Big data – small school

Oder: Likert-Skalen, Kaufempfehlungen, soziale Netzwerke, das Skalarprodukt und all das

Reinhard Oldenburg

1 Daten – ein Rohstoff? Auch für den Unterricht?

In der politischen Diskussion ist die Bedeutung der Digitalisierung zum Allgemeinplatz geworden. Was hinter big data, künstlicher Intelligenz und all den anderen schillernden Begriffen steckt, scheint aber nicht immer klar zu sein. Es gibt zwar einige populärwissenschaftliche Bücher (z. B. das empfehlenswerte von Mayer-Schönberger und Cukier 2013, und das etwas polemische, aber sehr informative von O'Neil 2016), aber auf dieser Basis versteht man vermutlich zu wenig, um einzuschätzen, welche Möglichkeiten es in Zukunft geben wird, sei es, um selbst mit big data Geld zu verdienen, oder sei es, um politische Regulationsbestrebungen einzuschätzen). Ein bemerkenswertes Vorhaben zur Elementarisierung einiger der Methoden dieses Bereichs unternimmt das Projekt ProDaBi (https://www.prodabi.de/) – allerdings überwiegend aus informatorischer Perspektive. Es bleibt also ein Bildungsbedarf und für den Mathematikunterricht die Frage, ob er auch etwas beitragen kann, um diese neuen Möglichkeiten zumindest teilweise zu erschließen.

Die digitale Welt konfrontiert uns mit vielen scheinbar intelligenten Prozessen. Verblüffenderweise reicht in vielen Fällen sogar Schulmathematik aus, zumindest Prinzipien zu verstehen. Kombiniert mit Fähigkeiten, die der Informatikunterricht vermitteln kann, können viele Ideen sogar selbst realisiert werden. Programmieren kostet aber Zeit, erst im Erlernen (einer der Gründe, weswegen es längst nicht in allen Bundesländern verpflichtend ist), dann in der Anwendung. Nicht nur das, es gibt viele Jugendliche, die gar nicht Programmieren lernen wollen. Kann man trotzdem konkrete und nützliche Vorstellungen über die Bedeutung der Mathematik in diesem Feld vermitteln?

R. Oldenburg (✉)
Institut für Mathematik, Universität Augsburg, Augsburg, Deutschland

Ein zweites Problem ist offensichtlich: Daten als Rohstoff schaufeln die großen Internetunternehmen im Terabytebereich täglich durch die Welt. Sie haben darauf Zugriff, weil Milliarden von Menschen im Gegenzug für gute Suchergebnisse oder einen scheinbar kostenlosen Auftritt in einem sozialen Netzwerk bereitwillig durch Klick „unterschreiben", dass ihre Daten genutzt werden dürfen. In Schulen aber ist man an Datenschutzgesetze gebunden – von beschränkter Rechenleistung und Speicherkapazität mal ganz zu schweigen.

Damit ergibt sich ein didaktisches Problemfeld: Die erste Frage ist, auf welcher „Auflösungsstufe" man die Zusammenhänge der digitalen Welt vermitteln sollte, wie Lernende sie verstehen sollten. Diese Frage, die der passenden Auflösungsstufe, betrifft auch andere Wissenschaften. In der Biologie kann man einen Organismus, seinen Stoffwechsel, seine Steuerungsprozesse auf molekularbiologischer Ebene verstehen, und das ist für einige Fragen die passende Auflösungsstufe, für andere ist sie aber völlig ungeeignet. Wenn man einen Organismus als Ganzes verstehen will oder sogar ganze Ökosysteme, die sich aus sehr vielen Individuen zusammensetzen, dann ist eine wesentlich größere Brille geeigneter, um die relevanten Dinge in den Blick nehmen zu können. Ganz ähnlich liegen die Dinge in der Informatik. In den siebziger und achtziger Jahren war man weitgehend der Meinung, dass man Schüler*innen die Funktionsweise von Computern von der elementaren physikalischen Hardware an erklären sollte. Ausgehend von dem Aufbau der Atome und dem Phänomen der elektrischen Leitfähigkeit, die sich bei Halbleitern durch äußere elektrische Felder leicht beeinflussen lässt, wurde vermittelt, wie sich damit Schalter realisieren lassen und wie mit diesen Schaltern elementare boolesche Funktionen berechnet werden. Aus zwei NAND-Gattern konnte man ein Flipflop bauen, das 1 Bit speichert. Diese wurden aggregiert zu größeren Speichern, die über einen Adressbus angesteuert werden konnten. Es wurde geklärt, wie man mit elementaren Gatterschaltungen die Grundrechenarten umsetzt und

