Visual Analytics von Mustern in hochdimensionalen Daten*

Andrada Tatu

tatu@dbvis.inf.uni-konstanz.de

Abstract: Der tägliche technologische Fortschritt ermöglicht die Speicherung und Verarbeitung riesiger Datenmengen. Die Verwendung solcher Datenarchive bringt neue Herausforderungen an Analysetechniken mit sich. Die große Anzahl der Datenpunkte sowie der beschreibenden Attribute (Dimensionen) erschweren die Datenanalyse. Diese Arbeit widmet sich numerischen hochdimensionalen Daten, die durch den *Fluch der Dimensionalität* und der exponentiellen Anzahl von Attributkombinationen neue Methoden der Analyse erforderlich machen. Mithilfe visueller Analysemethoden werden hier Qualitätsmaße vorgeschlagen, um Visualisierungen zu bewerten, die wichtige Strukturen aufweisen. Ebenso werden automatisch-interaktive Verfahren vorgestellt, die die Suche nach interessanten Mustern unterstützen.

1 Einleitung

Bedingt durch den technologischen Fortschritt der letzten Jahrzehnte sind kommerzielle Applikationen heute in der Lage, riesige Datenmengen zu erzeugen, zu speichern und zu verarbeiten. Diese Entwicklung beeinflusst auch die Natur der erzeugten Daten, d.h. dass für jeden Dateneintrag unterschiedliche Aspekte in der gleichen Datenbank gespeichert werden. Oft sind die beschreibenden Attribute numerisch. Datensätze, die mehr als fünf solcher Attribute (Dimensionen) beinhalten, nenne ich hochdimensional.

Meine Dissertation [Tat13] behandelt die Fragestellung, wie interessante Muster (bedeutende Informationen) in hochdimensionalen Räumen gefunden werden können. Diese Aufgabenstellung ist durch das Problem des *Fluches der Dimensionalität* äußerst herausfordernd. Dieses Problem besagt, dass Daten im hochdimensionalen Raum spärlich vorkommen.

Herkömmliche Analysetechniken werden dadurch beeinträchtigt. Automatische Methoden müssen die Datenkomplexität nicht nur im Hinblick auf ihre Laufzeit, sondern auch auf ihre Berechnungsfunktionen (z.B. Distanzfunktionen) einbeziehen. Außerdem wird die visuelle Exploration dieser Daten durch die Zweidimensionalität der Darstellungen beeinträchtigt. Die Herausforderung des Informationsgewinnungsprozesses aus hochdimensionalen Daten besteht darin, die Dimensionalität der Daten geschickt zu reduzieren – ein Prozess der auch für die Visualisierung dieser Daten ausschlaggebend ist.

*Englischer Titel der Dissertation: “Visual Analytics of Patterns in High-Dimensional Data”

Eine Möglichkeit der Dimensionsreduzierung ist die Selektion von Dimensionen. Die Visualisierung dieser Daten zur Informationsgewinnung beinhaltet, dass eine geschickte Abbildung zwischen Datendimensionen und Visualisierungsmerkmalen gefunden wird. Somit können hochdimensionale Daten auf den 2D Raum projiziert werden und weitere Merkmale auf weitere visuelle Attribute wie Farbe, Position, Richtung, etc. abgebildet werden. So kann beispielsweise ein zehndimensionaler Datensatz in einem Scatterplot dargestellt werden, indem jeweils zwei Dimensionen auf der x- bzw. y-Achse abgebildet werden und die dritte Dimension die Farbe und die vierte die Größe der Punkte bestimmt (siehe Abbildung 1). Somit ergeben sich für diesen Datensatz für die Selektion von vier Attributen und für die Darstellung mit vier visuellen Parametern über 5000 mögliche Abbildungen[1].

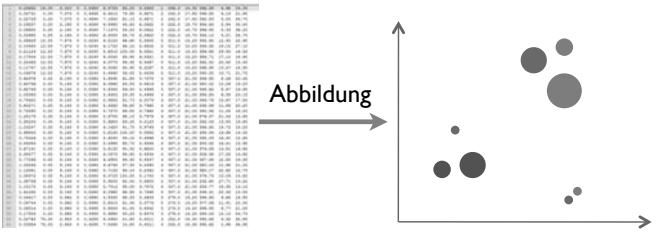

Abbildung 1: Ein 10 dimensionaler Datensatz hat über 5000 mögliche Abbildungen mit vier visuellen Parametern.

