Visualisierung und Analyse multi-dimensionaler Datensätze

Dirk J. Lehmann* Georgia Albuquerque† Martin Eisemann‡ Andrada Tatu§
Daniel Keim¶ Heidrun Schumann‖ Marcus Magnor** Holger Theisel††

Zusammenfassung

Für multi-dimensionale 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 Ihnen 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.

Schlüsselwörter: Visualisierung, Datenanalyse, Data Mining, Visual Analytics, Statistik

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 to you 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.

Keywords: Visualization, Data Analysis, Data Mining, Visual Analytics, Statistic

Einleitung

Im Sommer 1854 brach eine der schlimmsten Cholera-Epidemien in London aus; nach bereits vier Ausbrüchen innerhalb von nur 23 Jahren schritt dieser so schnell fort und war so tödlich wie noch keiner zuvor. Ohne Vorankündigung starben innerhalb weniger Tage allein im Stadtteil Soho weit über hundert 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 Miasmen-Theorie 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 Cholera-Opfer 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 Cholera-Opfer 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, den Schwengel der Pumpe abzumontieren, was die Bewohner zwang, ihr Wasser an anderen Pumpen zu holen. Innerhalb weniger Tage ging die Opferzahl 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, nicht-zufällig 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,

*dirk@isg.cs.uni-magdeburg.de; Universität Magdeburg
†georgia@cg.cs.tu-bs.de; TU Braunschweig
‡eisemann@cg.cs.tu-bs.de; TU Braunschweig
§tatu@dbvis.inf.uni-konstanz.de; Universität Konstanz
¶keim@dbvis.inf.uni-konstanz.de; Universität Konstanz
‖schumann@informatik.uni-rostock.de; Universität Rostock
**magnor@cg.cs.tu-bs.de; TU Braunschweig
††theisel@isg.cs.uni-magdeburg.de; Universität Magdeburg

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 multi-dimensionalen 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, oder ... Um der wahren Ursache eines phänomenologischen Zusammenhangs auf den Grund zu gehen, braucht es daher zweierlei: der Suche nach allen scheinbaren Zusammenhängen in einem Datensatz (exhaustive search) sowie eines Menschen, der die gefundenen Zusammenhänge kausal verknüpfen und Schein-Zusammenhänge auf ihre wahren Ursachen zurückführen kann. Das genannte historische Beispiel beschreibt einen

Abbildung 1: In der "Ghost Map" von John Snow sind die Wohnorte der Cholera-Opfer eingetragen; es wird deutlich, dass sich die Todesfälle um eine konkrete Wasserpumpe herum häufen.

einfachen Fall von multi-dimensionalen Daten. Ebenso einfach (und doch wirksam) ist die Wahl der Visualierungstechnik. 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 Kapiteln geben wir eine formale Definition von multi-dimensionalen Daten, beschreiben existierende Standardtechniken zur automatischen und visuellen Datenanalyse und zeigen an Beispielen einige aktuelle Arbeiten zur sinnvollen Kombination solcher Techniken.

Multi-dimensionale Datensätze

Um imstande zu sein ein (visualisiertes) Muster zu interpretieren bzw. um das Gesehene in einen Sinn-Kontext einzuordnen, ist ein allgemeines Verständnis von der Struktur zugrundeliegender multi-dimensionaler Datensätze unerlässlich. Aus diesem Grund werden sie in diesem Abschnitt eingeführt.

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, u.s.w. 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 ($\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{R}$, ...).
  - **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) handel 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 *multi-dimensionaler Datensatz* verstehen, als ein Datensatz mit mindestens zwei (oder mehr) skalaren Variabeln. 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 Datensatz, 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 zeitaufwändige Berechnungen parallel (gleichzeitig) auszuführen, anstatt seriell (nacheinander).

