Innovationen für die distanzbasierte Datenvisualisierung[1]

Andrej Gisbrecht[2]

Abstract: Die Arbeit adressiert das im Zuge stetig wachsender digitaler Datenmengen hoch relevante Gebiet der Datenvisualisierung durch Dimensionsreduktion. Im Zentrum stehen Problemstellungen für komplexe Situationen, wie etwa heterogene nichtvektrorielle Datensätze, rauschbehaftete oder falsch gestellte Problemstellungen, sowie extrem große Datenmengen. Im Rahmen der Arbeit wurden für diese Probleme jeweils effiziente Lösungen entwickelt und vergleichend evaluiert. Zudem gelingt es in der Arbeit erstmalig, mathematisch exakte Garantien für den wichtigen Fall nichtmetrischer Daten bereitzustellen.

1 Einführung

Der technologische Fortschritt in Bereichen der Messtechnik, Speichermedien und der Netzwerkinfrastruktur hat dazu geführt, dass die Menge an Daten, welche gesammelt und elektronisch gespeichert werden können, dramatisch ansteigt [HL11]. Dies hat wichtige Auswirkungen in verschiedenen Anwendungsbereichen, wie z.B. in der DNA-Sequenzierung und der Entschlüsselung des Genoms, im Internet und dort entstehenden sozialen Netzwerken, bei der Wettervorhersage durch die gemessenen Umweltbedingungen, bei der Digitalisierung von Büchern, Musik und Videoaufnahmen, als auch in vielen anderen Bereichen. Dieses gewaltiges Wachstum an digitalen Informationen und die damit verbundenen hohen potentiellen sozialen und ökonomischen Auswirkungen, haben dazu geführt, dass der Begriff 'Big Data' als eines der wichtigsten Probleme des Jahrhunderts benannt wurde [Ho13]. Die zunehmende Digitalisierung der Gesellschaft bringt sowohl Möglichkeiten als auch Herausforderungen mit sich: einerseits, verspricht die Analyse der gesammelten Daten neue Einsichten in bestehende Vorgänge und die Eröffnung neuer Forschungszweige; andererseits, sind die kognitiven Fähigkeiten eines Menschen begrenzt, wie z.B. bei der Verfolgung von mehreren unterschiedlichen Objekten gleichzeitig, so dass sogar Experten mit unverarbeiteten Informationen, in Form von z.B. einer Tabelle, schnell überfordert werden. Eine der wichtigsten damit zusammenhängenden Fragestellungen, ist deswegen, wie man solche Mengen an digitalen Daten für Menschen begreifbar macht.

Eine Möglichkeit auf die Daten direkt zuzugreifen, dabei aber das Problem der großen und hochdimensionalen Datensätze zu umgehen, ist es die Daten zu visualisieren. Dabei kann man die beeindruckenden kognitiven Fähigkeiten eines Menschen, visuelle Informationen zu verarbeiten, ausnutzen: so können die Menschen strukturelle Formen in visuellen Darstellungen wie Gruppen oder Ausreißer sofort erkennen. Deswegen bietet die Visualisierung eine hervorragende Methodik um eine interaktive Analyse von großen Mengen

[1] Englischer Titel der Dissertation: "Advances in dissimilarity-based data visualisation".

[2] Cognitive Interaction Technology - Exzellenzcluster, Universität Bielefeld, AG Theoretische Informatik, Inspiration 1, 33619 Bielefeld, Deutschland, agisbrecht@techfak.uni-bielefeld.de

digitaler Information zu ermöglichen. Es existieren unterschiedliche Ansätze um diese Aufgabe zu bewerkstelligen, z.B. durch die Benutzung von speziellen Formen wie Diagrammen, Kreisgrafiken, Graphen, parallelen Koordinaten oder ähnlichen. Solche Technologien können kombiniert werden mit passenden interaktiven Strategien, untersucht z.B. im Bereich der visuellen Analytik. Eine andere Möglichkeit die Daten zu visualisieren basiert auf der Aufgabe sehr hochdimensionale Daten passend zu verarbeiten, wobei sich hier die Dimensionalität im zweistelligen Bereich bewegen kann, bis hin zu einigen Millionen Dimensionen. Auf diesem Gebiet ist das Feld der nichtlinearen Dimensionsreduktion entstanden, welches einen sehr wichtigen Bereich des maschinellen Lernens darstellt [LV07].

