Erzeugung kalibrierter, metrischer Distanzen mittels multidimensionaler Skalierung

Thomas Böttcher, Ingo Schmitt

Brandenburgische Technische Universität Cottbus – Senftenberg
Walther-Pauer-Str. 2, 03046 Cottbus
tboettcher|schmitt@tu-cottbus.de

Abstract:
Für den Vergleich von Objekten, seien es Texte, Bilder etc, werden in der Regel Ähnlichkeiten bzw. Distanzen bzgl. verschiedener Eigenschaften (z.B. Kanten-, Farb-, Texturfeatures, GPS) genutzt. Werden mehrere Eigenschaften verwendet, führt dies zu einer verbesserten Ausdruckskraft. Problematisch sind hierbei die Eigenschaften der verwendeten Distanzmaße, insbesondere die Dreiecksungleichung. Die Verwendung effizienter Algorithmen, z.B. metrischer Indexsysteme erfordern jedoch diese Eigenschaften. Zusätzlich tritt z.B. bei unterschiedlichen Distanzverteilungen eine Dominanz eines Distanzmaßes auf, die das aggregierte Gesamtergebnis ungewollt verfälscht.
In dieser Arbeit präsentieren wir einen Lösungsansatz, der beide Probleme, mit Hilfe eines Verfahrens der multivariaten Statistik, der multidimensionalen Skalierung (MDS), löst. Wir zeigen wie die Dominanz einer Eigenschaft nachgewiesen und quantifiziert werden kann. Es wird zudem ein erweiterter MDS-Ansatz vorgestellt, der die Vergleichbarkeit verschiedener Distanzmaße gewährleistet. Unser Ansatz erlaubt dabei die Verwendung nicht-metrischer Distanzmaße. Eine Evaluierung auf unterschiedlichen Distanzverteilungen zeigt dabei eine fast vollständige Reduzierung der Dominanz.

1 Einleitung

Im Bereich des Information-Retrievals (IR), Multimedia-Retrievals (MMR), Data-Mining (DM) und vielen anderen Gebieten ist ein Vergleich von Objekten essentiell, z.B. zur Erkennung ähnlicher Objekte oder Duplikate oder zur Klassifizierung der untersuchten Objekte. Der Vergleich von Objekten einer Objektmenge basiert dabei in der Regel auf deren Eigenschaftswerten (Features). Im Bereich des MMR sind Features wie Farben, Kanten oder Texturen häufig genutzte Merkmale. Zur Ausnutzung einer optimalen Ausdruckskraft ist die Verwendung einer geeigneten Kombination verschiedener Features entscheidend.

Der (paarweise) Vergleich von Objekten anhand von Features erfolgt mittels eines Distanz- bzw. Ähnlichkeitsmaßes[1]. Bei mehreren Features lassen sich Distanzen mittels einer Aggregationsfunktion verknüpfen und zu einer Gesamtdistanz zusammenfassen. Problematisch sind hierbei die algebraischen Eigenschaften des Distanzmaßes und die mögliche Dominanz eines Features bzgl. des Aggregationsergebnisses:

1. *Dominanz einer Eigenschaft* - Für einen Ähnlichkeitsvergleich anhand mehrerer Features wird erwartet, dass die Einzelfeature gleichermaßen das Aggregationsergebnis

[1]Beide lassen sich ineinander überführen [Sch06], im Folgenden gehen wir daher von Distanzmaßen aus.

beeinflussen. Häufig gibt es jedoch ein Ungleichgewicht, so dass einzelne Features keinen oder nur geringen Einfluss besitzen.

2. *Nichterfüllung einer Metrikeigenschaft* - Verschiedene Distanzmaße erfüllen unterschiedliche algebraische Eigenschaften und nicht alle Distanzmaße sind für spezielle Probleme gleich geeignet. So erfordern Ansätze zu metrischen Indexverfahren oder Algorithmen im Data-Mining die Erfüllung der Dreiecksungleichung.

Fehlen algebraische Eigenschaften oder gibt es eine zu starke Dominanz, so können die Features und dazugehörigen Distanzmaße nicht mehr sinnvoll innerhalb einer Featurekombination eingesetzt werden.

In bisherigen Arbeiten wurde meist versucht jeweils eines der Probleme zu lösen, eine gemeinsame Betrachtung gibt es bisher nicht [Sko06, WCB06, MA01]. Aus diesem Grund sollen in diesem Paper beide Probleme im Kontext betrachtet werden. Wir werden zunächst ein neues Maß vorstellen, welches die Dominanz eines Features innerhalb der aggregierten Distanz misst. Anschließend soll mit einem angepassten Verfahren der multivariaten Statistik, der multidimensionale Skalierung (MDS), gezeigt werden, wie die genannten Probleme gelöst werden können. Mittels einer Erweiterung des rang-erhaltenden MDS-Verfahrens kann gezeigt werden, dass fehlende Metrikeigenschaften, insbesondere die Dreiecksungleichung, kompensiert werden können. Eine Evaluierung auf unterschiedlichen Distanzverteilungen zeigt dabei eine starke Reduktion der Dominanz.