Eine manuelle Durchsuchung dieser Darstellungsräume ist nicht möglich. Meine Dissertation beruht auf dem Konzept, dass die Suche nach interessanten Mustern in hochdimensionalen Datenmengen mit einem kombinierten Ansatz von automatischen, visuellen und interaktiven Methoden durchgeführt werden kann.

2 Qualitätsmaße zum Bewerten von Mustern in hochdimensionalen Räumen

Die hohe Anzahl an möglichen Abbildungen von hochdimensionalen Datensätzen erfordert eine automatische Durchsuchung der Darstellungen. Die Ausprägung der Muster einer Visualisierung kann durch sogenannte Qualitätsmaße gemessen werden. Durch diese automatischen Funktionen kann die große Menge an hochdimensionalen Visualisierungen eingegrenzt und dem Benutzer eine ausgewählte Menge zur weiteren Untersuchung zur Verfügung gestellt werden. Ich schlage Qualitätsmaße für Scatterplots [CLN86] und Parallele Koordinaten [Ins85] vor – den am häufigsten verwendeten Visualisierungen für hochdimensionale Daten. Dabei werden nichtklassifizierte und klassifizierte Daten in Betracht gezogen. Diese Maße konzentrieren sich auf unterschiedliche Aufgaben, wie die Identifikation von Gruppen oder Korrelationen. Dabei werden unterschiedliche Zielfunktionen behandelt und die Ergebnisse abhängig davon sortiert. In Abbildung 2 werden die

[1] $\binom{10}{4} \cdot 4! = \frac{10!}{6!}$

Ergebnisse von drei solchen Zielfunktionen für parallele Koordinaten mit den jeweils zwei besten Ergebnissen links und der schlechtesten gefundenen Abbildung rechts dargestellt. Diese Methoden selektieren die besten vier Dimensionen, um diese Funktionen zu maximieren.

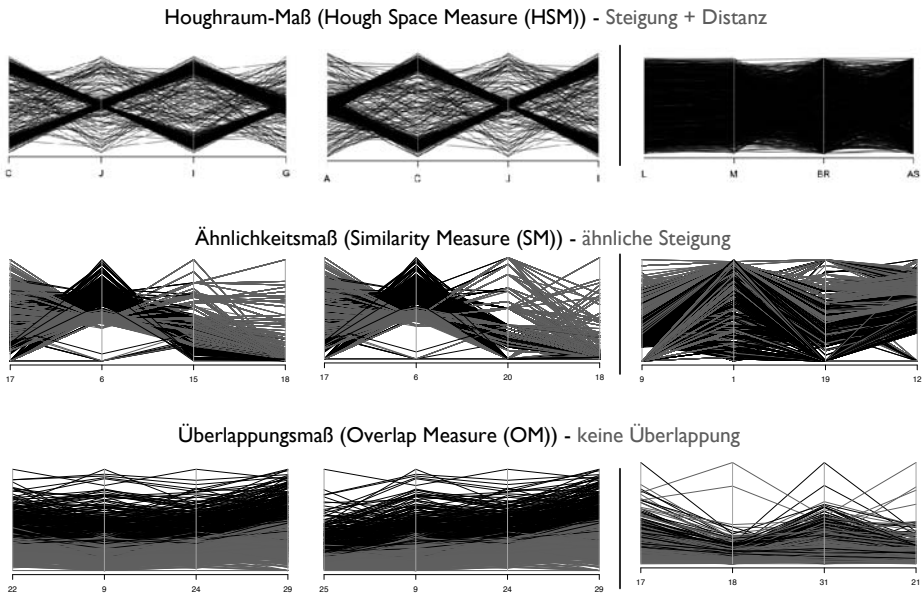

Abbildung 2: Qualitätsmaße für Parallele Koordinaten mit verschiedenen Zielfunktionen (blau) - jeweils die zwei Besten (links) und das schlechteste Ergebnis (rechts).

In Abbildung 3 werden die Ergebnisse der Qualitätsmaße für Scatterplots dargestellt. Dabei kann man auch erkennen, dass zur Identifikation von Gruppen unterschiedliche Kriterien für die automatische Suche eingesetzt werden können.