Standardmethoden zur Visualisierung multi-dimensioneller Daten

Bei den Methoden zur Datenvisualisierung wird zwischen der Visualisierung von physikalischen Daten (scientific visualization) und abstrakten Daten (information visualization) unterschieden: Dabei sind phys. 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. Ein 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 jedoch sind Tendenzen ersichtlich beide Teilgebiete mehr und mehr zu konsolidieren, um synergetisch weitere Fortschritte zu forcieren. Dessen ungeachtet, gibt es beiderseits etablierte Methoden multi-dimensionale Datensätze zu visualisieren. Im Weiteren stellen wir insbesondere drei typische Beispiele zur Visualisierung multi-dimensionaler Daten vor, die in Disziplinen, wie der Systembiologie oder der Meteorologie, Anwendung finden.

Tabellen Eine intuitive und wenig aufwändige 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. 2 (a) 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.

Graphisch Komprimierte Tabellen Dem letztgenannten Nachteil begegnen graphisch 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. Abb. 2 (b): 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 notwendig 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 die Abb. 3 (a) verdeutlicht. Aussagen über multivariate Zusammenhänge (z.B. multivariate Korrelation) sind allerdings kaum möglich, zudem gehen Informationen über die Daten-Anzahl verloren, die auf die gleiche Position im Scatterplot abgebildet werden. Transparenzen zu verwenden kann diesem Effekt bis zu einem gewissen Grad entgegenwirken. Für 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 Visualisierungs-Matrizen bzw. Panel Matrizen an. Dabei handelt es sich um eine

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

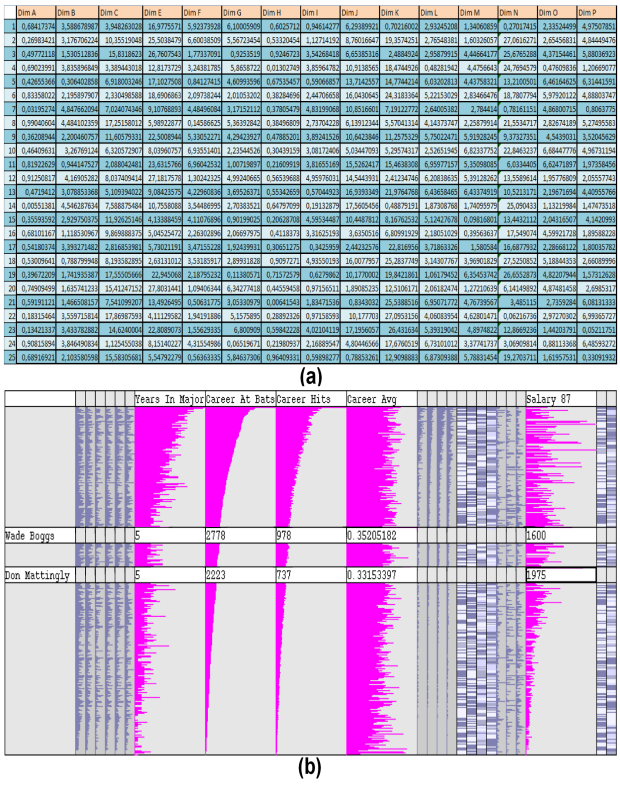


Abbildung 2: Daten Visualisierung mittels Tables und Graphical Compressed Tables: (a) Eine Table visualisiert 15 Variabeln mit 25 Daten. (b) Eine Graphical Compressed Table gleicher Auflösung visualisiert mit 25 Variablen und über 300 Daten wesentlich mehr Informationen als die Table auf einmal; weiterhin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