Die Arbeit ist dabei wie folgt aufgebaut. In Kapitel 2 werden grundlegende Definitionen von Distanzmaßen, Aggregation und Dominanz eines Features erläutert. Kapitel 3 liefert einen Überblick über den Stand der Technik. Kapitel 4 erläutert die Grundlagen der MDS und ihre Grenzen. Kapitel 5 beschreibt die notwendigen Erweiterungen der MDS zur Lösung genannter Probleme. Eine Evaluierung der Ergebnisse erfolgt in Kapitel 6. Kapitel 7 gibt eine Zusammenfassung sowie einen Ausblick über zukünftige Arbeiten.

2 Grundlagen

Das folgende Kapitel definiert die grundlegenden Begriffe und die Notationen, die in dieser Arbeit verwendet werden.

2.1 Grundlegende Definitionen

Eine Ähnlichkeitsberechnung erfolgt in der Regel auf der Grundlage einer Objektmenge $O \subseteq \mathbb{U}$ im Universum $\mathbb{U}$. Ein Distanzmaß zwischen zwei Objekten basierend auf einem Feature p sei als eine Funktion $d : \mathbb{U} \times \mathbb{U} \mapsto \mathbb{R}_{\geq 0}$ definiert. Zur Klassifikation der unterschiedlichen Distanzmaße werden folgende vier Eigenschaften genutzt: Selbstidentität: $\forall o \in O : d(o,o) = 0$, Positivität: $\forall o_r \neq o_s \in O : d(o_r,o_s) > 0$, Symmetrie: $\forall o_r,o_s \in O : d(o_r,o_s) = d(o_s,o_r)$ und Dreiecksungleichung: $\forall o_r,o_s,o_t \in O : d(o_r,o_t) \leq d(o_r,o_s) + d(o_s,o_t)$. Erfüllt eine Distanzfunktion alle vier Eigenschaften, so wird sie als Metrik bezeichnet [Sam06], ohne die Eigenschaft der Dreiecksungleichung wird sie als Semidistanzfunktion bezeichnet.

Für eine Distanzberechnung mit m Features $p = (p_1,\ldots,p_m)$ werden zunächst die partiellen Distanzen $\delta^j_{rs} = d^j(o_r,o_s)$ bestimmt. Anschließend werden die partiellen Distanzwerte δ^j_{rs} mittels einer Aggregationsfunktion $agg : \mathbb{R}^m_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ zu einer Gesamtdistanz aggregiert.

Die Menge aller aggregierten Distanzen für Objektpaare aus O mit $|O| = n$ sei als $\delta^j = \{\delta^j_1, \delta^j_2 \ldots, \delta^j_l\}$ mit $l = \frac{n^2-n}{2}$ bestimmt.

2.2 Dominanz eines Features

Für einen Ähnlichkeitsvergleich von Objekten anhand mehrerer Merkmale sollen die Einzelmerkmale gleichermaßen das Aggregationsergebnis beeinflussen. Offen ist dabei, wann eine Dominanz eines Features auftritt und wie der Grad der Dominanz gemessen werden kann. Betrachtet man in ein festes Intervall (z.B. $[0,1]$) normierte Distanzen, so kann eine unterschiedliche Distanzverteilung einer Stichprobe, insbesondere unterschiedliche Intervallgrößen, zu einer Dominanz eines Features führen [BS14]. Wir definieren den Grad der Dominanz mittels eines Korrelationsmaßes, wobei dieser als Abweichung zum gewünschten Zustand (kalibrierte Distanzverteilung) gemessen wird [BS14, Kub12]:

$$Cal_{err}(\delta^i, \delta^j, \delta^{agg}) = 1 - \frac{4}{\pi}\arctan\left(\frac{Corr(\delta^j, \delta^{agg})}{Corr(\delta^i, \delta^{agg})}\right). \quad (1)$$

Hierbei definiert $Corr(X,Y)$ ein geeignetes Korrelationsmaß, in unserem Fall genügt der Rangkorrelationskoeffizient von Spearman [Spe04]. Es wird immer die Kalibrierung von zwei Verteilungen bezüglich des Aggregationsergebnisses betrachtet. Die Korrelationswerte lassen sich als Punkt in $[0,1]^n$ auffassen. Die Verteilungen sind kalibriert ($Cal_{err} = 0$), wenn sie gleich stark mit dem aggregierten Wert korrelieren, d.h. auf der Winkelhalbierenden liegen. Bei allen Punkten unterhalb dieser Linie dominiert die erste Verteilung das Aggregationsergebnis, bei allen oberhalb die zweite Verteilung.