wie ein Taktgeber und ein paar weitere Speicher einen einfachen Prozessor ergeben, der zusammen mit dem Arbeitsspeicher einen universellen Computer darstellt. Es zeigte sich aber bald, dass dieses Verständnis der digitalen Hardware keineswegs nötig war, um mit Computern angemessen umgehen und diese sogar programmieren zu können. Statt eines solchen technischen Detailwissens braucht man offensichtlich nur geeignete mentale Modelle, um auf der Ebene der Interaktion mit dem System kompetent handeln zu können. Beispiel: Um über die elektronische Gesundheitskarte diskutieren zu können, benötigt man keineswegs technisches Wissen über Datenbanken, sondern es reicht aus, über die lokale Speicherung von Daten und die Kommunikation über das Netzwerk Bescheid zu wissen. Damit kann man beispielsweise beurteilen, welche Folgen unzureichend gewählte Passwörter in diesem Bereich haben können und kann mitentscheiden, welche Personengruppe Schreib- bzw. Leserechte für welche Teildatensätze bekommen sollten. Für den Mathematikunterricht ergibt sich die gleiche Fragestellung in einer noch dringlicheren Form: Auch wenn klar ist, dass die Digitalisierung ohne eine breite mathematische Fundierung nicht funktionieren würde, stellt sich die Frage, ob der Mathematikunterricht relevante mathematische Grundlagen vermitteln kann, ob er erfolgreich aufzeigen kann, welche Rolle die Mathematik in der Digitalisierung spielt, und ob dieses Wissen zu einem besseren Verständnis der Bedeutung der digitalen Transformation für die Gesellschaft als Ganzes führt – d. h. ob auf einer passenden Auflösungsstufe die Mathematik überhaupt sichtbar ist. Die Frage ist also letztlich, ob und wie Mathematik für eine digitalisierte Gesellschaft allgemeinbildend unterrichtet werden kann.

Die zweite Frage ist praktischer Natur und deutlich einfacher zu beantworten. Welche Themen lassen sich im Schnittfeld der rechtlichen und technischen Möglichkeiten der Schule und den mathematischen wie informatorischen Fähigkeiten von Schüler*innen intellektuell ehrlich vermitteln?

Diese beiden Fragen sind m. E. wichtig, aber auch kompliziert. Der vorliegende Beitrag wird keine finalen Antworten geben können, sondern exemplarisch eine gestufte oder modularisierte Strategie verfolgen, die eine mögliche Antwort darstellt. Diese Strategie der gestuften Komplexität besteht darin, dass man in einer Abfolge von whitebox-blackbox-Schritten Bausteine kombiniert und dabei die Auflösungsstufe wechselt. Diese Art der Wissensmodularisierung entspricht der Modularisierung großer Softwareprojekte und ist von daher auch ein methodologisch legitimes Lehrziel.

Exemplifiziert wird dies im vorliegenden Beitrag durch das Themenfeld der Empfehlungen auf der Basis sozialer Daten.