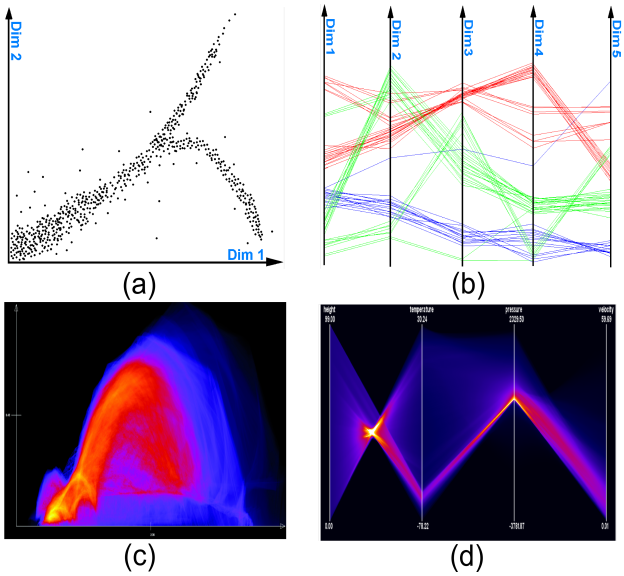


Abbildung 3: Scatterplots und Parallele Koordinaten als diskrete (a-b) und als kontinuierliche Datenvisualisierung [3, 9] (c-d).

Menge von Visualisierungen, die als rechteckiges Schema angeordnet sind. Es ergibt sich somit eine vollständige Visualisierung des Datensatzes. Die Visualisierungs-Matrizen untereinander unterscheiden sich in der Wahl der verwendeten Visualisierungsmethode.

Streudiagramm Matrizen Eine Streudiagramm Matrix bzw. Scatterplot Matrix (SPLOM) [5] eines Datensatzes mit k Variabeln ist eine symmetrische $k \times k$ Matrix 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 Matrix-Element der Position $M(i, j)$ ein Scatterplot ist, der die beiden Variabeln i und j darstellt. Derart werde alle orthogonalen Projektionen des Datensatzes in der unteren und in der oberen Dreiecksmatrix visualisiert, wie Abb. 4 (a) aufzeigt. Der direkte Vergleich von Scatterplots unterschiedlicher Variabeln unterstützt insbesondere die Hyphothesenbildung 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 Achsen-Schnittpunkt 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. Abb. 3 (b) 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 relevanten Parallelen Koordinaten zu finden.

Parallele Koordinaten Matrizen Um das Anordungsproblem von Parallelen Koordinaten zumindest teilweise zu lösen, wurde in [1] die Parallelen Koordinaten Matrix (PACOM) eingeführt. Dabei handelt es sich um eine $k \times p$ Matrix, 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. 4 (b) illustriert. Ein Zeile kann dabei bis zu $p := (k-1)/2$ unterschiedliche Anordnungselemente enthalten. Auch eine PACOM ist für den Laien nur schwer interpretierbar.

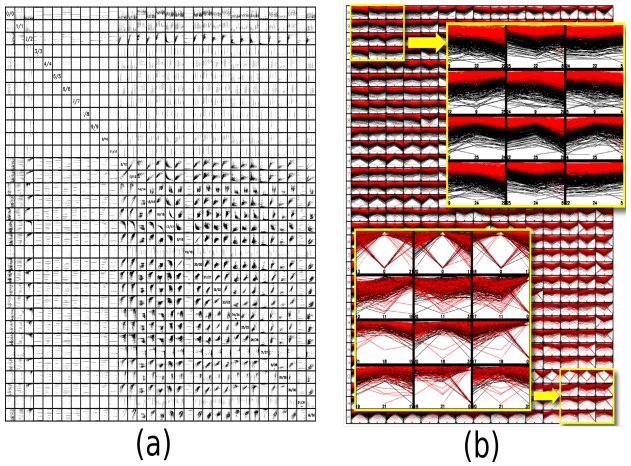


Abbildung 4: SPLOM (a) und PACOM (b) für einen Datensatz mit 32 Variabeln in der Gegenüberstellung.

Sowohl für Scatterplots als auch für Parallele Koordinaten existieren zudem kontinuierliche – jedoch weniger performante – Darstellungsmethoden [3, 9] (Abb. 3 (c-d)). Sie erlauben es Lücken in den Daten zu

"überbrücken" und werden von dem Nutzer meist als intuitiver empfunden als diskrete Darstellungen.

Bezüglich der Visualisierungs-Matrizen ist zu beachten, dass sie schlecht mit zunehmender Variablenanzahl skalieren und den Nutzer daher zunehmend überfordern: 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. Dem interessierten Leser sind thematisch weiterführende Werke, wie [27], [5] oder [21] sehr zu empfehlen.