3 Stand der Technik

Das Problem fehlender Metrikeigenschaften, insbesondere der Dreiecksungleichung, ist bereits aus verschiedenen Gebieten bekannt. Skopal [Sko06] stellt mit seinem Tri-Gen-Algorithmus einen Ansatz bereit, um eine Semi-Metrik in eine Metrik zu transformieren. Hierzu wird eine Teilmenge der Daten und deren Distanzverteilung genutzt. Ist die Teilmenge jedoch zu klein, kann nicht sichergestellt werden, dass der Tri-Gen-Ansatz eine vollständige Metrik liefert. Einen zur MDS sehr ähnlichen Ansatz stellt FastMap [FL95] dar. Dieser Ansatz bildet Daten im k-dimensionalen Raum ab. Durch die Verwendung einer L_p-Norm können auch die Eigenschaften einer Metrik sichergestellt werden. FastMap erreicht dabei eine verbesserte Laufzeit $O(n \log n)$ gegenüber $O(n^2)$ bei der MDS. Die Reihenfolge der Distanzen wird bestmöglich erhalten - eine Garantie gibt es jedoch nicht.

Score-Normalization stellt einen zweiten wichtigen Themenbereich dieser Arbeit dar. Die Evaluierung solcher Ansätze erfolgt in vielen Fällen, vor allem im Bereich des IR, direkt über die Auswertung der Qualität der Suchergebnisse. Dieses Vorgehen liefert aber kaum Anhaltspunkte, warum sich einige Normalisierungsansätze besser für bestimmte Anwendungen eignen als andere [Kub12]. Eine Übersicht über die zahlreichen Ansätze (nicht-)linearer Normalisierungen findet sich in [WCB06, Kub12, BS14]. Für das Problem der Dominanz lässt sich einfach zeigen, dass diese Ansätze keinen direkten Einfluss auf die Distanzverteilung haben. Es werden maximal zwei statistische Merkmale (Minimum, Maximum, Median etc.) genutzt, um eine Normalisierung durchzuführen [MA01]. Besonders

problematisch ist die mangelnde Robustheit gegenüber Ausreißern.

Ansätze, die Distanzverteilung als Grundlage zur Normalisierung heranziehen, versuchen Distanzen aus unterschiedlichen Quellen so abzubilden, dass sie möglichst exakt gleiche Verteilungen besitzen. Die Ansätze von Manmatha [MRF01] und Fernandez [FVC06] analysieren dabei das probabilistische Verhalten von Suchmaschinen unter der Annahme, dass relevante Dokumente eine Normalverteilung und irrelevante eine exponentielle Verteilung besitzen. Diese Ansätze bieten zwar eine optimierte Normierung, erfordern aber gleichzeitig Informationen über die Relevanz von Dokumenten, die häufig nicht vorhanden sind.

4 Multidimensionale Skalierung

Die Multidimensionale Skalierung (MDS) ist ein Verfahren der multivariaten Statistik und dient ursprünglich dem Vergleich einer Vielzahl von Objekten oder Datenpunkten anhand eines Nachbarschaftsmaßes [GR72]. Mittels MDS wird versucht, eine Menge von Objekten O mit Hilfe der paarweisen Distanzen in einem t-dimensionalen Raum anzuordnen [BG05]. Eine solche Anordnung bezeichnen wir als *Konfiguration*. In einigen Fällen wird die MDS auch zur Visualisierung der Daten ($t = 2$) eingesetzt.

In diesem Kapitel soll erläutert werden, welche Anforderungen die MDS erfüllen muss, um die Probleme der Dreiecksungleichung und der Kalibrierung lösen zu können. Anschließend werden das Prinzip und die Grenzen der nicht-metrischen MDS beschrieben.

4.1 Anforderungen

Zur Lösung des Problems fehlender Dreiecksungleichung muss aus einer nicht-metrischen Ausgangsdistanzmatrix D eine metrische Distanzmatrix D' erzeugt werden. Das Verfahren der Multidimensionalen Skalierung arbeitet mit Konfigurations- und Distanzmatrizen. Die durch die MDS erzeugte Konfiguration besteht aus Vektoren im mehrdimensionalen euklidischen Raum. Wir müssen zunächst zeigen, dass die aus der Konfiguration abgeleitete Distanzmatrix D' die Eigenschaften einer Metrik, insbesondere die der Dreiecksungleichung erfüllt. Zusätzlich muss nachgewiesen werden, dass die Rangfolge der Ursprungsdistanzen erhalten bleibt, d.h. $\text{Rang}(D(o_r, o_s)) = \text{Rang}(D'(o_r, o_s))$.

Für die Lösung des Dominanzproblems müssen die erzeugten Distanzmatrizen gleichermaßen das Aggregationsergebnis beeinflussen. Erzeugte Distanzverteilungen von D'_j sollten daher unabhängig von der Eingangsdistanzmatrix D_j möglichst identisch[2] sein.