2 Daten – Rohstoff-Gewinnung

Wenn man Menschen Empfehlungen geben will, muss man etwas über diese Menschen wissen. Es gibt verschiedene Möglichkeiten, an Daten zu kommen, die dieses Wissen liefern. Die offenste Form ist die einer Befragung. Wenn man mit Schüler*innen diskutiert, wie Studien zur Ermittlung von Präferenzen aussehen könnten, ergeben sich meist zwei einfache Formen. Eine Möglichkeit ist, dass man die Teilnehmer*innen an der Studie einfach aufschreiben lässt, welche Dinge oder Eigenschaften sie gerne haben. Das Ergebnis ist für jede Teilnehmer*in eine Menge von Objekten (oder Begriffen). Für eine systematische oder auch mit dem Computer automatisierte Auswertung ist es sinnvoll, diese Datenstruktur formal zu fassen: Man ermittelt für jede Proband*in eine *Menge* von Begriffen oder Objekten. Eine andere sehr verbreitete Möglichkeit ist die, Begriffe oder Aussagen vorzugeben und die Probanden auf einer Skala einschätzen zu lassen, wie sehr sie diese Dinge mögen oder bestimmten Aussagen zustimmen. Die Skalen können frei sein, sodass die Probanden eine beliebige Zahl (aus einem vorgegebenen Intervall) angeben können, oder man gibt ihnen eine Skala von Zustimmungsmöglichkeiten vor, zum Beispiel „ich stimme der Aussage ganz zu“, „ich stimme der Aussage weitgehend zu“, „ich bin neutral“ und so weiter. Solche Skalen nennt man Likert-Skalen. Die Ergebnisse lassen sich leicht als Zahlen fassen.

Zwei Beispiele für die beiden Typen:

Beispiel 1: Jeder Lernende der Klasse (es seien nur fünf) nennt Sportarten, die ihm/ihr gefallen.

Anna: Tennis, Ski, Schwimmen, Minigolf, Schach
Berta: Joggen, Radfahren, Schwimmen, Inliner, Ski
Clara: Schwimmen, Tennis, Ski, Schach, Volleyball
Dora: Joggen, Judo, Radfahren, Inliner
Elena: Inliner, Radfahren, Judo, Klettern, Schwimmen

Beispiel 2: Wie sehr magst Du die Sportarten? 0 = „gar nicht“… 5 = „sehr“

SoS	Ski	Tennis	Schach	Judo	Klettern	Inliner	Radfahren	Schwimmen
Anna	4	5	3	1	2	0	0	5
Berta	1	2	0	2	1	3	4	4
Clara	5	5	4	2	1	2	1	5
Dora	2	1	1	5	0	4	5	0
Elena	3	2	0	3	4	0	3	4

Damit sind zwei Datenstrukturen gewonnen, die man in Umfragen erheben kann. Beiden Optionen ist gemein, dass

man eine bestimmte Anzahl n von Personen oder Fällen hat, denen etwas zugeordnet wird:

1. Jedem Probanden $i \in \{1, ..., n\}$ wird eine Menge von Objekten M_i (z. B. Begriffe) zugeordnet.
2. Jedem Probanden $i \in \{1, ..., n\}$ wird ein Vektor $v_i \in \mathbb{R}^k$ zugeordnet, bei dem die j-te Komponente $v_{ij}, j \in \{1, ..., k\}$ den Skalenwert des i-ten Probanden bei der j-ten Frage angibt.

Bei der ersten Art der Daten könnte man noch eine Variante erfinden, bei der Mehrfachnennungen erlaubt sind (was z. B. eine Gewichtung ausdrücken kann: „Ich mag Schokolade, Schokolade, Schokolade und nix“), sodass bei jedem Begriff oder Objekt auch eine Anzahl gespeichert wird (technisch gesprochen: Man ersetzt Mengen durch Multimengen).

Bei der zweiten Art der Datenerhebung kann sich das interessante Problem ergeben, dass eine Proband*in auf eine Frage nicht antwortet. Dann muss man sich überlegen, wie man damit umgeht. Entweder die Proband*in wird nicht in die Auswertung einbezogen, oder man nimmt beispielsweise einfach einen mittleren Wert statt der fehlenden Antwort an. Zu welchen Verzerrungen diese Möglichkeiten führen, ist ein interessanter Diskussionsanlass. Auch in der professionellen Statistik ist dieses Problem vielfältig untersucht (Stichwort: Imputation).