Der Fokus der hier zusammengefassten Dissertation liegt auf den Methoden der Dimensionsreduktion. Das heißt die Daten sind gegeben als Punkte im hochdimensionalen Vektorraum und das Ziel ist es eine niedrigdimensionale Projektion dieser Daten zu konstruieren, so dass so viel Struktur wie möglich erhalten bleibt. Dadurch dass die Daten nach 2D oder 3D abgebildet werden können, werden diese für Menschen direkt visuell zugreifbar. Da es nicht a priori klar ist, was genau der Term 'Struktur' bedeutet, sind viele Formalisierungen entstanden, welche unterschiedliche Ziele verfolgen, angefangen bei der Distanzerhaltung, über Mannigfaltigkeitsextraktion, bis hin zur der Nachbarschaftserhaltung. Des Weiteren unterscheiden sich die Methoden im Bezug auf die Datenrepräsentation, so sind nicht alle Daten explizit als Vektoren gegeben, stattdessen können sie in Form von paarweisen Relationen wie Ähnlichkeiten und Unähnlichkeiten angegeben werden. In diesem Fall besteht eine zusätzliche Herausforderung darin, dass solche relationale Daten quadratischen Speicherplatz benötigen, da die Anzahl an paarweisen Relationen quadratisch mit der Anzahl der Objekte steigt.

Die steigende Komplexität der Daten bringt nicht nur Herausforderungen für die Benutzer mit sich welche diese inspizieren wollen, sondern auch für die dafür verwendeten Algorithmen: unterschiedliche Anwendungsgebiete führen zu spezialisierten Problemstellungen, welche ausgeklügelte Techniken zur Lösung erfordern. Deswegen stellt sich die Frage, wie die Methoden der Dimensionsreduktion umgewandelt werden können damit sie die Anforderungen der jeweiligen Aufgabenstellung erfüllen können. Große Datensätze können nicht mit Methoden analysiert werden, welche mehr als konstanten Speicherplatz und mehr als lineare Laufzeit benötigen. Deswegen müssen solche Verfahren angepasst werden, um diese Einschränkungen zu erfüllen. Des Weiteren entstehen immer mehr komplexe Daten auf den Gebieten der Bioinformatik und der Textverarbeitung. Diese Daten können mit Vektoren häufig nicht zuverlässig repräsentiert werden, was dazu führt dass sich die Algorithmen nicht auf vektorielle Darstellungen beschränken dürfen, sondern auch mit den problemspezifischen ähnlichkeitsbasierten Daten umgehen können müssen. Außerdem, sind in großen Datensätzen häufig Rauschen oder irrelevante Informationen enthalten, weswegen Methoden benötigt werden, welche sich auf relevante Informationen fokussieren um eine zuverlässige Datenverarbeitung zu garantieren.

In der hier zusammengefassten Dissertation wurden Methoden erarbeitet, welche mit solchen Problemstellungen umgehen können. Die wissenschaftlichen Beiträge der Dissertation befassen sich mit den folgenden Problemstellungen:

- Wie integriert man Vorwissen in den Prozess der Datenvisualisierung? Es werden zwei Ansätze vorgestellt: Das Lernen der Metrik durch das Relevanzlernen und Metrikadaption basierend auf der Fisher-Information.
- Wie erweitert man vektorielle Methoden der Dimensionsreduktion für nichtvektorielle Daten, welche nur in Form von paarweisen Unähnlickeiten gegeben sind? Es wird eine Erweiterung der generativen topographischen Abbildung (generative topographic mapping - GTM) vorgestellt.
- Wie erzeugt man bei nicht-parametrischen Methoden eine explizite Abbildungsfunktion für neue Daten? Es wird eine generelle Technik vorgestellt, welche für jede Visualisierungsmethode eine explizite Abbildung berechnet. Dies wird anhand von t-Distributed Stochastic Neighbor Embedding (t-SNE) demonstriert.
- Wie verbessert man die Recheneffizienz der nichtlinearen Methoden? Es wird die Nyström-Approximation in unähnlichkeitsbasierte Algorithmen integriert um lineare Laufzeitkomplexität zu erreichen. Dazu wird ein formaler Beweis vorgestellt, welcher die Nyström-Approximation auch für indefinite symmetrische Matrizen generalisiert. Darauf basierend wird ein Verfahren vorgestellt, um in linearer Zeit die Eigenwertzerlegung zu berechnen und die Eigenwertkorrektur durchzuführen.