4.2 Nicht-Metrische Multidimensionale Skalierung

Der nicht-metrische Ansatz der MDS nach Kruskal [Kru64a] arbeitet ausschließlich auf den Rängen der Distanzmatrix. Ausgangspunkt der MDS ist eine symmetrische[3] Distanzmatrix D bestehend aus jeweils paarweisen Distanzen δ_{ij}. Sei $\hat{D}$ die obere Dreiecksmatrix von D. Wir definieren nun ausgehend von $\hat{D}$ die zu erhaltende Rangfolge:

[2] Identisch bezieht sich hierbei auf die Größe des Intervalls in dem die Distanzen verteilt sind, so dass die Korrelation zu den Aggregationsdistanzen gleich ist.

[3] Eine Transformation einer nicht-symmetrischen in eine symmetrische Distanzmatrix ist nicht Bestandteil dieser Arbeit. Ein möglicher Ansatz findet sich z.B. in [Kru64a].

$$\hat{\delta}_1 \leq \hat{\delta}_2 \leq \cdots \leq \hat{\delta}_K, \tag{2}$$

wobei $\hat{\delta}_1$ hierbei der kleinsten Distanz $\delta_{ij} \in \hat{D}$ entspricht.

Die Repräsentation der Objekte im t-dimensionalen Raum erfolgt mittels Konfigurationen. Eine Konfiguration für n Objekte ist nach Kruskal durch $C = (x_1, \ldots, x_n), x_i \in \mathbb{R}^t$ definiert. Für die Erzeugung einer solchen Konfiguration genügen die Ränge der Distanzwerte. Eine initiale Konfiguration C wird typischerweise zufällig erzeugt. Zur Erhaltung der ursprünglichen Rangfolge müssen die aus der Konfiguration abgeleiteten Distanzen d_{ij} die gleiche Rangfolge ergeben. Um dies zu gewährleisten, definieren wir folgende Monotoniebedingung für alle Objektpaare:

$$\delta_{ij} > \delta_{kl} \Rightarrow d_{ij} > d_{kl}. \tag{3}$$

Die Distanzen d_{ij} werden hierbei durch die Verwendung der L_2-Norm bestimmt[4].

Um eine Konfiguration bzgl. (3) bewerten zu können, nutzen wir den von Kruskal eingeführten Stresswert $S = (\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2 / \sum_{i<j} d_{ij}^2)^{\frac{1}{2}}$, wobei d_{ij} die Konfigurationsdistanzen und $\hat{d}_{ij}$ die Disparitäten zwischen den Objekten o_i und o_j [Kru64a] sind. Eine optimale Konfiguration (Erfüllung von (3)) ist bei einem Stresswert $S = 0$ erreicht. Ist die Monotoniebedingung zwischen zwei benachbarten Distanzen gemäß (3) verletzt ($S > 0$), so erfolgt eine Anpassung der Distanzen als Disparitäten mittels Pool-Adjacent Violators (PAV) [RDJH73, ABE+55]. Dabei wird zunächst für die benachbarten Distanzen ein Pool gebildet und beide Distanzen durch den Mittelwert ersetzt. Wurde dadurch eine weitere Verletzung zu der vorherigen Distanz erzeugt, so wird diese ebenfalls in den Pool aufgenommen und der Mittelwert bestimmt. Dies wird solange fortgesetzt, bis keine Monotonieverletzung mehr vorliegt. Allerdings gibt es bei diesem Verfahren häufig Distanzen mit gleichem Wert.

Da in der Regel die initiale Konfiguration nicht genügt, muss diese modifiziert werden. Das Finden einer optimalen Konfiguration ist nicht trivial, da es möglich ist, dass mehrere oder (im Fall $t \ll n$) keine Lösungen existieren können, die (3) erfüllen. Um eine näherungsweise Lösung zu finden, schlägt Kruskal einen iterativen Prozess vor. Erhalten wir nach der Stressberechnung $S < \epsilon$, wobei ϵ ein vordefinierter maximaler Stresswert ist, so ist eine Zielkonfiguration gefunden. Andernfalls erfolgt die Berechnung einer modifizierten Konfiguration $\hat{C}$ aus den Koordinaten x_{it} (Wert von Objekt o_i in Dimension t). Hierfür werden zwei Objekte o_i und o_j genutzt, um die Position von o_i in Bezug auf o_j in der Dimension r durch $x'_{ir} = x_{ir} + \alpha(1 - \frac{\hat{d}_{ij}}{d_{ij}}) \cdot (x_{jr} - x_{ir})$, für $i \neq j, r = (1, \ldots, t)$ neu zu berechnen [Kru64b]. Eine Neupositionierung des Objektes o_i bezüglich aller anderen Objekte innerhalb einer Dimension r kann wie folgt abgeleitet werden:

$$x'_{ir} = x_{ir} + \frac{\alpha}{n-1} \sum_{j=1}^{n} (1 - \frac{\hat{d}_{ij}}{d_{ij}}) \cdot (x_{jr} - x_{ir}),\ r = (1, \ldots, t). \tag{4}$$

Der Parameter α definiert eine Schrittweite, zur Anpassung der Koordinaten von x_{ir}. Ein kleinerer Wert erlaubt eine höhere Genauigkeit der Lösung, jedoch auf Kosten von mehr Iterationen [Kru64a].