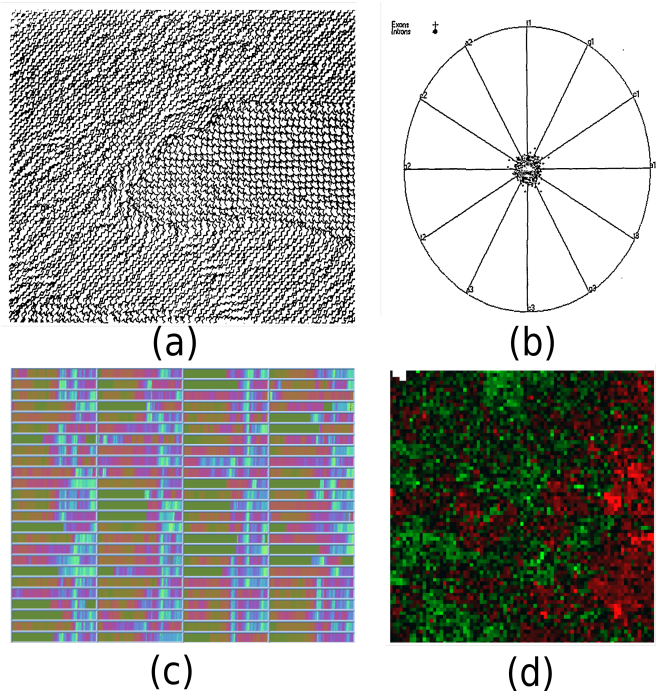


Abbildung 5: Exemplarische Visualisierungen: (a) Iconisierte Darstellung - versteckte Muster treten "zutage" [19], (b) RadViz - große Variabelnanzahl im direkten Vergleich [11], (c) Recursive Pattern - sich wiederholende Strukturen werden deutlich [14], (d) Jigsaw Map - zeigt u.a. Cluster einer Variablen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

Standardmethoden zur automatischen Datenanalyse

Bei einer Datenvisualisierung besteht leider auch immer das Problem, dass es zu Verwirrungen bei der Interpretation kommen kann (visual clutter): 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 u.a. Visualisierungs-Methoden zu falsifizieren oder eine multivariate statistische Datenanalyse zu erhalten. Letzteres ist eine erste Annäherung an eine vollständige explorative Analyse. Ziel ist 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 Datensatz-Strukturen näherer erläutert.

Korrelation Eine Korrelation ist eine (mathematisch-statistische) Beziehung zwischen mehreren Variabeln. Die Stärke dieser, 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 berechnet (durch Wahl geeigneter Koeffizienten). 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, mit Hilfe 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 Objektes – bezüglich eines Anfrageobjektes – mit steigender Dimensionalität, nicht so schnell zunimmt, wie die Distanz zum (metrisch) nächsten Nachbarobjekt (Fluch der Dimensionalität bzw. curse of dimensionality):

$$\lim_{d\to\infty} \frac{dist_{max} - dist_{min}}{dist_{min}} = 0.$$

Dies bedeutet, dass die Unterscheidung zwischen dem Nächstem und dem entferntesten Objekt an Bedeutung verliert: Ergebnisse des Clusterings werden somit kontinuierlich schlechter, mit Zunahme der Variablenanzahl. 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 um möglichst strukturerhaltend eine Transformation der original Daten zu ermöglichen. Üblicherweise wird dabei zwischen dimensionsreduzierenden und dimensionsselektierenden Techniken unterschieden:

Principal Component Analysis Die Principal Component Analysis (PCA) [6] projiziert 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 nicht-linearer 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 nicht linearer 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 [7], [17], [8], [22] und [2] genannt. Etwas allgemeiner ist eine Vertiefung in die Disziplinen der multivariaten Statistik, des Data Mining und des Machine Learning sehr zu empfehlen.

