Herausforderungen in der Anwendbarkeit von Metriken: Bias, Effizienz und Hubness[1]

Abdel Aziz Taha[2]

Abstract: Metriken spiegeln die Ähnlichkeiten bzw. Unterschied zwischen Objekten in Merkmalräumen wider. Es gibt in dieser Hinsicht drei Hauptprobleme: Erstens existieren hunderte von Metriken, die verschiedene Aspekte der Ähnlichkeit berücksichtigen, was den Bedarf an einer formalen Auswahlmethodik für Metriken motiviert. Für dieses Problem präsentieren wir eine detaillierte Analyse von 20 Metriken und präsentieren eine neue formale Methode für Metrikauswahl vor. Zweitens gibt es rechenintensive Metriken, deren ineffiziente Laufzeit in Verbindung mit großen Objekten ein Problem darstellt. Wir schlagen einen neuen beinahe zeit-linearen Algorithmus zur Berechnung der exakten Hausdorff-Distanz zwischen beliebigen Punktwolken vor. Drittens taucht in hoch-dimensionalen Featurespaces eine Kategorie von Schwierigkeiten auf, die als *curse of dimensionality* bekannt ist. Eine dieser Schwierigkeiten ist Hubness. Wir präsentieren eine neue Erklärung für die Ursache von Hubness, die auf einem neuen Modell der Distanzstruktur in hoch dimensionalen Datenräumen beruht. Auf Grundlage dieser Erklärung leiten wir einen Schätzer für Hubness ab, bzw. schlagen wir Verfahren zur Verringerung von Hubness vor.

1 Einführung

Metriken sind Funktionen, die auf Featurespace definiert sind, um Ähnlichkeiten bzw. Unterschiede zwischen Objekten (Punktwolken) widerzuspiegeln. Es gibt hunderte von Metriken, von denen jede nur bestimmte Aspekte der Ähnlichkeit misst. Daher repräsentieren Metriken verschiedene Sichtweisen der Realität. Das resultiert aus den unterschiedlichen Sensitivitäten bzw. Biases (Verzerrungen) der Metriken, was den Bedarf an Methodik für Metrikauswahl motiviert. Als erster Teil dieses Dokuments präsentieren wir in Abschnitt 2 Lösungen, betreffend Metrikbias und Metrikauswahl: Zuerst präsentieren wir eine Analyse eines Sets von 20 Metriken für die Evaluierung medizinischer Bilder und schließen diese Analyse mit Metrik-Auswahlrichtlinien. Dann präsentieren wir eine generelle formale Methodik für Metrikauswahl zwecks Evaluierung beliebiger Objekte.

Ein anderes Problem mit Metriken ist die Effizienz der Berechnung rechenintensiver Metriken, z.B. solche, die die Abstände zwischen allen möglichen Punkt-Paaren berücksichtigen. Die Berechnung kann hier extrem ineffizient sein, insbesondere wenn Objekte verglichen werden, die aus einer enormen Anzahl von Punkten bestehen. Ein Beispiel ist die Berechnung der Hausdorff-Distanz zwischen Magnetresonanztomographiebildern (MRI). Solche Bilder können aus bis zu 200 Mio Punkten (z.B. ganz Körper MRI Images) bestehen. Wir schlagen in Abschnitt 3 einen neuen Algorithmus für Berechnung von der

[1] Englischer Titel der Dissertation: Addressing metric challenges: Bias and Selection - Efficient Computation - Hubness Explanation and Estimation

[2] Vienna University of Technology, Institute of Software Technology and Interactive Systems.

Hausdorff-Distanz zwischen beliebigen Punktwolken vor, welcher eine beinahe lineare Zeitkomplexität hat.

In hoch-dimensionalen Featurespaces tritt eine neue Kategorie von Problemen in Verbindung mit Metriken auf, welche als Fluch der Dimensionalität (curse of dimensionality) bekannt ist, z.B. Sparsity (Spärlichkeit), Distanz-Konzentration und Hubness. Diese Schwierigkeiten können auch als Sonderfälle der Metrik-Sensitivität (asymptotische Neigung) betrachtet werden, welche entstehen, wenn die Dimensionalität ausreichend hoch ist. Sie führen dazu, dass Metriken in IR und ML Algorithmen nicht mehr genau sind. Wir präsentieren in Abschnitt 4 eine neue Erklärung der Ursache von Hubness, die auf einem neuen Modell beruht, welches die Distanzstruktur in hoch-dimensionalen Featurespaces repräsentiert. Aufgrund dieser Erklärung leiten wir einen zeit-linearen Schätzer für Hubness ab und schlagen wir ein neues Verfahren zur Verringerung von Hubness vor.