[4]Die Verwendung nicht-euklidischer Distanzmaße wird in [Kru64a] diskutiert.

4.3 Grenzen der nicht-metrischen MDS

Für die Erzeugung einer metrischen Distanzmatrix wird die aus der MDS erzeugte Konfiguration C genutzt. Es gibt hierbei zwei Lösungsansätze. Es können die aus den Konfigurationen bestimmten Distanzen d_{ij} direkt genutzt werden. Alternativ kann aus der Konfiguration eine positiv semidefinite Ähnlichkeitsmatrix $A = CC^T$ erzeugt werden. Die mittels des Skalarprodukts erzeugte Ähnlichkeitsmatrix kann anschließend rangerhaltend in eine metrische Distanzmatrix transformiert werden. Problematisch ist hierbei, dass unter Verwendung von Gleichung (4) die Objekte einer Konfiguration beliebig im t-dimensionalen Raum verteilt sein können und somit die maximal mögliche Distanz zwischen zwei Objekten nicht bestimmt werden kann. Hinzu kommt, dass die Verwendung des Pool-Adjacent Violator-Verfahrens nur eine schwache Monotonie zwischen Distanzen und Disparitäten beinhaltet, d.h. $d_{ij} > d_{kl} \Rightarrow \hat{d}_{ij} \geq \hat{d}_{kl}$ sowie $\delta_{ij} > \delta_{kl} \Rightarrow \hat{d}_{ij} \geq \hat{d}_{kl}$. Zur Erzeugung einer linearen Abbildung der Distanzen unabhängig von den Ursprungsdaten benötigen wir jedoch eine strikte Ordnung von Distanzen und Disparitäten: $\delta_{ij} > \delta_{kl} \Rightarrow \hat{d}_{ij} > \hat{d}_{kl}$. Zusätzlich müssen die Distanzen auf ein festes Intervall beschränkt werden. Der bisherige PAV-Algorithmus genügt somit nicht für die Lösung des Dominanzproblems und muss daher erweitert werden.

5 Erweiterung der Multidimensionalen Skalierung

Im vorherigen Kapitel haben wir gezeigt, dass die nicht-metrische MDS in der ursprünglichen Form nicht ausreicht, um unsere definierten Probleme lösen zu können. Im folgenden Kapitel werden mehrere Anpassungen an das ursprüngliche Verfahren erläutert. Anschließend zeigen wir, dass diese Modifikationen genügen, um sowohl das Problem der Dreiecksungleichung als auch das Problem der Dominanz zu lösen.

5.1 Konfigurationsnormalisierung

Da die ursprüngliche Konfigurationsberechnung für das Problem der Dreiecksungleichung nicht geeignet ist, wird das MDS-Verfahren durch eine Konfigurationsnormalisierung erweitert. Wir werden hierzu jeden Punkt x_i der Konfiguration C auf die Oberfläche der Hyperkugel mit einem Radius $r = 1$ und t Dimensionen projizieren. Als initiale Konfiguration wurde bisher eine zufällige Verteilung der Objekte in $\mathbb{R}^t$ angenommen. O.B.d.A. können wir die initiale Konfiguration auch auf den ersten Orthanten beschränken, d.h. alle Koordinaten sind stets positiv. Für eine gegebene Konfiguration C definieren wir die normalisierte Konfiguration als $C' = (x_1', \ldots, x_n'), x_i' \in \mathbb{R}^t$, wobei $x_i' = \frac{x_{it}}{l_i}$ und $l_i = \sqrt{x_{i1}^2 + \cdots + x_{it}^2}$. Betrachten wir die Koordinaten eines Objektes o_i als Vektor v_i, so erhalten wir stets einen Vektor der Länge 1 ($|v_i| = 1$). Zusätzlich wird die maximale Distanz zwischen zwei Objekten beschränkt, d.h. $\forall o_i, o_j : d(o_i, o_j) \leq \sqrt{2}$.

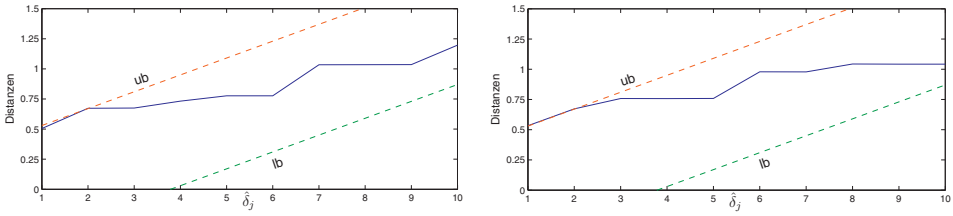
5.2 Erweiterung des PAV-Algorithmus

Um die Vergleichbarkeit erzeugter Distanzmatrizen zu gewährleisten, genügt die Normalisierung der Konfiguration nicht. Hierzu erweitern wir das PAV-Verfahren [RDJH73], um einen linearen Anstieg der Distanzwerte (vgl. Abbildung 1) zu erzeugen. Wir definieren