Die behandelten Problemstellungen sind schematisch in der Tabelle 1 dargestellt.

Problem	Relationale	Neue	Effizienz		Relevanz
Methode	Daten	Daten	vektoriell	relational	Learnen
GTM	?	$\surd$	$\surd$	?	?
t-SNE	$\surd$	?	?		?

Tab. 1: In der Dissertation behandelten Fragestellungen.

2 Relevanzlernen

Für die Demonstration des Relevanzlernens wird GTM verwendet [BSW98]. Dies ist eine probabilistische alternative zu der selbstorganisierenden Karte (SOM). GTM erzeugt ein generatives stochastisches Modell der Daten $\mathbf{x} \in \mathbb{R}^D$ indem eine Mixtur der Gauß-Verteilungen, induziert durch latente Punkte $\mathbf{w}$ auf einem regulären Gitter, die Datenverteilung erzeugt.

Das Prinzip des Relevanzlernens wurde eingeführt in [HV02] als eine besonders einfache und effiziente Methode um prototypenbasierte Klassifikatoren an besondere Aufgabenstellungen anzupassen. Es wird die Relevanz der einzelnen Dimensionen bewertet indem die Euklidische Metrik durch ihre gewichtete Form substituiert wird

$$d_\lambda(\mathbf{x},\mathbf{t}) = \sum_{d=1}^{D} \lambda_d^2 (x_d - t_d)^2 \,, \tag{1}$$

wobei die Metrikparameter λ, parallel zum Lernen der Prototypenpositionen, basierend auf der Klassifikationsaufgabe trainiert werden. Dieses Prinzip kann generalisiert werden

[SBH09], indem man die Kombinationen zwischen je zwei Dimensionen bewertet. In diesem Fall wird eine quadratische Matrix Ω als Parameter verwendet und die generalisierte Metrik hat die Form

$$d_\Omega(\mathbf{x},\mathbf{t}) = (\mathbf{x}-\mathbf{t})^T \Omega^T \Omega (\mathbf{x}-\mathbf{t}). \tag{2}$$

In der Dissertation wurde GTM mit dem Prinzip des Relevanzlernens erweitert.

Die Metrikparameter werden bezüglich einer Kostenfunktion gelernt, wobei es verschiedene Möglichkeiten gibt. Es wurden Varianten basierend auf generalized relevance und robust soft learning vector quantization untersucht. So ist z.B. bei robust soft das Ziel das statistische Modell, welches die Daten beschreibt, zu optimieren [SO03]. Es wird die Wahrscheinlichkeit maximiert, dass ein Punkt vom Prototypen der korrekten Klasse generiert wurde im Bezug auf die Gesamtwahrscheinlichkeit:

$$E(\Theta) = \sum_n E_n(\Theta) = \sum_n \log\left(\frac{\sum_{k|c(\mathbf{t}_k)=l_n} p(\mathbf{w}_k)p(\mathbf{x}_n|\mathbf{w}_k,\mathbf{W},\beta)}{p(\mathbf{x}_n|\mathbf{W},\beta)}\right), \tag{3}$$

wobei l_n das Label des n-ten Datenpunkts, $p(\mathbf{x}_n|\mathbf{w}_k,\mathbf{W},\beta)$ die Wahrscheinlichkeit, dass k-tes Prototyp n-tes Punkt erzeugt, $p(\mathbf{x}_n|\mathbf{W},\beta)$ die Wahrscheinlichkeit für n-tes Punkt, $c(\mathbf{t}_k)$ und $p(\mathbf{w}_k)$ das Label bzw. die Wahrscheinlichkeit des k-ten Prototyps sind.

Der Vorteil dieses Ansatzes wird bei der Visualisierung des *Phoneme* Datensatzes in Abbildung 1 klar erkennbar. Dieser Datensatz beschreibt 13 unterschiedliche Phoneme mit 3656 Punkten mit jeweils 20 Dimensionen. Wie man sieht, sind beim klassischen GTM jedem Prototypen die Punkte aus unterschiedlichen Klassen zugeordnet, während bei Relevanz GTM die Metrik so gelernt wurde, dass nur Punkte aus der selben Klasse in einem rezeptiven Feld sind.