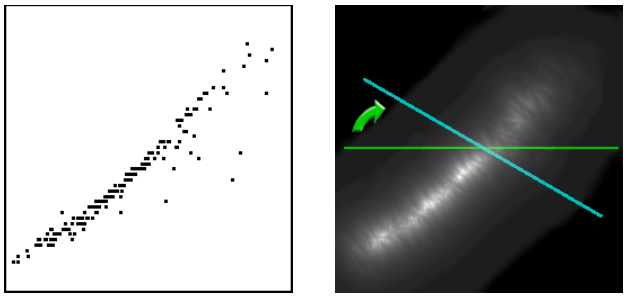
Kombination von Methoden der Datenvisualisierung und Datenanalyse: Beispiele und Chancen

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 Dilemma 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.

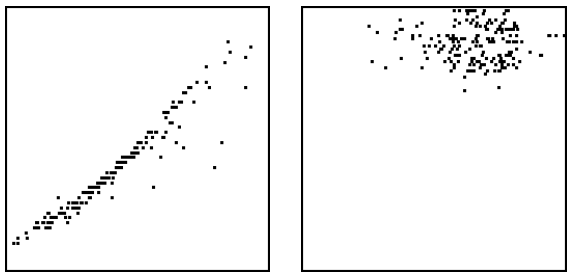
Es ist dabei das Ziel, automatisch, bestimmte Visualisierungen aus der Gesamtheit aller zu ermitteln, wie z.B. in [13]: Insbesondere solche, welche ein Visualisierungsziel (Korrelation, Cluster, Assoziation, etc.) vermeintlich am besten darstellen. Die Methoden, die im Bildraum der Visualisierungen selbst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

Fünf Quality Measures, am Beispiel der Scatterplots und der Parallelen Koordinaten multi-dimensionaler Daten, führen wir nun auf, die bezüglich der Korrelation, Klassenseparierthеit und der Cluster hin bewerten.



(a) (b)

Abbildung 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.



$RVM = 100$ $RVM = 0$

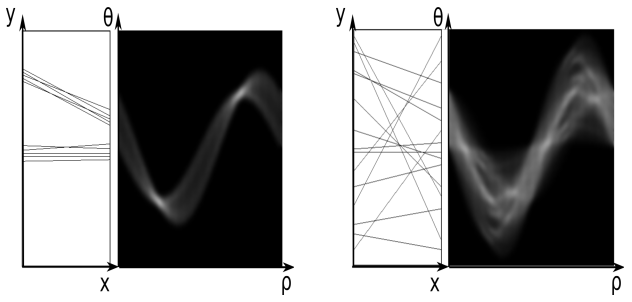
Abbildung 7: Bewertung von Scatterplots bezüglich der Korrelation seiner Variabeln: Ein hoher RVM entspricht einem Scatterplot mit stark korrelierenden Variablen, ein niedriger RVM-Wert hingegen deuten auf schwach korrelierende Variablen hin.

Rotating Variance Measure (RVM) [24] ist ein Maß, um lineare und nicht-lineare Korrelationen in Scatterplots zu bewerten. Um das RVM zu berechnen, wird zunächst ein kontinuierliches Dichtefeld aus dem Scatterplot ermittelt. Für ein Pixel $\mathbf{p}$ der Position $\mathbf{x} = (x, y)$ wird die maximale Distanz r zum nächsten Punkt im Scatterplot berechnet, zudem die lokale Dichte $\rho = 1/r$. Dieser Schritt ist für weitere Berechnungen essentiell und schliesst Ausreißer aus der Bewertung aus. Stark korrelierende Dichtefelder zeigen in der Regel eine auffällige 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 $\mathbf{p}$ berechnet (Abb. 6). Der beste Wert jeder Bildspalte und Richtung wird als Referenz für das RVM verwendet (Gleichung 1); je größer, desto besser ist die Korrelation, wie aus Abb. 7 ersichtlich ist:

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

mit der Massenverteilung $\nu(x, y)$.