Daten dieser Bauarten können auch im Kurs oder der Klasse leicht erhoben werden. Umfragen des ersten Typs beispielsweise können erheben, welche Sportarten oder Musikstücke man mag. Fragen, die sich für den zweiten Typ eignen sind etwa, wie sehr man einige vorausgewählte Künstler*innen mag oder wie man zu politischen Aussagen steht. Nicht mehr Sache der Mathematik, aber trotzdem wichtig und auch in der Schule behandelbar, ist die Frage, wie man im größeren Maßstab an solche Daten herankommt. Im Internet fallen ganz natürlich viele solcher Datenmengen an. Daten des ersten Typs beispielsweise gewinnt eine Versandhändler*in, indem er/sie betrachtet, welche Produkte jede Kund*in bisher gekauft hat. Die Variante, bei der auch Häufigkeiten berücksichtigt werden können, eignet sich beispielsweise für eine Mail-Provider*in, die/der auszählt, welche Worte wie häufig von seinen Kund*innen genutzt werden. Daten des zweiten Typs gewinnt ein soziales Netzwerk, wenn es für jede Benutzer*in notiert, wie häufig er/sie bestimmte Angebote nutzt. Im Informatikunterricht kann besprochen werden, wie man solche Daten aus dem Internet legal auch im Schulunterricht erheben kann – das ist für den Mathematikunterricht nicht zentral.

Exkurs Big data ist ein sehr variabler Begriff, der sich letztlich auf alle Bereiche erstreckt, in denen große (und von Hand nicht mehr beherrschbare) Datenmengen mithilfe von Computern verarbeitet werden können. An große Datenmengen heranzukommen ist keineswegs schwierig. Eine Möglichkeit sind Webcams: Diese liefern zu jedem Zeitpunkt eine Matrix von Helligkeitswerten der drei Farbkanäle. Einfache statistische Kenngrößen bekommen dabei eine greifbare Bedeutung: Der Mittelwert der Werte ist das, was man in der Technik als Bildhelligkeit kennt, und die Standardabweichung ist der Kontrast. Die Daten können in das zweite Szenario eingebettet werden: Jedem Zeitpunkt i wird ein Vektor mit Pixeldaten zugeordnet.

Andere Daten verdankt man dem Trend zu open data: Viele politische und wissenschaftliche Institutionen stellen Daten bereit. Beim Deutschen Wetterdienst kann man etwa für die meisten Stationen ein Tagesprotokoll abrufen, auf dem für jeden Tag der letzten Jahrzehnte eine Reihe von Wetterdaten (Tageshöchst- und -minimaltemperatur, Niederschlagsmenge, Windstärke und -richtung, Luftfeuchtigkeit, …) verzeichnet sind. Auch dieser Datensatz lässt sich dem obigen Format 2 zuordnen, und das wird in einem Beispiel weiter unten verwendet werden.

Soweit zur Datengewinnung: Es zeigt sich, dass Tabellen zwar nicht für alle Situationen, aber doch für recht viele zur Datenmodellierung geeignet sind. Damit kann die genaue Analyse der Situation in einer Blackbox verschwinden und man konzentriert sich auf Analysemethoden, die diese Form voraussetzen. Das zeigt, dass die dann erarbeiteten Methoden unabhängig von der konkreten Situation nützlich sein können. Lernende können also das Modularisierungsprinzip ein weiteres Mal erfahren, das die moderne Wissenschaft und Technik so erfolgreich gemacht hat. Aus didaktischer Sicht ist die Einsicht wichtig, dass die Abstraktion durch die Bildung der Blackbox keine Einbahnstraße sein sollte: Wenn man überlegt, wie man die Daten auswertet, kann es sinnvoll sein, wieder in die Blackbox hineinzuschauen.