In den nächsten Abschnitten erläutern wir diese Lösungen bündig, welche ausführlich in der Dissertation Taha [**?**] beschrieben sind.

2 Metrik-Bias und Metrik-Auswahl

Metrik Bias und Sensitivität stellen eine Herausforderung für Metrikauswahl dar. In diesem Abschnitt präsentieren wir eine Analyse von 20 Metriken für Validierung medizinischer Segmentierungen und eine generische Methode für die Auswahl von Evaluierungsmetriken.

2.1 Metriken für Segmentierung medizinischer Bilder

In diesem Abschnitt analysieren wir die Biases von 20 Metriken für die Validierung von Segmentierung medizinischer Bilder. Dieses Set von Metriken wurde anhand einer Literaturrecherche festgelegt, d.h. es wurden nur Metriken berücksichtigt, die häufig in diesem Gebiet angewendet werden. In dieser Analyse zeigen wir insbesondere, wie die verschiedenen Metriken unterschiedliche Sensitivitäten zu den unterschiedlichen Eigenschaften der Segmentierungen haben. Ausführliche Details über diese Analyse sind in Taha et al. [**?**].

2.1.1 Analyse Metrik-Eigenschaften

Der erste Ansatz der Analyse ist, die Korrelation zwischen den 20 Metriken zu überprüfen. Für diesen Zweck wurden 4833 maschinell erzeugten medizinischen Segmentierungen jeweils mit den zugehörigen Ground Truth Segmentierungen verglichen. Dies wurde unter Verwendung jeder der 20 Metriken wiederholt, und dann wurden die paarweisen Korrelationen zwischen den Metriken berechnet. Die Ergebnisse dieses Experimentes zeigen, dass Metriken sich nach Korrelation in drei Gruppen teilen, wobei Metriken in der jeweiligen Gruppe stark miteinander korrelieren, jedoch mit Metriken in anderen Gruppen

schwach, gar nicht oder sogar invers korrelieren. Die schwache (oder inverse) Korrelation zwischen Metriken, die in der gleichen Domain häufig benutzt werden, ist eine eindeutige Motivation für die Notwendigkeit an einer standardisierten Disziplin für Metrikauswahl.

Um die Biases und Sensitivitäten der Metriken zu verstehen, möchten wir unterschiedliche Korrelationen zwischen den Metriken zu erklären. Es läßt sich zeigen, dass es zwei unterschiedliche Kategorien von Ursachen der Korrelationsunterschied gibt.

- Ursachen in der Metrikdefinitionen: Ein tiefer Einblick in den Metrikdefinitionen zeigt, dass wir es mit drei Korrelationsgruppen mit Definitionsaspekten zu tun haben, z.B. ob die *True Negatives* in der Definition herangezogen werden oder nicht bzw. ob die Definition eine Behandlung von per Zufall erfolgter Segmentierung vorsieht oder nicht. Korrelationsunterschiede, die durch diese Kategorie von Ursachen verursacht sind, sind Metrik-charakteristisch, d.h. sie hängen nicht von den zugrundeliegenden Segmentierungen ab.
- Ursachen in den Segmentierungen: Die Metrikkorrelationen hängen auch von den validierten Segmentierungen ab. Betrachtet man z.B. die Korrelationen unter Berücksichtigung der Qualität der Segmentierungen (d.h. der durchschnittliche Überlappung zwischen Segmentierungen und Ground Truth), findet man, dass die Korrelation von der Qualität stark abhängt.

In den folgenden Paragraphen zeigen wir Beispiele, wie Eigenschaften der Segmentierungen die Resultate der verschiedenen Metriken beeinflussen und wie diese mit dem Zweck der Segmentierung zusammenhängen können.

Abgrenzungsfehler: Anatomische Strukturen können unterschiedlich komplexe Formen im Sinne der Ränder und Abgrenzungen haben. Manche sind rund und glatt, wie die Nieren, und manche haben komplexe Formen wie Adern und Blutgefäße. Metriken haben unterschiedliche Fähigkeiten Abgrenzungsfehler zu entdecken. Die Wichtigkeit der Abgrenzungsgenauigkeit hängt vom Zweck der Segmentierung ab. Wenn diese Segmentierung z.B. der Beobachtung des Fortschritts eines Tumors oder der Visualisierung eines Organs dient, könnte die Abgrenzungsgenauigkeit von Bedeutung sein. Überlappung- bzw. Volumen-basierte Metriken sind z.B. dafür nicht gut geeignet, weil Abgrenzungsfehler nicht abgefangen werden, da nur das Gesamtvolumen bzw. die Überlappung wichtig ist. Im Gegenteil sind Distanz-basierte Metriken in diesem Fall besser geeignet. Abgrenzungsgenauigkeit ist nicht immer erwünscht. Wenn die Segmentierung z.B. der Entfernung eines Tumors dient, ist die Abgrenzungsgenauigkeit weniger wichtig. Vielmehr ist hier die Recall relevant, d.h. dass die Segmentierung den gesamten Tumor enthält, damit er nicht erneut nachwachsen kann. Metriken, die Recall belohnen (z.B. mutual information), sind in diesem Fall gut geeignet.