zunächst einen festen Korridor, in denen die Konfigurationsdistanzen und Disparitäten liegen dürfen. Anschließend erzeugen wir eine totale Ordnung ($\delta_{ij} > \delta_{kl} \Rightarrow \hat{d}_{ij} > \hat{d}_{kl}$) auf den Disparitäten, so dass keine Disparitäten mehrfach auftreten. Um die Konfigurationsdistanzen innerhalb eines festen Korridors zu beschränken, wird ein Bereich (*range*) definiert, der sicherstellt, dass unabhängig von den Eingangsdaten alle Konfigurationsdistanzen auf das gleiche Intervall abgebildet werden. Hierzu wird eine obere Grenze $ub = min(\sqrt{2}, middle + \frac{range}{2})$ sowie eine untere $lb = max(0, middle - \frac{range}{2})$ definiert, wobei $middle = (ic - 0.5) \cdot \frac{\sqrt{2}}{K}$ ist. K definiert die Anzahl der Distanzen und ic ist ein Iterationszähler beim Erzeugen der Pools, um Distanzen gleichmäßig über das Intervall zu verteilen. Sollten Distanzen außerhalb der Grenzen (ub oder lb) liegen, wird ihnen der Grenzwert zugeordnet. Die Größe des Intervalls (*range*) liegt im Bereich $[0, \sqrt{2}]$ und variiert mit der Anzahl der Objekte.

Zur Erzeugung einer totalen Ordnung werden alle Distanzen eines Pools mit einem festen Wert inkrementiert. Sei $p(m)$ der m-te PAV-Pool mit $size(p(m))$ Elementen. Weiterhin sei $dist(p(m))$ der Distanzwert eines Pools m. Der Basiswert für die Inkrementierung für einen Pool sei durch $inc(m) = \frac{dist(p(m+1)) - dist(p(m))}{size(p(m))}$ bestimmt. Mit Hilfe des Basiswertes (inc) wird nun ein linearer Verlauf der Distanzen von einem Pool zum Nächsten ermöglicht. Sei $p(m)(i)$ die i-te Position im Pool, dann berechnet sich die i-te Disparität von Pool m durch $disparity(m,i) = dist(p(m)) + inc(m) \cdot (i-1), \forall i = 1, \ldots, size(p(m))$.

Abbildung 1 zeigt den erzeugten linearen Verlauf der Konfigurationsdistanzen gegenüber den Ausgangsdistanzen. Die Ursprungsdaten entsprechen dem Beispiel aus [BS14] und weisen initial durch zwei unterschiedliche Distanzverteilungen einen Kalibrierungsfehler von $Cal_{err} = 0.43$ auf. Unter Verwendung des erweiterten MDS-Algorithmus reduziert sich bei korrekter Erhaltung der Ursprungsrangfolge der Kalibrierungsfehler auf $Cal'_{err} = 0,01$.



(a) ν_1 (b) ν_2

Abbildung 1: Verteilung der Konfigurationsdistanzen unter Verwendung des erweiterten MDS-Algorithmus ($\alpha = 0.2, range = 0.7, t = 4$)

5.3 Zusammenfassung

Mit der Erweiterung des iterativen MDS-Verfahrens können wir nun eine Konfiguration erzeugen, die einerseits das Problem der Dreiecksungleichung und anderseits das Dominanzproblem löst.

Ausgangspunkt für die Erzeugung einer metrischen Distanzmatrix ist eine normalisierte Konfiguration C. Betrachtet man jedes Objekt als einen normierten Zeilenvektor $v_i =$

$(x_1, \ldots, x_t)$, kann eine normierte, positiv semidefinite Ähnlichkeitsmatrix $A = CC^T$ erzeugt werden, wobei die Ähnlichkeit zwischen zwei Objekten dem Skalarprodukt, also dem Kosinus des eingeschlossenen Winkels entspricht. Das Skalarprodukt von v_i zu sich selbst ist stets 1. Alle anderen Werte liegen im Intervall $[0, 1]$. Mit dem Kosinussatz kann aus dem Ähnlichkeitswert a_{ij} unter automatischer Erfüllung der Dreiecksungleichung eine Distanz $d'(i,j) = \sqrt{2 - 2\,a_{ij}}$ abgeleitet werden.

Erhalten wir durch die MDS eine Konfiguration mit einem Stresswert von 0 bzw. bei einem hinreichend kleinen ϵ mit $S < \epsilon$, so gilt auch (3). Damit wird eine exakte Abbildung der Ursprungsrangfolge durch die Konfiguration C gewährleistet.