3 Daten – Vermessung

Die Vielfalt der möglichen Datenauswertungen ist kaum überschaubar. Lernende verfügen in der Sekundarstufe II bereits über viele statistische Verfahren und Begriffe, die sich hier anwenden lassen: Mittelwerte, Häufigkeiten, etc., … Aber in diesem Kontext liegen viele Fragen nahe, die sich mit den üblichen Mitteln der Schulstatistik nicht gut beantworten lassen. In dem Sport-Beispiel aus dem vorhergehenden Abschnitt könnte man beispielsweise folgende Fragen stellen: Gibt es Gruppen von Schüler*innen, die besonders gut gemeinsam Sport betreiben können? Gibt es zueinander ähnliche Sportarten, die die gleichen Menschen ansprechen? Welche Sportarten, die sie noch nicht mögen oder betreiben, könnte man den Schüler*innen noch vorschlagen?

Der entscheidende Begriff, mit dem man hier weiterkommt, ist der der Ähnlichkeit. Wenn man aus den statistischen Daten ermitteln kann, wie ähnlich zwei Personen oder wie ähnlich zwei Sportarten einander sind, dann kann man Empfehlungen geben.

In Beispiel 1: Welche Schüler*innen sind in ihrem Sportverhalten einander besonders ähnlich? Bei solch kleinen Datenmengen kann man durch „darauf Schauen" einen guten Überblick bekommen und es mag an der Motivation fehlen, ein berechenbares Ähnlichkeitsmaß zu haben. Hier hilft die Perspektive auf „big data", man weiß ja, dass die großen Internet-Unternehmen gigantische Datenmengen bewerten müssen. Noch größer dürfte die Motivation sein, wenn im Informatikunterricht solche Auswertungen auch programmiert werden.

Wie könnte man mit einem Zahlenwert erfassen, wie ähnlich Anna und Berta einander in ihren freien Nennungen in Beispiel 1 sind? Eine naheliegende Idee ist, die Anzahl der Übereinstimmungen zu nehmen (das wären zwei). Wenn es nun aber eine weitere Schüler*in gäbe, die extrem viele Sportarten aufgeschrieben hätte, ergäbe sich automatisch eine hohe Überschneidungszahl, d. h. es ist sinnvoll, relativ zur Gesamtzahl der genannten Sportarten zu rechnen:

$$\text{Ähnlichkeit Anna, Berta} = \frac{2}{5+5-2} = \frac{2}{8} = 0{,}25$$

Diese Zahl nennt man den Tanimoto-Koeffizienten. Er misst die Ähnlichkeit zweier Mengen M_i, M_j gemäß der Formel: $T = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}, 0 \leq T \leq 1$.

Damit lässt sich eine Ähnlichkeitsmatrix aufstellen:

	Anna	Berta	Clara	Dora	Elena
Anna	1	0,25	0,67	0,00	0,11
Berta	0,25	1	0,25	0,50	0,43
Clara	0,67	0,25	1	0,00	0,11
Dora	0,00	0,50	0,00	1	0,50
Elena	0,11	0,43	0,11	0,50	1

Eine offene Fragestellung ist, wie man vorgeht, wenn neben den Objekten selbst auch eine Häufigkeit erfasst wird (z. B. „Nenne Sportarten und gib an, wie oft du sie im letzten Jahr betrieben hast"). Es gibt naheliegende Antworten, die hier aber nicht verraten werden sollen. Es ist gerade eine schöne Eigenschaft dieses Themas, dass es bei vielen Fragen nicht die eindeutig richtige Lösung gibt. Stattdessen kann man sich verhältnismäßig leicht Dinge ausdenken, und letztlich entscheidet die Praxis, was am erfolgreichsten ist.

Eine analoge Auswertung im zweiten Fall (Beispiel 2) erfordert ein Ähnlichkeitsmaß für Vektoren $u, v \in \mathbb{R}^k$. Dafür gibt es viele Kandidaten. Eine Idee ist, den euklidischen Abstand der Vektoren zu nehmen: $d_1(u, v) := |u - v|$. Das ist in vielen Anwendungen gut, aber man denke an folgende Konstellation: Zwei Schüler*innen bewerten vorgelegte Sportarten auf einer Skala von 0 bis 20 mit den Werten (9,15,0,12) und (12,20,0,16). Dann haben sie die gleichen Vorlieben, aber ihr Abstand erscheint doch groß. Deswegen kann es besser sein, die Vektoren erst zu normieren:

$d_2(u, v) := \left| \frac{u}{|u|} - \frac{v}{|v|} \right|$.

