spezielle Fragestellungen, wenn die Zeitabhängigkeit der Daten explizit untersucht wird. Die eigentliche Stärke visueller Datenanalysemethoden zeigt sich allerdings erst an interaktiven Softwaresystemen, bei denen unterschiedlichste Visualisierungen, Analysemethoden, Selektions- und Interaktionstechniken durch den Nutzer beliebig kombiniert eingesetzt werden können, um einen Datensatz (idealerweise) in Echtzeit zu explorieren und zu analysieren. Geprägt wurde hierfür u. a. in [25] der Begriff Visual Analytics, welches z. Z. im Umfeld der Visualisierung und des Mensch-Computer-Interfaces eines der größten und mit am stärksten wachsenden Forschungsfelder darstellt: An deren Ende steht eine ferne Vision von einem System, das in der Lage ist, alle interessanten Visualisierungen für jedes beliebige Visualisierungsziel eines beliebigen Datensatzes nach Bedarf liefern zu können.

Literatur

1. Albuquerque G, Eisemann M, Lehmann DJ, Theisel H, and Magnor M (2009) Quality-based visualization matrices. Proceedings of Vision, Modeling, and Visualization. Braunschweig
2. Asimov D (1985) The grand tour: a tool for viewing multidimensional data. J Sci Stat Comp 6(1):128–143
3. Bachthaler S, Weiskopf D (2008) Continuous Scatterplots. IEEE T Vis Comput Gr 16(6):1428–1435
4. Beyer SK, Goldstein J, Ramakrishnan R, Shaft U (1999) When is "nearest neighbor" meaningful? In: ICDT '99: Proceedings of the 7th International Conference on Database Theory, London, UK, pp 217–235, Springer
5. Cleveland SW (1993) Visualizing Data. Hobart Press, Summit, NJ
6. Everitt SB, Dunn G (1991) Applied Multivariate Data Analysis. Arnold

7. Fisherkeller AM, Friedman HJ, Tukey WJ (1987) Prim-9: an interactive multidimensional data display and analysis system. In: Sleveland WS (ed) Dynamic Graphics for Statistics. Chapman and Hall, New York
8. Friedman HJ (1987) Exploratory projection pursuit. J Am Stat Assoc 82:249–266
9. Heinrich J, Weiskopf D (2009) Continuous Parallel Coordinates. IEEE T Vis Comput Gr (Proceedings Visualization/Information Visualization 2009) 15(6):1531–1538
10. Hinneburg A, Aggarwal CC, Keim AD (2000) What is the nearest neighbor in high dimensional spaces? In: VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases, San Francisco, CA, USA, pp 506–515, Morgan Kaufmann Publishers Inc.
11. Hoffman P, Grinstein G, Marx K, Grosse I, and Stanley E (1997) Dna visual and analytic data mining. In: Proceedings of the 8th conference on Visualization, Phoenix, AZ, pp 437ff
12. Inselberg A (2009) Parallel Coordinates. Springer, Berlin
13. Johansson S, Johansson J (2009) Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE T Vis Comput Gr 15(6):993–1000
14. Keim D, Ankerst M, Kriegel H (1995) Recursive pattern: a technique for visualizing very large amounts of data. In: Proc. Visualization 1995 IEEE Computer Society Press, Washington, DC, pp 279–287
15. Kohonen T (1995) Self Organizing Maps. Springer
16. Mead A (1992) Review of the development of multidimensional scaling methods, vol 33. The Statistician 41:27–39
17. Moore SD, McCabe PG (1999) Introduction to the Practice of Statistics. WH Freeman, New York, NY
18. Nocke T (2007) Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung. Dissertation, Universität Rostock, Fakultät für Informatik und Elektrotechnik
19. Picket MR, Grindstein G (1988) Iconographics displays for visualizing multidimensional data. In: Proc. IEEE Conference on Systems, Man and Cybernetics, Beijing and Shenyang, pp 514–519
20. Rao R, Card KS (1994) The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp 318–322
21. Schumann H, Müller W (2000) Visualisierung: Grundlagen und allgemeine Methoden. Springer
22. Seo J, Shneiderman B (2005) A rank-by-feature framework for interactive exploration of multidimensional data. Inform Visual 4(2):96–113
23. Sips M, Neubert B, Lewis PJ, Hanrahan P (2009) Selecting good views of high-dimensional data using class consistency. Comput Graph Forum (Proc. EuroVis 2009) 28(3):831–838
24. Tatu A, Albuquerque G, Eisemann M, Schneidewind J, Theisel H, Magnor M, Keim D (2009) Combining automated analysis and visualization techniques for effective exploration of high dimensional data. In: IEEE Symposium on Visual Analytics Science and Technology, New Jersey, pp 59–66
25. Thomas JJ, Cook KA (2006) A Visual Analytics Agenda. IEEE Comput Graph 10–13
26. Wattenberg M (2005) A note on space-filling visualizations and space-filling curves. In: Proc. of the 2005 IEEE Symposium on Information Visualization, pp 181–186
27. Wong PC, Bergeron RD (1997) 30 Years of Multidimensional Multivariate Visualization. In: Scientific Visualization, Overviews, Methodologies, and Techniques. IEEE Computer Society Press, Washington, DC, pp 3–33