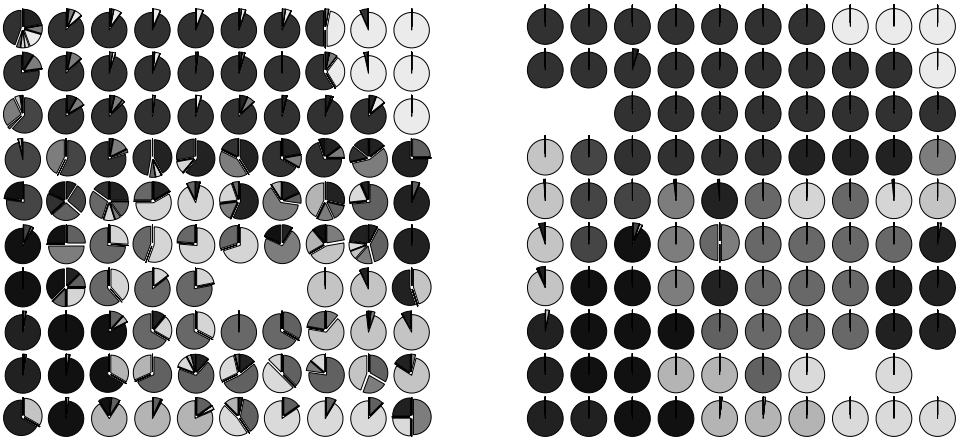

Abb. 1: Visualisierung des *Phoneme* Datensatzes mit GTM. Links ist die klassische Variante, rechts wurde Relevanzlernen verwendet. Die Kreisdiagramme geben an wie viele Punkte der jeweiligen Klasse im rezeptiven Feld des Prototypen sind. Wie man sehen kann, erreicht überwachtes Lernen eine bessere Separation der Klassen.

3 Relationale Daten

Die Komplexität der Daten steigt und immer häufiger können diese Daten nicht hinreichend gut in vektorieller Form dargestellt werden. Stattdessen werden die Daten in paarweisen (Un)Ähnlichkeiten repräsentiert. Das Problem dabei ist, dass klassische Verfahren mit solchen Daten nicht umgehen können. Deswegen wurde in der Dissertation die GTM diesbezüglich erweitert. Als Grundlage dient ein Trick [HB94], welcher es erlaubt die Prototypen implizit darzustellen auch wenn man die explizite vektorielle Form nicht kennt. Die Prototypen werden als eine lineare Kombination der Punkte dargestellt:

$$\mathbf{t}_k = \sum_{n=1}^{N} \alpha_{kn} \mathbf{x}_n \text{ wobei } \sum_{n=1}^{N} \alpha_{kn} = 1. \tag{4}$$

In weiteren Rechnungen können die Prototypen nun in Form von Koeffizienten α_n repräsentiert werden, so kann man die Distanz zwischen einem Prototypen und einem Datenpunkt berechnen als:

$$d(\mathbf{x}_n, \mathbf{t}_k) = [\mathbf{D}\alpha_k]_n - \frac{1}{2} \cdot \alpha_k^T \mathbf{D}\alpha_k, \tag{5}$$

wobei $\mathbf{D}$ die quadratische Matrix ist, welche die paarweisen Unähnlichkeiten zwischen allen Punkten angibt. Basierend darauf ist es nun möglich GTM neu zu formulieren so dass es relationale Daten visualisieren kann.

Ein Beispiel dafür ist in Abbildung 2 dargestellt. Hier wurden mit Relational GTM drei Datensätze abgebildet: *cat cortex* welches aus anatomischen Studien des Katzengehirns stammt und 65 kortikale Gebiete aus vier Klassen beschreibt; *proteins* welches 226 globine Proteine aus fünf Klassen, basierend auf evolutionärer Distanz vergleicht; und *aural sonar* welches Unähnlichkeiten zwischen 100 Sonarsignalen beschreibt und aus zwei Klassen, 'Treffer' und 'Rauschen', besteht.

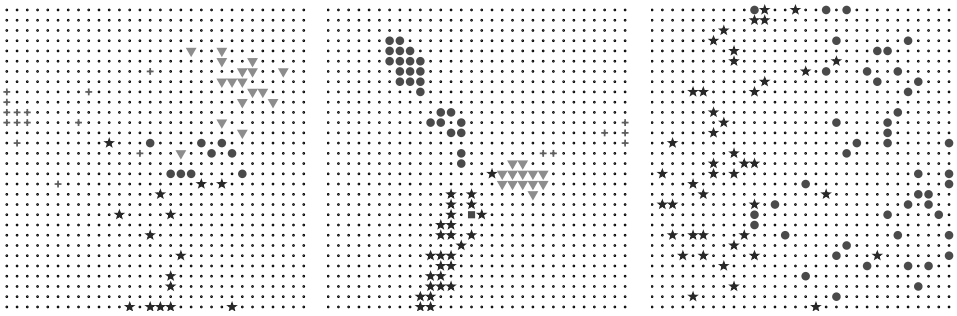

Abb. 2: Visualisierung der Datensätze *cat cortex*, *proteins* und *aural sonar* (von links nach rechts) mit Relational GTM.