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 nicht-Hintergrund Pixel eine sinusförmige Kurve in einer 2D Ebene, dem sogenannten Hough- oder Akkumulatorraum. Ein Schnitt dieser Kurven deutet darauf hin, dass die zugehörigen Pixel auf einer Geraden im Bildraum liegen. Abb. 8 zeigt zwei Bei-



(a) (b)

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

spiele vor und nach einer Hough-Transformation. Abb. 8(a) enthält zwei wohldefinierte Geraden-Cluster und ist für die Cluster-Erkennung besser geeignet als Abb. 8(b), 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" Visualisierung enthält wohldefinierte Cluster, wenn es Zellen mit hohen Werten im Houghraum gibt. Um solche Zellen zu erkennen, berechnen wir den Median m als Schwellwert, der die Akkumulatorfunktion $h(x)$ in zwei identische Teile teilt:

$$\begin{aligned} \frac{\sum h(x)}{2} &= \sum g(x) \quad \text{, mit} \\ g(x) &= \begin{cases} x & \text{wenn } x \leq m; \\ m & \text{sonst.} \end{cases} \end{aligned}$$

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

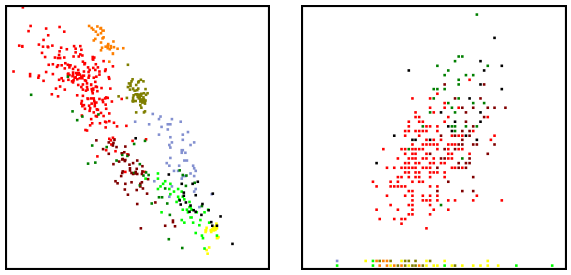
$$\text{HSM}_{i,j} = 1 - \frac{n_{cells}}{wh},$$

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:

$$\text{CDM} = \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \sum_{i=1}^{P} ||\mathbf{p}_k^i - \mathbf{p}_l^i|| \quad ,$$

wobei M der Menge der Dichtefelder, $\mathbf{p}_k^i$ dem i-tem Pixel im k-tem Dichtefeld und P der Menge an Pixeln entsprechen. Abb. 9 zeigt ein Beispiel mit den am besten und den am schlechtesten bewerteten Scatterplots eines Datensatzes.



$CDM = 100$ $CDM = 0$

Abbildung 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.

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 Klassenzentroiden $c_o(x_i)$ kleiner ist als der Abstand zu allen anderen Klassen-Zentroiden. Ansonsten ist die Marke "false". Ein Klassenzentroid meint den 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:

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

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

Distribution Consistency Measure (DC) [23] In einer ε-Umgebung jedes Datenwertes x_i werden die Anzahl $p_c(x_i)$ der Werte gleicher Klasse 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 Gleichung 2 kann ein globale Aussage über das Verteilungs- und Separierungsverhalten der Klassen getroffen werden:

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

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 nicht konvexe Klassenverteilungen geeignet ist, wie aus Abb. 10 (c-d) ersichtlich.

Quality Measures zeigen erstmals das Potential auf, welches die Kombination von Teildisziplinen 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 multi-dimensionaler Datensätze.

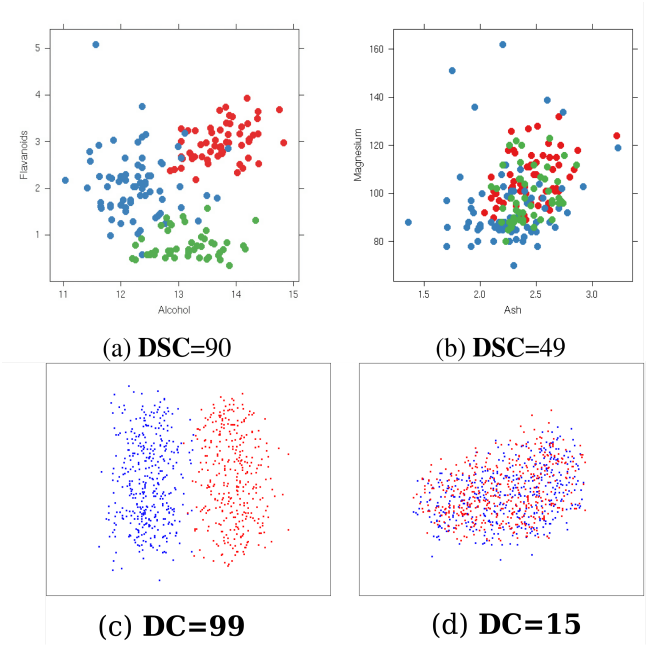


Abbildung 10: Bewertung von Separation farbkodierter Klassen (rot, blau, grün) in Scatterplots nach [23]: Ein hoher DSC/DC enspricht einem Scatterplot mit gut separierter Klassendarstellung (a und c); niedrigere DSC/DC Werte hingegen deuten auf eine schlechte Separierung hin (b und d).