Segment-Dichte: Abhängig von den verwendeten Methoden in den Segmentierungsalgorithmen können Segmente unterschiedlich dicht sein. Da die Dichte des Segments einen direkten Einfluss auf die Überlappung bzw. das Volumen hat, würden Überlappung- bzw. Volumen-basierte Metriken einen Algorithmus mit einer höhen Abgrenzungsgenauigkeit

benachteiligen, wenn er Segmentierungen mit weniger Dichte produziert. Distanz basierte Metriken sind in diesem Fall besser geeignet.

Segmentsgröße: Je kleiner das Segment ist, desto kleiner ist die Wahrscheinlichkeit einer Überlappung, weil der Freiheitsgrad der Segmentposition großer ist. Bei sehr kleinen, z.B. punktuellen, Segmenten, ist die Wahrscheinlichkeit, dass es überhaupt keine Überlappung gibt bzw. dass die Segmente weit voneinander sind, groß. Ist das der Fall, so versagen die Überlappung- bzw. Volumen-basierte Metriken in Unterscheidung zwischen nahen und weiten Segmenten, denn in beiden Fällen ist die Überlappung null. Distanz basierte Metriken sind in diesem Fall besser geeignet.

2.1.2 Rechtlinien für Metrikauswahl

In Abschnitt 2.1.1 zeigten wir anhand von Beispielen, dass Metriken verschiedene Biases zu den Eigenschaften der validierten Segmentierungen haben. Wir zeigten auch, dass diese Biases abhängig vom Zweck der Segmentierungen erwünscht oder nicht erwünscht sein können. Generell lässt sich sagen, dass Metrikbiases nur dann erwünscht sind, wenn sie erwünschte Segmentierungs-Eigenschaften belohnen oder unerwünschte Eigenschaften benachteiligen. Diese Überlegung haben wir angestellt, um ein formales Protokoll für die Metrikauswahl für die Evaluierung medizinischer Segmentierungen zu erstellen. Dafür wurden zuerst generische Eigenschaften für Segmentierungen definiert, wie Größe, Dichte, Komplexität der Abgrenzungen, etc und dann wurden generische Anforderungen an Segmentierungs-Algorithmen definiert, z.B. genaue Abgrenzungen, Maximierung von Recall, Sensitivität zu Ausreißer, etc. Schließlich wurde ein Protokoll erstellt, das jede Metrik entweder empfehlt oder davon abrät, und zwar basierend auf dem Vorhandensein von diesen generischen Eigenschaften und Anforderungen.

2.2 Ein Framework für automatischen Auswahl von Evaluierungsmetriken

Ist ein analytisches Verfahren unbedingt notwendig, um ein Metrrikbias festzulegen? In Abschnitt 2.1.2 haben wir Metrikbiases mit Hilfe von Analyse der Metrikdefinitionen, Konstruieren von Beispielen und Heranziehen von dokumentierten Beobachtungen in der Literatur festgelegt. In diesem Abschnitt präsentieren wir eine formale Methode für die Ermittlung von Metrikbias, ohne dabei analytische Verfahren zu verwenden. Basierend auf dieser Methode, schlagen wir eine Methodik für Metrikauswahl in einem beliebigen Validierungsprozess vor. Diese Methoden haben wir in Taha et al. [**?**] publiziert.