Durch Anpassung des PAV-Verfahrens kann mittels MDS eine Konfiguration erzeugt werden, deren abgeleitete Distanzen gleichmäßig über ein definiertes Intervall verteilt sind. Dies ermöglicht es, Distanzmatrizen zu erzeugen, die unabhängig von ihren Ursprungsdaten das Aggregationsergebnis gleichmäßig beeinflussen. Das Beispiel aus Abbildung 1 zeigt exemplarisch, dass die Dominanz nahezu vollständig beseitigt wurde. Eine umfangreichere Evaluierung erfolgt im nächsten Kapitel. Offen für zukünftige Arbeiten bleiben zunächst Lösungsansätze zur automatisierten Findung der Parameter α, $range$ und t.

6 Evaluierung

In diesem Kapitel soll der Grad der Dominanz mit Hilfe des Kalibrierungsfehlers evaluiert werden. Wir werden zunächst die Auswirkungen des erweiterten MDS-Verfahrens zeigen. Anschließend erfolgt ein Vergleich mit zwei häufig eingesetzten Normalisierungsverfahren. Zur Durchführung der Experimente wurden Matlab-Skripte implementiert.

Durch die Erzeugung einer Konfiguration mittels des erweiterten MDS-Algorithmus wird eine nahezu lineare Abbildung von Distanzen ermöglicht (vgl. Abbildung 1). Wir zeigen nun, dass die erzeugten Distanzverteilungen unabhängig von den Ausgangsdaten ähnlich in Bezug auf die Korrelation zum Aggregationsergebnis sind. Zur Evaluierung wurden in 1000 Durchläufen verschiedene Objektmengen ($n = 5, 10, 30, 50, 100$) aus Normal- und Gleichverteilungen normalisiert. Es zeigte sich, dass der Kalibrierungsfehler fast vollständig entfernt wurde und somit das Problem der Dominanz nach der Kalibrierung durch die MDS nicht mehr vorhanden ist. Tabelle 1 zeigt einige Auszüge der Ergebnisse, wobei neben den MDS-Parametern t, α und range auch die Stresswerte S_1 und S_2 nach der MDS sowie die Korrelationswerte ρ_1 und ρ_2 und die Kalibrierungsfehler vor und nach der Normierung dargestellt sind. Durchschnittlich konnte der Kalibrierungsfehler um 94% reduziert werden. Dies wird zudem anhand der Korrelationswerte ρ_i vor und nach der MDS deutlich.

Tabelle 1: Veränderung des Kalibrierungsfehlers nach Anwendung der MDS

Obj.	t	α	**range**	S_1	S_2	ρ_1	ρ_2	Cal_{err}	ρ_1'	ρ_2'	Cal_{err}'
5	4	0,22	0,95	0,0032	0,0072	0,939	0,454	**0,426**	0,757	0,696	**-0,053**
5	4	0,22	0,95	0,0004	0,0046	0,963	0,575	**0,314**	0,878	0,866	**0,008**
10	8	0,22	0,95	0,0112	0,0272	0,958	0,169	**0,777**	0,701	0,638	**0,063**
30	15	0,15	1,15	0,076	0,0833	0,928	0,294	**0,61**	0,659	0,667	**-0,008**
100	35	0,025	1,25	0,0746	0,0787	0,933	0,344	**0,55**	0,66	0,684	**0,022**

In einer zweiten Evaluierung sollen die Auswirkungen auf das Dominanzproblem genauer

untersucht werden. Hierzu vergleichen wir unseren Ansatz mit der Min/Max-[WCB06] und der Z-Score-Normalisierung [MA01]. Grundlage der Evaluierung bilden jeweils drei zufällig erzeugte Distanzverteilungen (Normalverteilung, Gleichverteilung und ein Mischverteilung aus Normal- und Gleichverteilung). Die hierfür genutzten Distanzen basierten auf zwei unterschiedlich großen Intervallen ($[0.2, 0.3]$ vs. $[0.2, 0.9]$). Darüber hinaus soll die Anfälligkeit gegenüber Ausreißern geprüft werden. Hierzu wurden der Verteilung mit dem kleinen Intervall (die dominierte Verteilung) künstlich Distanzen außerhalb des eigentlichen Intervalls hinzugefügt, um die Dominanz künstlich zu reduzieren. Die Verteilung enthielt dabei einen verschmutzten Anteil von 0.5% aller Distanzen. Abbildung 2 zeigt die Ergebnisse der Evaluierung. Zur Messung der Streuung wurden die Experimente 1000 mal wiederholt und der durchschnittliche Kalibrierungsfehler nach der Normalisierung gemessen. Besonders bei Verteilungen mit Ausreißern und bei unterschiedlichen Verteilungen zeigt der erweiterte MDS-Ansatz seine Vorteile. Leichte Abweichungen in den Ergebnissen des MDS-Verfahrens lassen sind durch die zufällig erzeugten Distanzen und die möglicherweise nicht optimal gewählten MDS-Parameter erklären (vgl. Kapitel 4.2). Dies und die Tatsache, dass Z-Score-Normalisierung beste Ergebnisse bei einer Normalverteilung erzielt, erklären auch den minimal höheren Kalibrierungsfehler bei Experimenten mit Normalverteilung ohne Outlier. Insgesamt kann das Verfahren dennoch als robust gegenüber Ausreißern oder verschiedenen Verteilungen bezeichnet werden.