Um sicherzustellen, dass die automatischen Qualitätsmaße das menschliche Handeln widerspiegeln, werden diese Techniken (1) hinsichtlich ihrer Fähigkeit, Cluster in unterschiedlichen realen und synthetischen Datensätzen zu finden, und (2) hinsichtlich ihrer Korrelation mit der menschlichen Wahrnehmung untersucht. Der ausführlichen Diskussion dieser Resultate folgen Überlegungen für die zukünftige Forschung. An dieser Stelle möchte ich eine dieser Überlegungen ansprechen. Diese Studie untersucht, inwieweit die automatischen Qualitätsmaße die menschliche Wahrnehmung widerspiegeln, und zeigt deutliche Korrelationen der Maße mit dem menschlichen Handeln, wobei diese Wahrnehmung als Optimum angesehen wird. Diese Annahme erfordert weitere Untersuchungen, da automatische Methoden auch dafür eingesetzt werden können, Muster zu identifizieren, die vom menschlichen Auge schwer erfasst werden können, jedoch für die Datenanalyse von großer Bedeutung sein können.

Da dieses Thema in den letzten Jahren großes Interesse geweckt hat, sind viele verschiedene Qualitätsmaße für eine Reihe weiterer hochdimensionaler Visualisierungen entwickelt

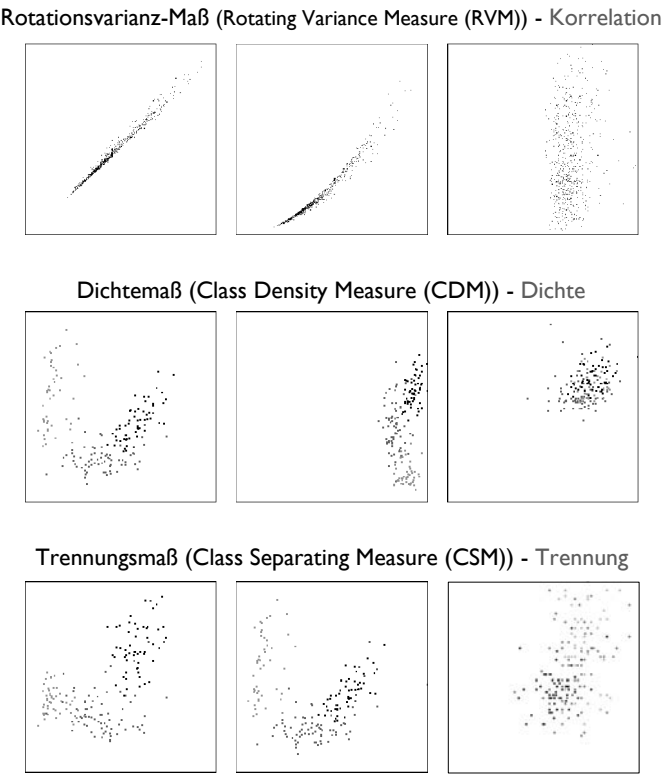

Abbildung 3: Qualitätsmaße für Scatterplots mit verschiedenen Zielfunktionen (blau) - jeweils die zwei Besten (links) und das schlechteste Ergebnis (rechts).

worden. Ich stelle in meiner Dissertation Vorschläge für deren Vernetzung und Weiterentwicklung vor. Hierfür wird eine Übersicht über die verschiedenen Ansätze erstellt, welcher eine Systematisierung zugrunde liegt, die aufgrund einer umfassenden Literaturauswertung zustande kam. Die Systematisierungsfaktoren sind in fünf Kategorien eingeteilt:

1. **Visualisierungsart** – z.B. Scatterplots, Parallele Koordinaten, Pixelvisualisierungen, Histogramme und andere.
2. **Was gemessen wird** – z.B. Korrelation, Gruppierung, Ausreißer, etc.
3. **Wo gemessen wird** – im Daten- oder Bildraum.
4. **Grund für die Messung** – z.B. die richtige Projektion zu finden, die richtige Anordnung der Koordinatenachsen, die richtige Abbildung von Datendimensionen auf visuellen Variablen, usw.
5. **Interaktionsmöglichkeit** – ob man mit den Maßen interagieren kann, z.B. durch die Auswahl von Parametern.