Sei M eine Menge von Metriken und sei O eine Menge von Objekten unter Evaluierung. Für jedes Objekt $o_j \in O$ gibt es ein Ground truth Objekt $\hat{o}_j$ mit dem es verglichen (validiert) wird. Sei s_{ij} der Wert des Vergleichs von o_j mit $\hat{o}_j$ unter Verwendung von Metrik m_i. Weiteres sei F eine Menge von Eigenschaften (Ausprägungen), die alle oder einige der Objekte o_i aufweisen. Das Bias von Metrik m_p (die Metrik unter Prüfung) zu der Eigenschaft $f \in F$ wird in zwei Schritten, wie folgt, ermittelt:

Schritt 1, Gruppierung: Die Objekte o_j werden auf zwei Weisen zu n Untermengen gruppiert: Die erste Gruppierung G_z erfolgt zufällig, und die zweite Gruppierung G_f erfolgt nach dem Grad der Ausprägung der Eigenschaft f. Rankt man nun die Gruppen in den beiden Fallen nach ihren durchschnittlichen Metrikwerten s_{ij}, dann bekommt man von jeder Metrik m_i zwei Gruppenrankings, ein Ranking bezüglich der zufälligen Gruppierung G_z, nennen wir es R_{iz} und ein Ranking bezüglich der Gruppierung G_f, nennen wir es R_{if}, d.h. es resultieren insgesamt $2.|M|$ Rankings. Die zwei Rankings, die von der Metrik unter Prüfung (m_f)resultieren, nennen wir R_{pz} und R_{pf}.

Schritt 2, Bias Inferenz: Die Kernidee hinter Bias Inferenz ist folgende: Hat Metrik m_p kein Bias zu Eigenschaft f, so wird erwartet, dass die durchschnittliche Korrelation zwischen dem Ranking, generiert durch die getestete Metrik (m_p) und den Rankings, generiert durch die restlichen Metriken, gleich bleibt, egal ob die zufällige Gruppierung G_z oder die Gruppierung nach der Eigenschaft G_f in Betracht genommen wird. Liegt ein Unterschied in der Korrelation zwischen den beiden Fällen, dann bedeutet das, dass die Metrik m_p Bias bezüglich der Eigenschaft f hat. Die Stärke dieses Bias entspricht dem Unterschied in der Korrelation. Formal heißt das, $B(m_p, f) \approx Corr(R_{pz}, R_{iz}) - Corr(R_{pf}, R_{if})$, wobei $B(m, f)$ ist das Bias von Metrik m zu Eigenschaft f und $Corr(x, y)$ ist die durchschnittliche Korrelation zwischen Metrik x und alle anderen Metriken y.

Um nun eine Metrik aus der Menge M zu selektieren, wird für jede Metrik das Gesamtbias (Summe der Biases über alle Eigenschaften in F) berechnet, und die Metrik mit dem kleinsten Bias wird ausgewählt. Die detaillierte Form der Methode berücksichtigt Eigenschaftsgewichtung bzw. die Richtung der Bias (Belohnung oder Benachteiligung), welche aus Platzmangel hier ausgelassen wurden.

3 Effiziente Metrikberechnung

Metriken zwischen zwei Punktwolken, die auf Messung von allen paarweisen Distanzen beruhen, z.B. die Hausdorff Distanz, sind sehr zeitaufwendig, insbesondere wenn die Punktwolken enorm groß sind. In diesem Abschnitt stellen wir eine Methode für zeitlineare Berechnung der Hausdorff Distanz zwischen beliebigen Punktwolken vor.

Die Hausdorff Distanz H zwischen zwei beliebigen Punktwolken A und B ist das Maximum der Distanzen zwischen jedem Punkt $x \in A$zu seinem nächsten Nachbarn $y \in B$. Das heißt:

$$H(A,B) = max_{x \in A}\{min_{y \in B}\{||x,y||\}\} \quad (1)$$

Wobei $||.,.||$ eine beliebige Distanz-Norm, z.B. die Euklidische Distanz, ist.

Eine naive Berechnung der HDD bedeutet das Durchlaufen zweier verschachtelter Schleifen. Die Erste (die innere Schleife) läuft über die Punktwolke B, um die nächsten Nachbarn zu finden, und die Zweite (die äußere Schleife) läuft über die Punktwolke A, um das Maximum zu finden. Das hat offenbar eine Zeitkomplexität von $O(|A|.|B|)$, was ein Effizienzproblem mit sehr großen Punktwolken bereitet. In diesem Abschnitt präsentieren wir eine zeit-lineare Berechnungsmethode der Hausdorff Distanz, welche wir in Taha et al. [**?**] publiziert haben.