4 Nyström-Approximation

Daten, welche durch eine (Un)Ähnlichkeitsmatrix gegeben werden, benötigen quadratische Speicherkapazität. Es ist klar, dass Algorithmen, welche auf diesen Matrizen operie-

ren, mindestens quadratische, häufig aber kubische Rechenkomplexität benötigen. In Hinblick auf 'Big Data' bedeutet das, dass die Datensätze gar nicht vollständig gespeichert werden können und die Algorithmen nicht praktikabel sind. Um dieses Problem zu umgehen wurde auf dem Gebiet der kernelbasierten Methoden die Nyström-Approximation eingeführt [WS01]. Diese Methode wurde jedoch nur für positiv definite Kernelmatrizen definiert, so dass nichtmetrische (Un)Ähnlichkeiten damit nicht approximiert werden konnten. In der Dissertation wurde ein Beweis geliefert, dass die Nyström-Approximation für beliebige symmetrische Matrizen anwendbar ist. So dass viele relationale Methoden nun in linearer Laufzeit durchgeführt werden können.

Relationale Daten, welche z.B. aus einem Sequenzalignment der Gensequenzen, generiert werden, erfüllen nicht notwendigerweise die metrischen Eigenschaften, wie z.B. die Dreiecksungleichung. Das bedeutet, dass solche Daten nicht im Euklidischen Raum eingebettet werden können. Sie können jedoch stets, mithilfe einer symmetrischen Bilinearform, in einem sogenannten pseudo-Euklidischen Raum repräsentiert werden [PD05]:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{p,q} = \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i = \mathbf{x}^\top \mathbf{I}_{p,q} \mathbf{y}, \tag{6}$$

wobei $\mathbf{I}_{p,q}$ eine Diagonalmatrix mit p Einträgen gleich 1 und q Einträgen gleich -1 ist. Somit erzeugen Skalarprodukte in diesem Raum die Ähnlichkeiten und Distanzen die Unähnlichkeiten. Es können auch Transformationen hergeleitet werden, welche diese beide Darstellungen ineinander umformen können. Wenn man von einer Ähnlichkeitsmatrix ausgeht, dann kann man mittels Eigenwertzerlegung die Darstellung aus der Gleichung 6 rekonstruieren. Falls man in dieser Zerlegung die Eigenwerte korrigiert, indem man die absoluten Beträge nimmt, oder negative Werte abschneidet, bekommt man eine positiv definite Matrix welche als Kernel weiter verwendet werden kann. Diese Transformationen sind schematisch in Abbildung 3 dargestellt. Wie man jedoch sehen kann, haben diese Transformationen quadratischen bis kubischen Aufwand, so dass sie in der Praxis zu langsam sind.

Die Nyström-Approximation erlaubt es eine Matrix mit einer anderen Matrix, welche aber einen kleinen Rang hat, zu approximieren [WS01]:

$$\hat{\mathbf{S}} = \mathbf{S}_{N,m} \mathbf{S}_{m,m}^{-1} \mathbf{S}_{m,N}, \tag{7}$$

wobei $\mathbf{S}_{N,m}$ ein linearer Teil der gesamten Matrix ist, welcher aus Relationen zwischen allen N Punkten und m Punkten, die auch als Landmarken bezeichnet werden, besteht. Die Matrix $\mathbf{S}_{m,m}^{-1}$ ist die Pseudoinverse der Matrix der Relationen zwischen den Landmarken. Die Darstellung aus der Gleichung 7 erlaubt es nun die Transformationen in Abbildung 3 durch geschicktes Klammern in linearer Laufzeit durchzuführen.

Mit diesen Transformationen und der approximierten Form der Matrix ist es möglich diverse relationale Algorithmen in linearer Zeit zu berechnen. Außerdem, können beliebige (Un)Ähnlichkeiten zu gültigen Kernen transformiert werden, so dass diese mit der großen Masse an Kernelmethoden verarbeitet werden können.