Das Hauptziel dieser Systematisierung ist, deskriptive Modelle zu bilden, um das Verständnis über dieses Gebiet zu erweitern, wie Qualitätsmaße heutzutage eingesetzt werden und wie sie eingesetzt werden könnten, um effektivere Visualisierungstechniken zu entwickeln.

3 Visuelle Suche von Mustern in Unterräumen

Nachdem sich der erste Teil meiner Arbeit der explorativen Aufgabe gewidmet hat, Muster in hochdimensionalen Räumen zu suchen und darzustellen, ist der Fokus im zweiten Teil auf die konkrete Aufgabe des Clustering gesetzt. Ein typisch angewandtes Paradigma für die automatische Mustersuche ist das Clustering, d.h. die Gruppierung von Punkten abhängig von ihrer Ähnlichkeit. Im hochdimensionalen Raum existieren manche Muster nur in verschiedenen Unterräumen des Datenraumes. Subspace Clustering Algorithmen [KKZ09] wurden entwickelt, um Unterräume zu finden, in denen Cluster existieren, die durch traditionelle Clustering Algorithmen nicht gefunden werden würden. Obwohl diese Algorithmen spärlich mit Daten besetzte, hochdimensionale Räume gut explorieren können, ist das Entwickeln von effektiven Visualisierungstechniken, um diese Clusteringresultate zu analysieren, nicht trivial. Zusätzlich zu der Clusterzugehörigkeit von Elementen müssen die relevanten Attributmengen eines Clusters und die Objekt- und Dimensionsüberlappungen von Subspaceclustern dargestellt werden (siehe Abbildung 4).

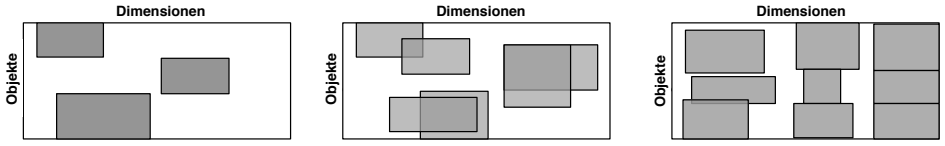

Abbildung 4: Mögliche Relationen von Clustern in hochdimensionalen Räumen, schematisch dargestellt. Jedes große Rechteck stellt eine Datentabelle dar und die blauen Rechtecke jeweils die Cluster in diesen Daten.

Auch wenn eine Reihe von Techniken für die Visualisierung von traditionellen Clustering Resultaten existiert (z.B. Scatterplots, Heatmaps, Dendrogramme, hierarchische Parallele Koordinaten), gibt es nur wenige Ansätze, um das Resultat von Subspace Clustering Algorithmen zu visualisieren. Außerdem wurden bisher erstaunlich wenige Ansätze vorgestellt, die eine visuelle Analyse der Subspace Clustering Ergebnisse unterstützen können, obwohl im Bereich der Subspace Clustering Algorithmen viel Forschung betrieben wurde. In meiner Dissertation stelle ich eine mögliche Visualisierungsumgebung für dieses Problem vor. Angemessene Visualisierungstechniken, die von speziellen Methoden zur Extraktion von Informationen unterstützt werden, würden nicht nur die Nachverfolgung der Clustering Ergebnisse ermöglichen, sondern auch Fachleuten dabei helfen, den Subspace Clustering Prozess so zu steuern, dass relevante Muster zum Vorschein kommen. Dieses Ziel vor Augen stelle ich in Abschnitt 3.1 ein Konzept vor, das Subspace Clustering Algorithmen mit interaktiven skalierbaren Visualisierungen kombiniert. Meine Ansätze widmen sich deshalb der Aufgabe der Visualisierung zum Vergleich von alternativen Clustergruppen, die durch Nutzerfeedback gesteuert werden.

3.1 Visuelle Analyse zum Auffinden von Alternativen Clustern

In den folgenden Abschnitten möchte ich einen interaktiven Ansatz für die Subspace-Analyse (Unterraum-Analyse) vorstellen. Ziel dieses Ansatzes ist es, verschiedene Aspekte und Muster in unterschiedlichen Subspaces (Unterräumen) hochdimensionaler Daten zu finden und deren Relationen zu verstehen.