Die vorgeschlagen HDD Berechnungsmethode basiert auf zwei Hauptstrategien, die im Folgenden kurz erläutert sind:

Das frühzeitige Brechen (*early break*): Hier werden unnötige Berechnungen vermieden, indem die innere Schleife immer unterbrochen wird, sobald festgestellt wird, dass der Rest der Schleife das Ergebnis nicht ändern kann. Das ist genau der Fall, wenn eine Distanz gemessen wird, die kleiner als die derzeitige HDD (*cmax*) ist. Das wird so erklärt: *cmax* ist das aktuelle Maximum, das als Ergebnis der Iterationen der äußeren Schleife resultiert. Da die äußere Schleife eine Maximierungsschleifen ist, bedeutet das, dass *cmax* monoton steigend ist. Nun unter der Betrachtung, dass die innere Schleifen eine Minimierungsschleifen ist, wenn in einer Iteration der inneren Schleife eine Distanz d gemessen wird, die kleiner als *cmax* ist, bedeutet das, dass diese innere Schleife sofort unterbrochen werden kann, denn das Durchlaufen dieser Schleife kann nie eine Distanz finden, die größer als *cmax* ist.

Die Randomisierung (*randomization*): Diese Strategie nutzt das Lokalitätsprinzip, um die Auswirkung vom *early break* zu maximieren, indem die Reihenfolgen der Abarbeitung in den Schleifen optimiert werden. Wir erklären das wie folgt: Wir wollen, dass das *early break* so oft wie möglich auftritt, und zwar jedes Mal möglichst früh in der inneren Schleife. Wie oben erwähnt, tritt kein early break, wenn die gemessene Distanz gr??er als *cmax* ist. Werden nun die Punkte in der natürlichen Reihenfolge (z.B. Scannen der Pixels eines Bildes von links nach rechts und von oben nach unten) abgearbeitet, dann gilt folgendes: Wenn in einer Iteration in der inneren Schleife kein *early break* auftritt, dann ist es wahrscheinlich, dass in der nachfolgenden Iteration ebenfalls kein *early breaks* wird. Deshalb schlagen wir die Randomisierung der Abarbeitungsreihenfolgen in den inneren und äußeren Schleifen vor, was die Chance für *early breaks* maximiert.

Die Anwendung von *early break*, kombiniert mit *randomization*, erzielt eine erhebliche Effizienzsteigerung gegenüber state-of-the-art Algorithmen. Der vorgeschlagene Algorithmus war in Tests (i) mit Zufallsvariablen um 4,8, (ii) mit realen medizinischen 3D Bilder um 7,8, und (iii) mit realen Straßen-Netzwerken um 30 Mal schneller als die state-of-the-art Algorithmen.

4 Formale Analyse von Hubness

Hubness ist ein Phänomen, das auftritt, wenn k-Nächst-Nachbar (k-NN) Algorithmen auf hoch-dimensionalen Datenräumen angewendet wird. Hubness ist durch Auftritten von Hubs gekennzeichnet. Diese sind Objekte, die weit öfter als erwartet als nächste Nachbarn von anderen Objekte gefunden werden. Unti-Hubs sind im Gegenteil Objekte, die sehr selten oder gar nicht als nächste Nachbarn gefunden.

Sei X eine Punktwolke mit $x \in X$. Wendet man nun einen k-NN Algorithmus auf X an, so definieren wir die Funktion $N_k(x)$ als die Anzahl der Punkte in X, für die x als nächster Nachbar gefunden wurde. Wir definieren Hubness als die Verzerrung (Asymmetrie) von der Verteilung der Funktion $N_k(x)$. Bitte beachte, dass die Funktion $N_k(x)$ den Werk k

als Erwartungswert hat. Nun sind Hubs Objekte, welche wesentlich weit über k mal als nächste Nachbarn gefunden werden.

In Abschnitt 4.1, präsentieren wir eine neue Erklärung für die Entstehung von Hubness, basierend auf einem neuen Modell für Distanzen in einem hoch-dimensionalen Raum. In Abschnitt 4.2 schlagen wir Applikationen vor, die auf dieser Erklärung basieren, nämlich einen Schätzer für Hubness und ein Verfahren für Hubness-Verringerung. Dieser Abschnitt ist detaillierter in Taha et al. [**?**] erläutert.

4.1 Entstehung von Hubness

Der begriff *distance concentration* (DC) bezeichnet ein besonderes Verhalten der Distanz im hoch-dimensionalen Raum, demzufolge die Distanz zum Mittelpunkt und die paarweise Distanz konvergieren, wenn die Dimensionalität ausreichend groß ist. Das heißt, Punkte sind (i) auf dem gleichen Abstand zueinander und (ii) auf dem gleichen Abstand zum Mittelpunkt.