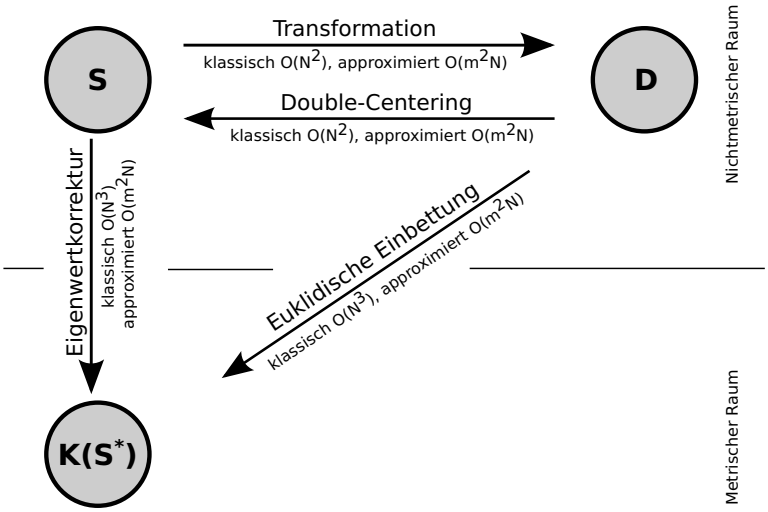

Abb. 3: Transformationen zwischen metrischen und nichtmetrischen (Un)Ähnlichkeiten. Mit der Approximation wird dabei ein linearer Aufwand erreicht.

5 Explizite Abbildung

Obwohl die Nyström-Approximation bei vielen Methoden den Rechenaufwand verbessern kann, ist sie leider nicht immer anwendbar. Unter anderem nicht bei dem t-distributed stochastic neighbour embedding (t-SNE) [vdMH08], welches eine der besten und populärsten Methoden für Dimensionsreduktion ist. Es berechnet für alle Paare von Punkten die Wahrscheinlichkeit, dass diese Punkte benachbart sind, sowohl im hochdimensionalen als auch im niedrigdimensionalen Raum. Die Projektionen der Punkte werden optimiert, indem man die Kullback-Leibler-Divergenz der beiden Wahrscheinlichkeiten minimiert. Da t-SNE einen quadratischen Aufwand hat, ist es leider nicht für sehr große Datensätze geeignet. Zudem generiert t-SNE eine Visualisierung nur für gegebene Daten, falls neue Daten hinzukommen, muss die ganze Abbildung neu trainiert werden. Dies ist ein sehr großer Nachteil bei Anwendungen mit kontinuierlich aufgenommenen Messwerten.

Um dieses Problem zu lösen, wurde ein allgemeines Verfahren vorgestellt, welches eine explizite Abbildungsvorschrift lernt, nicht nur für t-SNE, sondern für jede beliebige Dimensionsreduktion. Die explizite Abbildung kann nicht nur bei wachsenden Datensätzen angewendet werden, sondern auch bei sehr großen Datenmengen: die Projektion wird auf einer kleinen Teilmenge gelernt, und auf verbleibende Daten in linearer Zeit erweitert. Damit kann also für beliebig komplexe Methoden ein linearer Aufwand erreicht werden.

Sind hochdimensionale Daten $\mathbf{x}_j$ gegeben, sowie deren niedrigdimensionale Abbilder $\mathbf{y}_j$, lässt sich eine Abbildung von $\mathbf{x}$ nach $\mathbf{y}$ definieren:

$$\mathbf{x} \mapsto \mathbf{y}(\mathbf{x}) = \sum_j \alpha_j \cdot \frac{k(\mathbf{x}, \mathbf{x}_j)}{\sum_l k(\mathbf{x}, \mathbf{x}_l)}, \tag{8}$$

wobei α_j Parameter sind, welche gelernt werden müssen, und $k(\mathbf{x},\mathbf{x}_j) = \exp(-0.5\|\mathbf{x} - \mathbf{x}_j\|^2/\sigma_j^2)$ ein Gauß-Kern mit Varianz σ_j^2 ist. Die Parameter dieser Abbildung lassen sich mithilfe der Methode der kleinsten Quadrate bestimmen:

$$\mathbf{A} = \mathbf{K}^{-1} \cdot \mathbf{Y}, \tag{9}$$

wobei $\mathbf{A}$ die Matrix der Parameter α_j ist, $\mathbf{Y}$ Matrix der Projektionen $\mathbf{y}_j$ und $\mathbf{K}^{-1}$ die Inverse der Matrix $\mathbf{K}$, gegeben durch $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)/\sum_l k(\mathbf{x}_i,\mathbf{x}_l)$.