Die Tatsache, dass interessante Muster in hochdimensionalen Räumen nur in Subspaces zu finden sind, erschwert die Suche nach diesen erheblich, da die Anzahl der Subspaces exponentiell zu der Anzahl der Dimensionen des Datenraumes ansteigt. Eine Suche in all diesen Subspaces ist somit zeitintensiv, daher praktisch nicht möglich. Ich schlage deshalb eine analytische Pipeline vor, die aus automatischen und visuellen Schritten aufgebaut ist und es ermöglicht, einzelne Sichten der Daten auszuwählen und zu analysieren.

Abbildung 5 verdeutlicht die Schritte dieser Pipeline. Die hochdimensionalen Daten werden mit einem Subspacesearch-Algorithmus durchlaufen, der vielversprechende Unterräume zur Mustersuche selektiert. Wegen der Redundanz der Ergebnisse werden Gruppierungsfunktionen basierend auf der Ähnlichkeit der Datentopologie angewendet, um ähnliche Muster zu gruppieren. Der Benutzer kann dann interaktiv steuern, wie viele Gruppen zur weiteren Analyse dargestellt werden sollen. Als repräsentativer Subspace einer Gruppe wird derjenige gewählt, der vom Subspacesearch-Algorithmus den höchsten Interssantheitswert hat. Die darin identifizierten Muster können mit Hilfe von interaktiven Werkzeugen markiert und verglichen werden. So können sowohl übereinstimmende als auch komplementäre, sogenannte alternative Muster identifiziert werden.

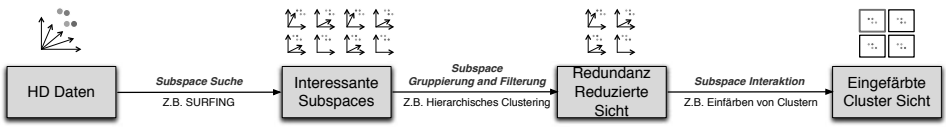

Abbildung 5: Analysepipeline zur visuellen Subspacesuche zur Identifikation von Alternativen Clustern.

Zur Veranschaulichung möchte ich auf die visuellen Schritte im Einzelnen eingehen und das Ergebnis in jedem Schritt der Pipeline erläutern. Dafür wird der zwölf-dimensionale Datensatz aus [FBT+10] verwendet. Im ersten Schritt werden mit Hilfe von einem Subspace-search-Algorithmus die Subspaces identifiziert, die potentiell interessante Muster aufweisen können. An dieser Stelle wurde SURFING [BPK+04] ausgesucht, da dies ein praktisch parameterloser State-of-the-Art Algorithmus ist, der die Subspaces als interessant bewertet, deren k-NN Distanzen nicht uniform verteilt sind [BPK+04]. Dies führt dazu, dass die Punkte im Raum nicht gleich verteilt sind, sondern sich unterschiedlich im Raum gruppieren. Jeder Subspace wird durch ein Scatterplot dargestellt, der die projizierten Daten zeigt (z.B. mit Hilfe von MDS) und einen Dimensionsglyph, der in den grauen Zellen die vorhandenen und in den weißen die nicht vorhandenen Dimensionen darstellt. Das Ergebnis des ersten Schrittes ist in Abbildung 6 zu sehen.

Ein erster Blick auf diese Abbildung lässt Redundanzen im Bezug auf die vorhandenen