Nun stellt sich die Frage, welche Distanzstruktur es in hoch-dimensionalem Raum gibt, die die Merkmale (i) und (ii) gelten lässt? In die Dissertation [**?**] zeigen wir mit Methoden, die aus Platzgründen hier nicht präsentiert werden können, dass die Distanz im hoch-dimensionalen Raum so konvergiert, dass die Datenpunkte nur in bestimmten Orten relativ zu den Orthants des Datenraums residieren können, sodass die Datenpunkte die Ecken eines Hyperwürfels bilden. Wir verstehen unter einem Orthant hier einen Teilraum im Hyperraum analog zu einem Quadrant im zwei-dimensionalen Raum. Abbildung 1 (A) veranschaulicht im zwei-dimensionalen Raum, wie Punkte, die in bestimmten Orten relativ zu den Quadrants residieren, ein Quadrat bilden. Offenbar gibt es 2^d Orthants in einem

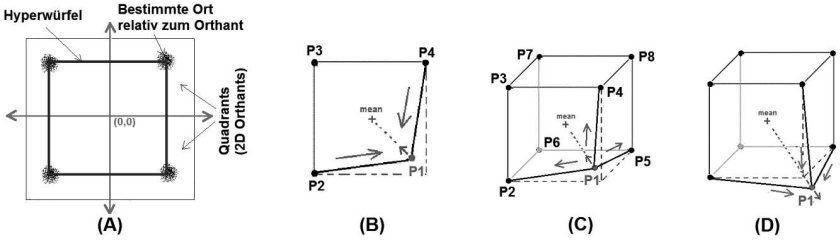

Abb. 1: Illustration von Hyperwürfel-Modell in 2D: (A) Punkte, die in bestimmten Orten relativ zu den Orthants (Quadrants) konzentriert sind, bilden die Ecken eines Hyperwürfels (Quadrat). (B) Ein Punkt, der vom Eck in Richtung Mittelpunkt abweicht, wird näher zu den benachbarten Punkten. (C), wie in (B), aber in 3D. (D) Ein Punkt, der vom Eck in Richtung nach außen abweicht, wird weiter von den benachbarten Punkten.

d-dimensionalen Raum. Berücksichtigt man nun die enorme Anzahl von Orthants in einem hoch-dimensionalen Raum (z.B. $2^{100} \approx 10^{30}$), so kann man davon ausgehen, dass in einem Orthant höchstens ein einziger Punkt residiert. Wir fassen bisher unser Modell für Distanzstruktur in hoch-dimensionalen Datenräumen so zusammen: Hoch-dimensionale Datenpunkte besetzen die Ecken eines Hyperwürfels, wobei es auf einer Ecke höchstens

einen Punkt gibt. Beachten Sie, dass die Punkte nur dann exakt auf den Ecken stehen, wenn die Dimensionalität unendlich ist. In der Praxis ist sie es aber nicht. Deshalb können die Punkte von den Hyperwürfel Ecken abweichen. Ab jetzt werden wir dieses als das Hyperwürfel-Modell bezeichnen.

Nun möchten wir Hubness, basierend auf dem Hyperwürfel-Modell, erklären, wobei wir zur Illustration einen zweidimensionalen Raum betrachten. Stellen wir uns vier Punkte vor, die exakt auf den Ecken eines Quadrats stehen. Da die Abstände zwischen diesen Punkten gleich sind, erwarten wir kein Hubness, da jeder Punkt die gleiche Chance hat, als nächster Nachbar gefunden zu werden. Weil die Punkte in der Praxis nicht unbedingt exakt auf den Ecken stehen müssen, nehmen wir an, dass ein Punkt von der Ecke in Richtung Mittelpunkt abweicht, wie es Abbildung 1 (B) veranschaulicht. Das führt dazu, dass Punkt 1 näher zu Punkt 2 und 4 liegt und als ihr nächster Nachbar gilt. Daher wird P1 zu einem Hub. Wie schaut es aus, wenn wir einen dreidimensionalen Raum (also eine Dimension mehr), als in Abbildung 1 (C), betrachten? Wenn P1 aus der Würfelecke in Richtung Mittelpunkt abweicht, wird P1 diesmal näher zu drei Punkten werden, nämlich P2, P4, und P5, also um einen Punkt mehr als in 2D-Raum. Hyperwürfel-Geometrie besagt, dass ein Eck in einem d-dimensionalen Hyperwürfel mit d Kanten verbunden ist. Deshalb wird ein vom Eck in Richtung Mittelpunkt abweichender Punkt näher zu d Ecken werden und vermutlich für alle auf diesen Ecken stehenden Punkten als nächster Nachbar gelten. Das erklärt, warum Hubness mit steigender Dimensionalität stärker wird. Eine Abweichung in die Gegenrichtung, also weg vom Mittelpunkt, wie es in Abbildung 1 (D) veranschaulicht ist, wirkt umgekehrt: P1 wird in diesem Fall weiter weg von den drei benachbarten Ecken werden und gilt daher zu keinem als nächster Nachbar. P1 wird deshalb zu einem Unti-Hub.