Stellt man sich die Datenvektoren als geometrische Vektoren in einem hochdimensionalen Raum vor, könnte Ähnlichkeit bedeuten, dass die Vektoren in die gleiche Richtung zeigen. Aus dieser Idee folgt die Messung der Ähnlichkeit über das Skalarprodukt: $d_3(u, v) := -u \cdot v$. Das negative Vorzeichen polt den Wert so, dass wie bei d_1, d_2 kleine Werte für hohe Ähnlichkeit stehen. Auch hier gibt es aus den gleichen Gründen die Idee, die Vektoren erst zu normieren:

$d_4(u, v) := -\frac{u}{|u|} \cdot \frac{v}{|v|}$.

Das Beispiel mit den Vektoren (9,15,0,12) und (12,20,0,16) zeigte, dass die absolute Länge u. U. nicht relevant ist. Bei Likert-Skalen zeigen einige Menschen die Tendenz, höhere Werte anzukreuzen als andere, obwohl sie aber vermutlich etwas Ähnliches meinen. Das Problem kann als ein Faktor auftreten, der wie oben durch Division eliminiert wird, es kann aber auch als Summand auftreten. Es kann daher ratsam sein, die Einträge des Vektors so zu verschieben, dass der Mittelwert 0 entsteht. Dazu definiert man für $u \in \mathbb{R}^k$ den Mittelwert als $\overline{u} := (u_1 + \ldots + u_k)/k$ und den zum Mittelwert 0 verschobenen Vektor als $N(u) := (u_1 - \overline{u}, \ldots, u_k - \overline{u})$. Modifiziert man d_4 entsprechend, erhält man $d_5(u, v) := d_4(N(u), N(v))$. Dieses Maß nennt man (negativen) Korrelationskoeffizienten.

Jetzt stehen fünf Maße zur Verfügung und für jedes dieser Maße kann man eine entsprechende Ähnlichkeitsmatrix wie oben ausrechnen. Welches Ähnlichkeitsmaß in der Praxis verwendet werden sollte, wird am besten auf Basis von Erfahrungen beantwortet wird. Man kann aber aufgrund der mathematischen Eigenschaften der definierten Maße ein paar Schlüsse ziehen. Segaran (2011) empfiehlt alle Programme so zu schreiben, dass man das Ähnlichkeitsmaß leicht austauschen kann. In vielen Fällen aber, z. B. bei der Klassifikation von News-Blogs, bevorzugt er aus seiner Erfahrung die Korrelation.

Exkurs Es gibt noch viele weitere wichtige Ähnlichkeitsmaße, je nachdem welche Art von Daten vermessen werden sollen. Die Edit-Distanz bestimmt die Ähnlichkeit von Wörtern: Wie viele elementare Vorgänge (ein Zeichen löschen, eines einfügen, zwei aufeinanderfolgende Zeichen vertauschen) muss man mindestens durchführen, um vom einen Wort zum anderen zu kommen? Das ist nicht nur für die Rechtschreibkorrektur wichtig zu wissen (welches Wort könnte jemand gemeint haben), sondern auch für die Molekularbiologie, wo man die Ähnlichkeit von Lebewesen

durch die Ähnlichkeit ihrer DNS-Sequenzen ermitteln kann. Der Hamming-Abstand ist ein weiteres Ähnlichkeitsmaß für diesen Zweck, das vor allem dann geeignet ist, wenn Verschiebungen unwahrscheinlich sind.

Eine etwas exotische und nicht sehr gut, aber immerhin halbwegs funktionierende Anwendung von Ähnlichkeitsmaßen ist die Folgende: Wie wird das Wetter morgen? Dazu besorgt man sich die Wetterdaten für jeden Tag des Jahres der letzten Jahrzehnte. Dann gibt man die heutigen Daten ein und sucht einen Tag in der Vergangenheit, an dem das Wetter möglichst ähnlich war wie heute und schaut, wie es damals am Folgetag war. Das funktioniert leidlich gut. Professionelle Wettervorhersagen werden nicht so gemacht – und man kann gleich lernen, dass big data nicht alle Probleme löst.