Nicht erwähnt wurde die Einbeziehung applikationsspezifischen Wissens, welche zu spezialisierten Ansätzen z.B. für medizinische oder biologische Daten führen (siehe weitere Artikel in diesem Heft). Auch nicht diskutiert werden konnten Fragen der Performance, speziell die Frage, welche Möglichkeiten die rasante Entwicklung der Graphik-Hardware 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.Zt. 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 on demand liefern zu können.

Literatur

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Quality-based visualization matrices. *Proceedings of Vision, Modeling, and Visualization (VMV 2009)*, 2009.

[2] D. Asimov. The grand tour: a tool for viewing multidimensional data. *Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[3] S. Bachthaler and D. Weiskopf. Continuous Scatterplots . *IEEE Transactions on Visualization and Computer Graphics* , 1428–1435 , 2008 .

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 217–235, London, UK, 1999. Springer-Verlag ISBN 3-540-65452-6.

[5] W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, New Jersey ISBN 0-9634884-0-6, 1993.

[6] B. S. Everitt and G. Dunn. *Applied Multivariate Data Analysis*. Arnold, 1991.

[7] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. *Prim-9: An interactive multi-dimensional data display and analysis system*, volume In W. S. Sleveland, editor. Chapman and Hall, New York, 1987.

[8] J. H. Friedman. Exploratory projection pursuit. *Jurnal of the American Statistical Association*, (82):249 – 266, Mar. 1987.

[9] J. Heinrich and D. Weiskopf. Continuous Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2009)*, 15(6), 2009.

[10] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 506–515, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[11] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. Dna visual and analytic data mining. *Proceedings of the 8th conference on Visualization*, page 437 ff, 1997.

[12] A. Inselberg. *Parallel Coordinates*. Springer Verlag Berlin, ISBN 0387215077, 2009.

[13] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[14] D. Keim, M. Ankerst, and H. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. *Proc. Visualization 1995 IEEE Computer Society Press*, pages 279–287, 1995.

[15] T. Kohonen. *Self Organizing Maps*. Springer Verlag, 1995.

[16] A. Mead. *Review of the development of multidimensional scaling methods*, volume 33. The Statistician, 1992.

[17] D. S. Moore and G. P. McCabe. *Introduction to the Practice of Statistics*. W. H. Freeman & Co., New York, NY, USA, 1999.

[18] T. Nocke. *Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung*. Dissertation, Universität Rostock, Fakultät für Informatik und Elektrotechnik, 2007.

[19] R. M. Picket and G. Grindstein. Iconographics displays for visualizing multidimensional data. *Proc. IEEE Conference on Systems, Man and Cybernetics*, pages 514–519, 1988.

[20] R. Rao and S. K. Card. 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*, 1994.

[21] H. Schumann and W. Müller. *Visualisierung: Grundlagen und allgemeine Methoden*. Springer Verlag, ISBN 3-540-64944-1, 2000.

[22] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multi-dimensional data. *Information visualization*, 4(2):96–113, 2005.

[23] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis 2009)*, 28(3):831–838, 2009.

[24] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high dimensional data. *IEEE Symposium on Visual Analytics Science and Technology*, pages 59–66, 2009.

[25] J.J. Thomas and K.A. Cook. A Visual Analytics Agenda. *IEEE Computer Graphics and Applications* , 10-13 , 2006 .

[26] M. Wattenberg. A note on space-filling visualizations and space-filling curves. *Proc. of the 2005 IEEE Symposium on Information Visualization*, 2005.

[27] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. *IEEE Computer Society Press*, pages 37–44, 1992.