Eigentlich können Punkte in beliebigen Richtungen abweichen. Bisher haben wir nur von Abweichungen gesprochen, die sich auf der Hyperwürfels-Achse (zum Mittelpunkt und weg vom Mittelpunkt) bewegen. Wir bezeichnen diese als die *radiale Abweichung*. Betrachten wir nun Abweichungen orthogonal dazu (normal auf die radiale Richtung). Offenbar handelt es sich hier um einen $(d-1)$-Teilraum. Wir bezeichnen diese Richtung als die *tangentiale Abweichung*.

Wie wirken tangentiale Abweichungen auf Hubness? Ein Punkt, der tangential abweicht, wird zwar näher zu einigen Ecken, aber gleichzeitig weiter von anderen Punkten werden. Statistisch gesehen hebt sich die Wirkung von tangentialer Abweichung auf, d.h der Hubness des tangential abweichenden Punktes wird dadurch nicht verändert.

Auf jeden Fall wirken viele tangentiale Abweichungen Hubness-mindernd auf andere radial abweichende Punkte. Wir erklären das so: Betrachten wir noch einmal Abbildung 1 (C) und unterscheiden wir zwischen zwei Fällen, während P1 in Richtung Mittelpunkt abweicht. Fall 1: alle anderen Punkte stehen exakt auf den Ecken, und Fall 2: viele Punkte weichen in beliebigen tangentialen Richtungen ab. Im Fall 1 wird P1 definitiv nächster Nachbar zu allen benachbarten Punkten, egal wie klein die Abweichung ist. Im Fall 2 wird dieser Effekt jedoch durch die tangentiale Abweichung geschwächt, und hängt von dem Verhältnis zwischen der radialen Abweichung und der mittleren tangentialen Abweichung ab. D.h. Man kann dieses Verhältnis als ersten groben Schätzer für Hubness nehmen, d.h.

das Hubness eines Datensatz X ist

$$Hubness(X) \approx \frac{mittlere\ radiale\ Abweichung}{mittlere\ tangentiale\ Abweichung} \tag{2}$$

Selbstverständlich können Abweichungen in beliebiger Richtung (nicht nur radial und tangential) sein, aber egal in welcher Richtung sie erfolgen, sie können immer in zwei Komponenten zerlegt werden, eine Komponente in die radiale Richtung und eine in die tangentiale Richtung.

4.2 Hubness Schätzung und Minderung

In diesem Abschnitt erläutern wir in Kürze, wie die Hubness-Erklärung im Abschnitt 4.1 verwendet wird, um zeit-lineare Hubness Schätzer zu definieren. Formel 2 besagt, dass das Verhältnis zwischen mittlerer radialen und mittlerer tangentialen Abweichung ein Indikator von Hubness ist. Anstatt die Abweichungen der einzelnen Punkte direkt zu messen, kann man die mittleren Abweichungen leichter und effizienter mithilfe von Statistiken der Distanz zum Mittelpunkt (DTM) und der paarweisen Distanz (PWD), wie folgt, schätzen:

Die mittlere radiale Abweichung: Der Varianz der DTM ist offenbar ein direkter Schätzer für den Umfang der mittleren radialen Abweichung, was leicht zu zeigen ist. Ein weiteres wichtiges Merkmal der DTM Verteilung sind Ausreißer (Outliers). Wenn Outliers auf der linken Seite der Verteilung liegen, dann weisen sie auf Punkte hin, die stark in Richtung Mittelpunkt abweichen, die also mit hoher Wahrscheinlichkeit Hubs sind. Wenn die Outliers im Gegenteil auf der rechten Seite liegen, dann weisen sie auf Unti-Hubs hin. Schließlich ist die Wölbung der DTM Verteilung ein wichtiger Indikator für den Umfang der mittleren radialen Abweichung, denn eine stark gewölbte Verteilung besteht aus zwei Teilen: (i) Eine Masse in der Mitte, die darauf hinweist, dass viele Punkte auf den Ecken konzentriert sind. (ii) Zwei relativ langen Ausläufern auf den Seiten, die darauf hinweisen, dass einige Punkte von den Ecken radial abweichen.

Mittlere tangentiale Abweichung: Der Varianz der PWD ist ein Schätzer der gesamten mittleren Abweichung (radial und tangential) und kann daher verwendet werden, um die mittlere tangentiale Abweichung zu schätzen. Da der PWD Varianz zeit komplex ist, reicht eine in linearer Zeit gerechnete Stichprobenschätzung, welche sich in den Tests als völlig ausreichend erwiesen hat.