Die didaktischen Bemerkungen am Ende des letzten Abschnitts lassen sich hier sinngemäß wiederholen.

4 Daten – Ähnlichkeit nutzen

Die Details der elementaren Mathematik der Vektoren kann man weitgehend vergessen in ihren Anwendungen: Ab dieser Stelle braucht man nur noch irgendeine Distanzfunktion. Das genaue Wissen darüber, wie diese definiert ist, kann in einer Blackbox verschwinden.

Der Stand ist jetzt also, dass es eine Menge O von Objekten (z. B. Menschen) gibt, deren „Abstand“ berechnet werden kann. Was lässt sich damit machen? Zunächst kann man durch Mittelwertbildung auch Mengen solcher Objekte einen Abstand zuordnen:

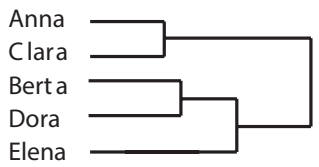
$$A, B \subset O, d(A,B) := \frac{\sum_{a\in A}\sum_{b\in B} d(a,b)}{|A| \cdot |B|}$$

Allerdings gibt es sehr viele Teilmengen (ihre Zahl steigt ja exponentiell mit der Zahl der Objekte). Man braucht also eine Strategie, die Ähnlichkeitsstruktur mit geeigneten Teilmengen darzustellen. Das gelingt durch Clusterbildung.

Eine solche Ähnlichkeitsstruktur (Cluster) kann ganz einfach rekursiv definiert werden: Man bildet im ersten Schritt die Menge aller einelementigen Mengen von Objekten: $C_1 := \{x | x \in O\}$ Dann sucht man die beiden Teilmengen mit dem geringsten Abstand und verbindet sie zu einer zweielementigen Menge, sodass die neue Menge C_2 aus einer zweielementigen und vielen einelementigen Mengen besteht. Nun sucht man wieder die beiden Teilmengen, die sich am ähnlichsten sind und verbindet sie. So fährt man fort bis man letztlich nur noch eine Menge hat. Angewendet auf die Sportangaben aus Beispiel 1 ergibt sich folgende Folge von Mengen:

1. {{Anna}, {Berta}, {Clara}, {Dora}, {Elena}}
2. {{Anna,Clara}, {Berta}, {Dora}, {Elena}}
3. {{Anna,Clara}, {Berta,Dora}, {Elena}}
4. {{Anna,Clara}, {Berta,Dora,Elena}}
5. {{Anna,Clara,Berta,Dora,Elena}}

Eine schöne Darstellung ist ein Dendogramm, bei dem die fortschreitende Vereinigung dargestellt wird.



Auf Basis dieser Erkenntnisse kann man jetzt folgende Empfehlungen geben:

- Anna und Clara könnten gut zusammen Sport machen, ebenso Berta und Dora
- Das beste Dreierteam ist Berta, Dora, Elena
- Man könne Anna Volleyball vorschlagen, denn die ihr ähnliche Clara mag das.

Der Transfer von dieser Situation auf das Empfehlen von Produkten oder Beiträgen in Journalen oder Musik ist nicht schwierig.

Neben dieser Strategie der Clusterung gibt es noch viele andere. Beim k-means-Algorithmus gibt man die Zahl k der Cluster vor (siehe (Oldenburg 2011) für eine einfache Implementierung). Das kann man natürlich für jedes $k = 2, ..., n-1$ machen und sich ein Maß überlegen, das angibt, wie gut eine Clusterung zu den Daten passt. Auch da gibt es viele Möglichkeiten und in der Praxis zählt Fantasie und Erfahrung.

Clusterbildung ist auch die wissenschaftliche Grundlage für das Schubladendenken: Einige empirische Studien in den Sozialwissenschaften funktionieren so, dass eine Gruppe von Menschen auf Basis bestimmter Merkmale (z. B. Fragebogenantworten) in einige wenige Cluster-Schubladen gesteckt werden, deren Gemeinsamkeiten dann von Menschen (das leisten die Algorithmen nicht) in „Typen“ gefasst werden.