Durch diese effiziente Abbildung können Ansätze realisiert werden, welche zwar theoretisch sehr interessant sind, jedoch aufgrund von hoher Komplexität nicht praktikabel. So kann z.B. das Konzept des Relevanzlernens zu einer Riemannschen Metrik generalisiert werden [PKK04]. Dabei wird an jedem Punkt $\mathbf{x}$ des Raumes eine Metrik definiert, welche durch die Fisher-Informations-Matrix gegeben ist:

$$\mathbf{J}(\mathbf{x}) = E_{p(c|\mathbf{x})}\left\{\left(\frac{\partial}{\partial \mathbf{x}}\log p(c|\mathbf{x})\right)\left(\frac{\partial}{\partial \mathbf{x}}\log p(c|\mathbf{x})\right)^T\right\}, \tag{10}$$

wobei E den Erwartungswert bezeichnet und $p(c|\mathbf{x})$ die Wahrscheinlichkeit für einen Punkt der Klasse c an der Stelle $\mathbf{x}$. Eine so definierte Metrik muss über Wegintegrale berechnet werden, kann jedoch stückweise approximiert werden, ist aber trotzdem zu aufwändig um praktikabel zu sein. Mithilfe der expliziten Abbildung kann diese Metrik effizient verwendet werden um eine diskriminative Dimensionsreduktion zu generieren.

Ein Beispiel für eine solche Methode ist in Abbildung 4 dargestellt. Hierbei wurde der *MNIST* Datensatz visualisiert. Dieser besteht aus Bildern von 60000 handgeschriebenen Zahlen von 0 bis 9, jedes Bild mit der Größe von 28×28 Pixel. Es wurde zuerst die Fisher-Distanz zwischen 2000 Punkten berechnet, mit t-SNE visualisiert, und dann mit der darauf gelernten Abbildung die verbleibenden 58000 Punkte zu projizieren.

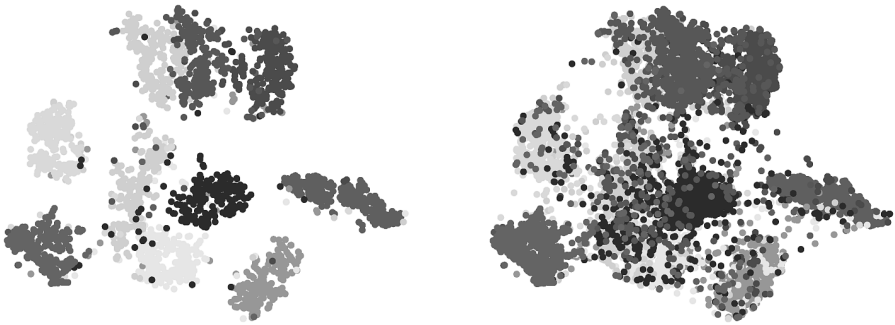

Abb. 4: Visualisierung des *MNIST* Datensatzes unter Verwendung der Fisher-Information. Links sind 2000 Punkte mit t-SNE visualisiert, rechts ist die Erweiterung auf alle 60000 Punkte.

6 Zusammenfassung

Dimensionsreduktion ist ein mächtiges Werkzeug des Data Mining, welches große Datenmengen für menschliche Benutzer zugänglich macht. Als solches muss es den Algorithmen möglich sein, umfangreiche Mengen von beliebig komplexen Daten zu verarbeiten, und daraus leicht interpretierbare Repräsentationen zu erstellen. Diese Dissertation trägt dazu bei, die Lücke zwischen existierenden Ansätzen und effizienten universellen Methoden zu schließen, indem eine Auswahl von verschiedenen Techniken präsentiert wird.