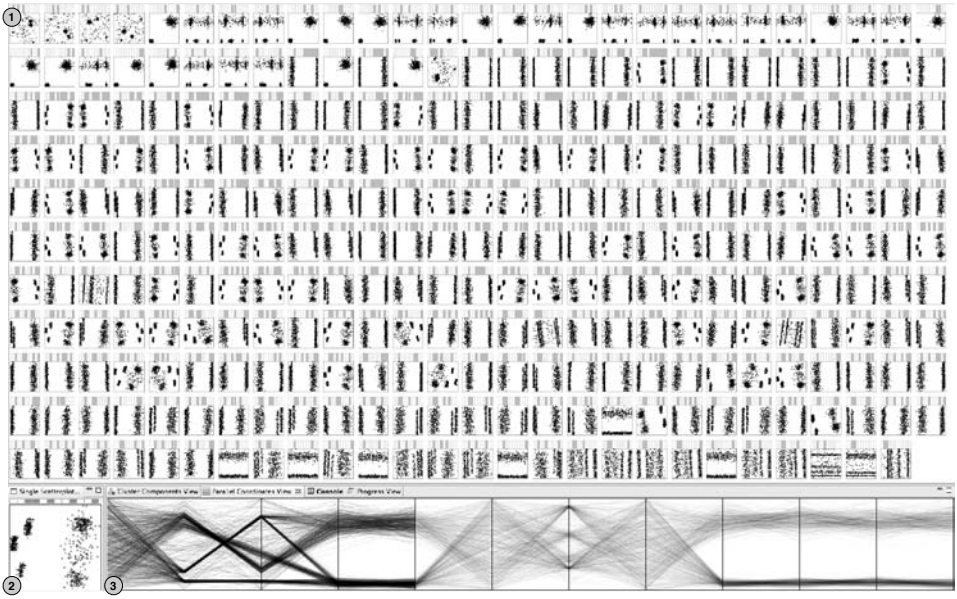

Abbildung 6: (1) Sicht der von SURFING als interessant gewerteten 296 Subspaces des 12D synthetischen Datensatzes aus [FBT+10], sortiert nach dem Interessantheitswert. Der selektierte Subspace in dieser Sicht wird in (2) in einer Einzelsicht dargestellt, die eine interaktive Selektion der Datenpunkte ermöglicht. In (3) ist eine Parallele Koordinaten Ansicht dargestellt, mit den Dimensionen dieses Subspace als hervorgehobene erste Achsen und die übrigen Datendimensionen am Ende zum Vergleich aufgereiht.

Muster erkennen. Deshalb werden Mechanismen benötigt, diese zu reduzieren und die vorhandenen Unterräume zu gruppieren. Ich schlage deshalb im zweiten Schritt Gruppierungsverfahren vor, die die Unterräume entweder basierend auf der Objektähnlichkeit oder basierend auf der Dimensionsüberlappung gruppieren. Details über diese Maße können meiner Dissertation [Tat13] entnommen werden.

Einen Unterraum alleine zu betrachten, ist in vielen Fällen nicht sehr aussagekräftig in Bezug auf die Beschaffenheit der Daten. Unterschiedliche Subspaces können übereinstimmende, komplementäre oder widersprechende Relationen zwischen Datenpunkten aufweisen. Über die Gruppierungsfunktionen können Unterräume verglichen werden und die darin enthaltenen Muster besser interpretiert werden. So identifiziere ich vier verschiedene Arten von Relationen (siehe Abbildung 7), in denen Unterräume zueinander stehen. So sind sie **redundant**, wenn sie sowohl ähnliche Dimensionen als auch topologische Merkmale aufweisen. Sie sind **übereinstimmend**, wenn sie die gleiche Topologie, jedoch unterschiedliche Unterräume aufweisen. Der Fall ähnlicher Dimensionen mit unterschiedlichen Topologien heißt **dominante Dimension**, da die unterschiedlichen Dimensionen das Muster bestimmen. Im letzteren Fall unterschiedlicher Topologien und Dimensionen weisen die Subspaces **komplementäre** Muster auf.

Nach der Gruppierung können im letzten Schritt der Analyse durch Interaktion die Daten

		Enthaltene Dimensionen	
		ähnlich	nicht ähnlich
Topologie der Daten	ähnlich	redundant	übereinstimmend
	nicht ähnlich	dominante Dimensionen	komplementär

Abbildung 7: Relationen von Unterräumen abhängig von ihrer Dimensionsüberlappung und Datentopologie, die durch die Gruppierungs- und Filterfunktionen entdeckt werden können.

in Bezug auf diese Muster untersucht werden. Die Ansichten in Abbildung 8 helfen zur Identifizierung der Relationen. Im vorherigen Beispiel sind nach der Gruppierung in Abbildung 8(1) sechs Gruppen zu sehen, jeweils mit einer Farbe umrandet. In Abbildung 8(5) sind die gruppenzugehörigen Unterräume der orangenen, grünen und lilanen Gruppe zu sehen. Rechts daneben sind die Subspaces abhängig von ihrer Dimensionsähnlichkeit positioniert. Mit Hilfe von Interaktion mit den verlinkten Ansichten und den unterschiedlichen Perspektiven auf die Daten können die bislang besprochenen Relationen identifiziert werden. Ein Beispiel dieser Relationen ist auch anhand eines zweiten Datensatzes einer Ernährungsdatenbank in Abbildung 9 zu sehen.