Ein zweites Beispiel soll diesen Punkt noch erweitern. Oben wurde bereits darauf hingewiesen, dass man beim Deutschen Wetterdienst für sehr viele Stationen Tag-genaue Tabellen der Wetterverhältnisse der letzten Jahrzehnte bekommen kann. Ein sehr naiver Ansatz zur Wettervorhersage ist, in dieser Liste der alten Daten einen Tag zu suchen, bei dem das Wetter so ähnlich war wie heute, und zu vermuten, dass dann der morgige Tag so ähnlich wird wie der damals folgende Tag. Aus der Fülle der Wetterdaten, die der Wetterdienst notiert, sucht man sich dabei ein paar aussagekräftige heraus, die man auch selbst für den heutigen Tag leicht bestimmt kann, z. B. Minimal- und Maximaltemperatur m, M, ungefähre Windgeschwindigkeit w in m/s,

Sonnenscheindauer s und Niederschlagsmenge n in mm und Luftdruck p in mbar. Bei der Beurteilung der Ähnlichkeit zweier daraus gebildeter Vektoren $(m, M, w, s, n, p) \in \mathbb{R}^6$ wird man auf eine weitere wichtige Technik gestoßen: Normierung. Es ist nämlich ziemlich schnell offenbar, dass ein Unterschied zwischen einer Sonnenscheindauer von 0 und 10 h erheblich ist, ein Unterschied im Luftdruck von 1000 mbar zu 1010 mbar aber nicht ganz so wichtig ist. Daher normiert man die Daten zuerst, d. h. man bildet sie z. B. linear auf das Intervall [0,1] ab. Dabei ermittelt man die minimalen und maximalen Werte für jede Größe, z. B. $p_{\min}, p_{\max}$ und berechnet den normierten Luftdruck als $p' := \frac{p - p_{\min}}{p_{\max} - p_{\min}}$. Mit den so normierten Vektoren $\left(m', M', w', s', n', p'\right)$ gibt dann der Euklidische Abstand ein brauchbares Ähnlichkeitsmaß ab. Auch hierbei zeigt das Vorgehen eine Reihe von denkbaren Alternativen auf: Es gibt viele andere Normierungsmethoden und die Frage, welches Ähnlichkeitsmaß in der konkreten Anwendung das Beste ist, lässt sich nicht so einfach beantworten. Das Beispiel zeigt aber exemplarisch, wie mathematische Reflexion über die Eigenschaft von Größen zu verbesserten Verfahren führt. In der konkreten Situation der Wettervorhersage sind die Ergebnisse allerdings nicht sehr gut. Wie Wettervorhersagen tatsächlich berechnet werden, ist vom Standpunkt der Schulmathematik aus leider viel zu schwierig. Die Dynamik der Atmosphäre wird durch gekoppelte partielle Differenzialgleichungssysteme beschrieben.

5 Daten – Ausblick

Die oben dargestellten Themen sind nur ein kleiner Teil der mathematischen Modellierung im Bereich der angewandten Datenwissenschaften. Weitere verbreitete Methoden sind Neuronale Netze, Klassifikatoren (Entscheidungsbäume, bayesches Klassifizieren, Support-vector-machines) und Regressionsrechnung mit all ihren Variationen (z. B. Faktorenanalyse). Aber ein erster Einblick ist, wie gezeigt, relativ elementar möglich und kann hoffentlich die Fantasie anregen, was man alles machen könnte – und ob man es sollte!

Literatur

Meyer-Schönberger, V., Cukier, K.: Big data. John Murray, London (2013)

O'Neil, C.: Weapons of Math Destruction. Crown Publisher, New York (2016)

Oldenburg, R.: Mathematische Algorithmen für den Unterricht. Vieweg, Wiesbaden (2011)

Segaran, T.: Kollektive Intelligenz. O'Reilly, Köln (2011)