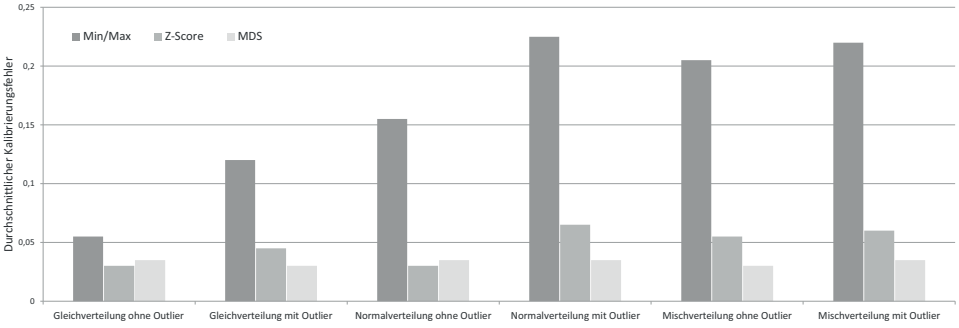


Abbildung 2: Durchschnittlicher Kalibrierungsfehler nach der Normalisierung

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Ansatz vorgestellt, der das Problem fehlender Metrikeigenschaften sowie der Dominanz eines Features bzgl. einer Aggregation mit Hilfe einer Erweiterung der multidimensionalen Skalierung löst. Es konnte gezeigt werden, dass unter Erhaltung der Reihenfolge eine Distanzmatrix erzeugt werden kann, die alle Eigenschaften einer Metrik erfüllt, insbesondere der Dreiecksungleichung. Weiterhin konnte gezeigt werden, dass die MDS auch das Dominanzproblem reduzieren und im optimalen Fall beseitigen kann.

In zukünftigen Arbeiten soll das Verfahren vor allem im Hinblick auf die Laufzeit weiter verbessert werden. Eine Kombination mit dem MDS-nahen Ansatz *FastMap* könnte eine effizientere Berechnung auch auf größeren Datenmengen ($n > 10000$) ermöglichen.

Literatur

[ABE+55] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid und Edward Silverman. An Empirical Distribution Function for Sampling with Incomplete Information. *Ann. Math. Statist.*, 26(4):641–647, 1955.

[BG05] I. Borg und P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[BS14] Thomas Böttcher und Ingo Schmitt. Dominanzproblem bei der Nutzung von Multi-Feature-Ansätzen. In *Grundlagen von Datenbanken*, 2014.

[FL95] C. Faloutsos und K. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In *Proceedings of 1995 ACM SIGMOD*, Seiten 163–174, 1995.

[FVC06] Miriam Fernández, David Vallet und Pablo Castells. Probabilistic Score Normalization for Rank Aggregation. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika und Alexei Yavlinsky, Hrsg., *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings*, Jgg. 3936 of *Lecture Notes in Computer Science*, Seiten 553–556. Springer, 2006.

[GR72] Paul E. Green und Vithala R. Rao. *Applied multidimensional scaling; a comparison of approaches and algorithms*. Holt, Rinehart and Winston New York, 1972.

[Kru64a] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, Marz 1964.

[Kru64b] J.B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964.

[Kub12] Robert Kuban. Analyse von Kalibrierungsansätzen für die CQQL-Auswertung. Bachelor's thesis, University of Cottbus-Senftenberg, Germany, 10 2012.

[MA01] Mark Montague und Javed A. Aslam. Relevance Score Normalization for Metasearch. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, Seiten 427–433, New York, NY, USA, 2001. ACM.

[MRF01] R. Manmatha, T. Rath und F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, Seiten 267–275, New York, NY, USA, 2001. ACM Press.

[RDJH73] Barlow R.E., Bartholomew D.J., Bremner J.M. und Brunk H.D. Statistical inference under order restrictions. *Statistica Neerlandica*, 27(4):189–189, 1973.

[Sam06] Hanan Samet. *Foundations of Multidimensional And Metric Data Structures*. Morgan Kaufmann, 2006/08/08/ 2006.

[Sko06] Tomás Skopal. On Fast Non-metric Similarity Search by Metric Access Methods. In *EDBT*, Seiten 718–736, 2006.

[Spe04] C. Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.

[WCB06] Shengli Wu, Fabio Crestani und Yaxin Bi. Evaluating Score Normalization Methods in Data Fusion. In *Proceedings of the Third Asia Conference on Information Retrieval Technology*, AIRS'06, Seiten 642–648, Berlin, Heidelberg, 2006. Springer-Verlag.