Ein Hubness-Schätzer, basierend auf die oben genannten Statistiken, wird in linearer Zeit gerechnet und hat eine starke Korrelation mit dem naiven sehr Zeit-komplexen Algorithmus, der auf Messung von Nächst-Nachbar Listen von allen Punkten beruht.

Wir stellen eine Methode für Hubness-Minderung vor, die auf der Entfernung von den Punkten mit den höchsten Hubness-Werten aus dem Datensatz basiert. Diese Punkte können mithilfe eines Hubness-Schätzers ermittelt werden, der dazu definiert ist, um Hubness eines einzigen Punktes anstatt des gesamten Datensatzes zu schätzen. Empirische Tests mit Zufallsvariablen bzw. realen Textdaten zeigen, dass die Entfernung von wenigen Hub-Punkten (ca. 1% von dem gesamten Datensatz) zu erheblichen Hubness-Minderung führt.

5 Conclusion

Dieses Dokument fasst die Dissertation in [**?**] zusammen, in der Lösungen für Schwierigkeiten in der Verwendung von Metriken in Featurespaces aufgezeigt werden. Diese Lösungen liegen in drei Gebieten. Im ersten Gebiet wird eine umfassende Analyse von 20 Metriken für die Validierung von Segmentierungen medizinischer Bilder durchgeführt. Diese Metriken werden nach ihren Eigenschaften und die Korrelation zwischen ihnen analysiert. Weiteres werden zusätzlich die Zusammenhänge zwischen generischen Eigenschaften der Segmentierungen und generischen Anforderungen analysiert, um Richtlinien für Auswahl von Metriken zu definieren. Die Auswahlmethodik wird für andere Domänen in Form von einer formalen Methodik für die Metrikauswahl generalisiert. Im zweiten Gebiet wird ein zeit-linearer Algorithmus für die Berechnung des Hausdorff Distanz präsentiert, welcher allgemein und an beliebigen Punktwolken anwendbar ist. Im dritten Gebiet wird eine neue Erklärung des Hubness-Phänomens präsentiert. Basierend auf dieser Erklärung, werden Schätzer für den Umfang von Hubness von einem Datensatz bzw. von einem bestimmten Punkt vorgeschlagen. Weiters wird eine Methode für Hubness-Minderung vorgestellt, die auf der Entfernung von den Top-Hub-Punkten beruht, welche durch Einsatz des vorgeschlagenen Schätzers ermittelt werden können.

Literaturverzeichnis

[Ta15] Taha, Abdel Aziz: Addressing metric challenges: Bias and Selection Efficient Computation Hubness Explanation and Estimation. Dissertation, Vienna University of Technology, December 2015. http://ifs.tuwien.ac.at/˜taha/phd_thesis.pdf.

[TH15a] Taha, Abdel Aziz; Hanbury, Allan: An Efficient Algorithm for Calculating the Exact Hausdorff Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37:2153–2163, Mar 2015.

[TH15b] Taha, Abdel Aziz; Hanbury, Allan: Metrics for Evaluating 3D Medical Image Segmentation: analysis, selection, and tool. BMC Medical Imaging, 15:29, August 2015.

[THJ14] Taha, Abdel Aziz; Hanbury, Allan; Jimenez del Toro, Oscar: A Formal Method for Selecting Evaluation Metrics for Image Segmentation. In: 2014 IEEE International Conference on Image Processing (ICIP) (ICIP 2014). Paris, France, S. 932–936, okt 2014.

Abdel Aziz Taha wurde am 20. Dezember 1970 geboren. Er besuchte die Schule im Palästina (Westjordanland), wo er 1989 auch die Matura erfolgreich abschloss. Von 1992 bis 1996 absolvierte er eine Ausbildung für Elektrotechnik auf der HTL Mödling. Bis 2001 arbeitete er als Elektroingenieur im Industriebereich. Von 2002 bis 2011 arbeitete er in der Softwareentwicklung in industriellen Projekten. Parallel dazu studierte er Informatik auf der TU Wien. 2006 absolvierte er das Bakkalaureatsstudium für Technische Informatik und 2008 das Magisterstudium für Informatikmanagement. Im Jahr 2012 begann er sein Doktoratsstudium auf der TU Wien, wo er auch parallel dazu in der Forschung in der Technischen Universität Wien arbeitete. Ende 2015 schloss er sein Doktoratsstudium mit Auszeichnung ab.