Eine dieser Techniken ist die Integration des Relevanzlernens, welches eine grundlegende Möglichkeit darstellt, um die Zielsetzung der Dimensionsreduktion zu lenken. Dabei wird diese mittels Zusatzinformationen konkretisiert, so dass wichtige Aspekte der Daten bezüglich dieser Zusatzinformationen herausgestellt werden. Dies führt zu Visualisierungen, bei denen die relevanten Anteile der Daten für den Benutzer priorisiert werden. Des Weiteren wurde eine Erweiterung der GTM für die Verwendung von relationalen Daten vorgeschlagen. Da diese Art der Datenrepräsentation eine paarweise Verarbeitung von Objekten bedingt, bringt die Erweiterung den Nachteil einer quadratischen Zeit- und Speicherkomplexität mit sich, was für große Datenmengen nicht realisierbar ist. Es wurde gezeigt, wie dieses Problem mit Hilfe der Nyström-Approximation umgangen werden kann, wodurch der Algorithmus eine lineare Komplexität behält. Die Charakteristiken von relationalen Daten wurden in der Arbeit ausführlich diskutiert, und eine generelle Methode zur Umwandlung nichtmetrischer (Un)Ähnlichkeitsmatrizen in metrische Matrizen wurde vorgestellt. Dieses Prinzip erlaubt die Verarbeitung von beliebigen relationalen Daten mit allen Ähnlichkeitsbasierten und Kernelbasierten Methoden. Als letztes, wurde eine generelle Vorgehensweise vorgestellt, welche es erlaubt für beliebige Dimensionsreduktionen eine explizite Abbildung zu lernen. Dadurch wird es möglich die Visualisierung auf einer kleinen Menge der Punkte zu lernen und die verbleibenden Punkte dann auf die selbe Art und Weise effizient runter zu projizieren.

Literaturverzeichnis

[BSW98] Bishop, Christopher M.; Svensén, Markus; Williams, Christopher K. I.: GTM: the generative topographic mapping. Neural Computation, 10(1):215–234, Januar 1998.

[Gi15] Gisbrecht, Andrej: Advances in dissimilarity-based data visualisation. Dissertation, Bielefeld University, 2015.

[HB94] Hathaway, Richard J.; Bezdek, James C.: Nerf c-means: Non-Euclidean relational fuzzy clustering. Pattern Recognition, 27(3):429–437, 1994.

[HL11] Hilbert, Martin; López, Priscila: The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025):60â65, 2011.

[Ho13] Horvath, Sabine: , Aktueller Begriff - Big Data. Wissenschaftliche Dienste des Deutschen Bundestages, November 2013.

[HV02] Hammer, B.; Villmann, Th.: Generalized Relevance Learning Vector Quantization. Neural Networks, 15(8-9):1059–1068, 2002.

[LV07] Lee, John A.; Verleysen, Michel: Nonlinear dimensionality reduction. Springer, 2007.

[PD05] Pekalska, Elzbieta; Duin, Robert P.W.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, 2005.

[PKK04] Peltonen, Jaakko; Klami, Arto; Kaski, Samuel: Improved Learning of Riemannian Metrics for Exploratory Analysis. Neural Networks, 17:1087–1100, 2004.

[SBH09] Schneider, P.; Biehl, M.; Hammer, B.: Adaptive relevance matrices in learning vector quantization. Neural Computation, 21:3532–3561, 2009.

[SO03] Seo, Sambu; Obermayer, Klaus: Soft learning vector quantization. Neural Computation, 15(7):1589–1604, Juli 2003.

[vdMH08] van der Maaten, L. J. P.; Hinton, G.E.: Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9:2579–2605, 2008.

[WS01] Williams, Christopher; Seeger, Matthias: Using the Nyström Method to Speed Up Kernel Machines. In: Advances in Neural Information Processing Systems 13. MIT Press, S. 682–688, 2001.

Andrej Gisbrecht wurde am 28. März 1983 in Iskitim (Russland) geboren. Er hat in Deutschland das Abitur gemacht und hat an der Technischen Universität Clausthal Informatik studiert. Im August 2009 hat er das Studium mit Auszeichnung abgeschlossen. Sein Doktorandenstudium hat er in Clausthal begonnen und ist im April 2010 zur Universität Bielefeld gewechselt. Dort arbeitete er als wissenschaftlicher Mitarbeiter an der Technischen Fakultät in der Arbeitsgruppe Theoretische Informatik. Im September 2014 hat er seine Doktorarbeit mit Auszeichnung verteidigt. Während des Doktorandenstudiums wurde er zu mehreren Forschungsseminaren in Dagstuhl, zum MPI in Dresden und zu diversen Forschungsgruppen eingeladen. Mit seiner Forschung hat er zu den wissenschaftlichen Bereichen der Visualisierung und Big Data beigetragen.