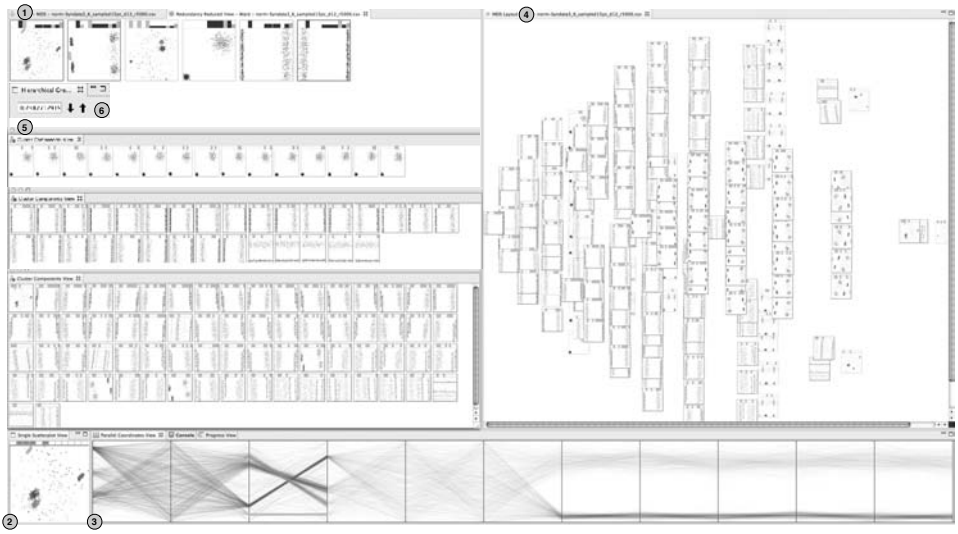

Abbildung 8: (1) Sechs Gruppen nach der Gruppierung basierend auf die Ähnlichkeit der Topologie der Daten. Die erste Gruppe in dieser Sicht wird in (2) in einer Einzelsicht dargestellt, die eine interaktive Selektion der Datenpunkte ermöglicht. In (3) ist eine Parallele Koordinaten Ansicht dargestellt, mit den Dimensionen dieses Subspace als hervorgehobenen ersten Achsen und den übrigen Datendimensionen am Ende zum Vergleich aufgereiht. In (4) werden die Subspaces mit Hilfe von MDS basierend auf die Ähnlichkeit ihrer Dimensionen im 2D Raum positioniert. In (5) sind die Komponenten von drei Gruppen aus (1) zu sehen. (6) stellt die Navigationsknöpfe für das Durchlaufen der Hierarchie dar.

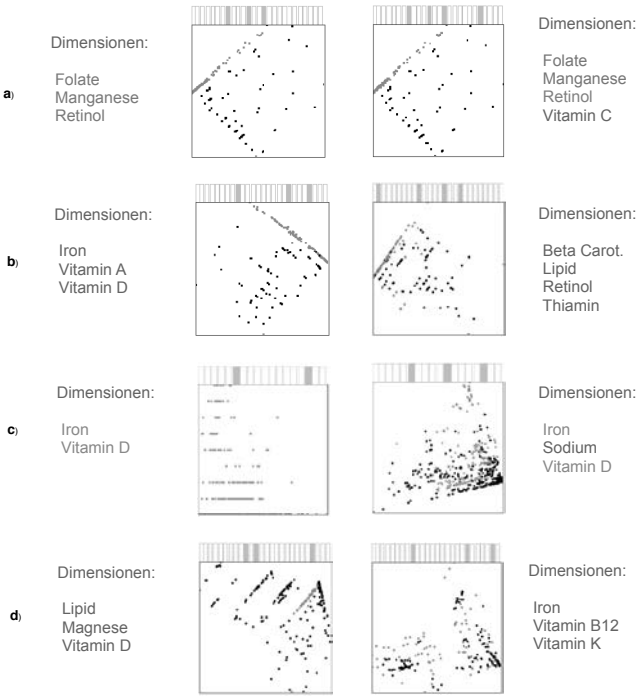

Abbildung 9: Relationen von Unterräumen abhängig von ihrer Dimensionsüberlappung und Datentopologie, die durch die Gruppierungs- und Filterfunktionen entdeckt werden können.

Der vorgestellte Ansatz ist für die Erforschung der Relationen in Subspaces ein möglicher Ausgangspunkt. Erweiterungen dieses Ansatzes können in Betracht gezogen werden. Im Bereich der Datenanalyse wären dies zum einen die Skalierbarkeit der Algorithmen mit anwachsenden Datenmengen sowie unterschiedliche Interessantheitsmaße für das Filtern der Unterräume im ersten Schritt, zum anderen die Sensitivität der verschiedenen Komponenten im Umgang mit Rauschen. Auf der Seite der Visualisierung werden Verbesserungen bei der Darstellung der Unterräume benötigt, um erstens deren visuelle Skalierbarkeit zu gewährleisten, um zweitens die Projektionsstabilität zu beeinflussen und um drittens möglicherweise eine neue Darstellung, die nicht auf Scatterplots beruht, vorzuschlagen. Ebenso könnten automatische Clusteringalgorithmen in den verschiedenen Unterräumen eingesetzt werden, um den manuellen Schritt zu ersetzen und so gleichzeitig eine Vielfalt von möglichen Gruppierungen zu vergleichen.

4 Schlussfolgerung

Die Mustersuche in hochdimensionalen Räumen ist heute ebenso komplex wie von großer Wichtigkeit. Die Verwendung von Visualisierungen, um numerische Daten darzustellen,

von automatischen Methoden, um interessante Elemente zu berechnen, und von Interaktion, um durch die Informationsräume zu navigieren, kann dazu führen, wichtige Muster zu finden. Durch die Komplexität der Daten ist ein Bereich allein nicht ausreichend, um diese zu analysieren, aber in der Kombination können diese sehr nützlich sein.

Literatur

[BPK+04] Christian Baumgartner, Claudia Plant, Karin Kailing, Hans-Peter Kriegel und Peer Kröger. Subspace Selection for Clustering High-Dimensional Data. In *Proceedings of the Fourth IEEE Conference on Data Mining (ICDM '04)*, Seiten 11–18. IEEE CS Press, 2004.

[CLN86] Daniel B. Carr, Richard J. Littlefield und Wesley L. Nichloson. Scatterplot Matrix Techniques for Large N. In *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics*, Seiten 297–306. Elsevier North-Holland, Inc., 1986.

[FBT+10] Bilkis J. Ferdosi, Hugo Buddelmeijer, Scott Trager, Michael H. F. Wilkinson und Jos B. T. M. Roerdink. Finding and Visualizing Relevant Subspaces for Clustering High-Dimensional Astronomical Data Using Connected Morphological Operators. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '11)*, Seiten 35–42. IEEE CS Press, 2010.

[Ins85] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

[KKZ09] Hans-Peter Kriegel, Peer Kröger und Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD '09)*, 3(1):1–58, 2009.

[Tat13] Andrada Tatu. *Visual Analytics of Patterns in High-Dimensional Data*. Dissertation, University of Konstanz, Juli 2013.

Andrada Tatu wurde am 23. Dezember 1984 in Sibiu (Rumänien) geboren. Nachdem sie dort die deutsche Schule besucht hatte, studierte sie in Konstanz von Okt. 2004 bis Sept. 2007 im Studiengang Information Engineering und erlangte den Bachelor of Science, danach weitere zwei Jahre, um im Sept. 2009 den Master of Science an der Universität Konstanz zu erlangen. In dieser Zeit war sie ab dem dritten Bachelorsemester Stipendiatin der Studienstiftung des deutschen Volkes und Hilfswissenschaftlerin (HiWi) am Lehrstuhl für Datenanalyse und Visualisierung von Prof. Dr. Daniel A. Keim. Durch den engen Kontakt zur Forschungsgruppe begann sie im Okt. 2009 eine Promotion und schloss diese an diesem Lehrstuhl im Juli 2013 ab. Nach einer halbjährigen Tätigkeit als Postdoktorandin entschied sie sich für die Wirtschaft und arbeitet seit Januar 2014 bei PricewaterhouseCoopers AG als Senior Consultant im Bereich Forensic Services / Datenanalyse.