FACHHOCHSCHULE
ERFURT UNIVERSITY
OF APPLIED SCIENCES
Angewandte
Informatik

Masterarbeit
in der Angewandten Informatik
Nr. AI-2021-MA-014

Generisches Clustern hoch-komplexer Produktdaten

Hannes Dröse

Abgabedatum: 29.04.2022

Prof. Dr. Ines-Kerstin Rossak
Dipl.-Inform. Robert Queck

Zusammenfassung

Die Kurzfassung wird zumeist als letzter Abschnitt geschrieben. Sie soll auf einer Seite den Hintergrund bzw. die Motivation sowie die zentralen Ergebnisse der Arbeit zusammenfassen. Der Zweck dieses Texts ist es, einem Recherchierenden oder Informationssuchenden zu einem bestimmten Thema zu signalisieren, ob er in diesem Werk für ihn relevante Informationen finden wird: Lohnt es sich, die Arbeit zu bestellen oder zu kaufen? Was ist neu?
Die Kurzfassung ist das sichtbare Aushängeschild der Arbeit und dient der Aufnahme in Referenzdatenbanken, in Online-Literatur-Shops u. Ä. Schauen Sie sich die Abstracts in der Fachliteratur (etwa in Fachbüchern, am Anfang von Zeitschriftenartikeln, in Konferenz- Proceedings, in ACM- oder IEEE-Literaturdatenbanken, im OPAC der Bibliothek usw.) an, um ein Gefühl dafür zu entwickeln, was hineingehört (siehe Abschnitt 2.4 Materialsammlung, S. 2).
Was definitiv nicht hineingehört, sind Gliederungsübersichten, chronologische Arbeitsberichte u. dgl.

Abstract

The abstract has to be provided in english as well.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Hintergrund

Diese Masterarbeit ist in Kooperation mit der adesso SE entstanden. adesso ist ein deutsches IT-Beratungs- und Dienstleistungsunternehmen. Sie ist 1997 gegründet worden mit Hauptsitz in Dortmund. Seit der Gründung ist die Firma sehr kontinuierlich gewachsen. Mittlerweile sind über 5800 Mitarbeiter an 44 Standort in 10 verschiedenen europäischen Ländern hier beschäftigt.

Über die Jahre sind nicht nur die Mitarbeiterzahlen gewachsen, sondern auch die abgedeckten Branchen, in denen das Unternehmen tätig ist. In den letzten Jahren sind zunehmend Projekte im E-Commerce-Sektor umgesetzt worden. Mit Beginn von 2022 kümmert sich eine Abteilung des Unternehmens explizit um das Thema E-Commerce und Retail.

In diesen Bereich fällt die Programmierung und Betreuung von Online-Shops sowie das umfangreiche Thema des Product-Information-Managements (PIM). Dabei geht es um das Verwalten und Aufbereiten von Produktdaten für verschiedene Anwendungen wie Warenhaltung, Marketing oder Bestellabwicklung. adesso erwägt in Zukunft ein eigenes Product-Information-Management-System (PIM-System) zu entwickeln. In Vorbereitung dessen steht die Evaluierung der Machbarkeit verschiedener Ansätze nud Ideen rund um das Product-Information-Management an. Dies ist auch der Aufhänger für diese Masterarbeit.

1.2 Themenfindung

Die initiale Idee bestand darin zu überprüfen, ob es mögliche Anwendungen des Verfahrens der Clusteranalyse im Zusammenhang mit dem Product-Information-Management gibt. Beim Clustering handelt es sich um die Einteilung von Objekten in Gruppen (sog. Cluster), sodass sich die Objekte in der gleichen Gruppe ähnlicher sind als Objekte in den anderen Gruppen. [King15, Kap. 1.1 What Is a Cluster?]

In einem ersten Brainstorming sind verschiedene Anwendungen der Clusteranalyse diskutiert worden. Es folgte eine intensive Literaturrecherche zu den besprochenen Anwendungen, die im Folgenden kurz diskutiert werden.

Die erste Idee war das automatische Kategorisieren von Produkten. Zu dieser Anwendung gibt es kaum Literatur – lediglich einige hoch-experimentelle Ansätze. Z.B. nutzten Chen et al. [ChZhYi19] ein neuronales Netz zum Vorschlagen von Produktkategorien zu Produkten bestehend aus Titel und Beschreibung.

Eine weitere Idee war die Nutzung der Clusteranalyse zur Entwicklung eines Produktempfehlungsalgorithmus. Zu diesem Thema gibt es ganze Reihe an Arbeiten:

- Oh und Kim [OhKi19] implementierten einen Empfehlungsalgorithmus für Cloud Computing-Angebote. Die verschiedenen Anbieter wurden nach den Anforderung der auszuführenden Applikation geclustert, um so das am besten passende Angebot zu finden.
- Cui [Cui21] verbesserte die Klick- und Kaufraten in einem Online-Shop durch die Kombination von Assoziationsregeln mit Fuzzy Clustering. Dadurch konnten passende Warenkörbe schneller identifiziert und vorgeschlagen werden.
- Zuletzt sei die Arbeit von Kumar et al. erwähnt [KRRT01]. Sie näherten sich mit einem möglichst allgemeinen mathematischem Modell an das Thema. Auf dieser Basis identifizierten sie Faktoren, die ein Empfehlungsalgorithmus erfüllen sollte. Werden diese befolgt, so können bereits mit wenigen Daten über das Nutzerverhalten effektive Empfehlungen gegeben werden. Das Clustern der Produkte beschleunigt diesen Prozess, da Erkenntnisse zu einem Produkt auf die gesamte Gruppe übertragbar sind.

Die letzte Idee beschäftigte sich mit der Suche von Produkten. Ein sehr spannendes Paper hierzu verfassten Kou und Lou [KoLo12]. Sie verbesserten die Klickraten einer Websuchmaschine durch die Nutzung von Clustering. Ihr Algorithmus nutzte die am meisten geklickten Webseiten zu jedem Suchbegriff als initiale Schwerpunkte für die Cluster. Anschließend sind die übrigen Webseiten rund um diese Schwerpunkte geclustert worden. Dadurch konnten neu hinzugekommene Seiten direkt in die Suche integriert werden. Außerdem erhöhte sich die Relevanz der gefundenen Suchergebnisse.

Davon abgesehen gibt es für diesen Bereich kaum Arbeiten in denen Suchfunktionen mit Clustering kombiniert worden sind. Das liegt sicherlich daran, dass klassische probabilistische Suchalgorithmen (der bekannteste ist der BM25 [RoZa09]) sehr gute Ergebnisse liefern. Ansätze wie von Kou und Lou sind in dem Zusammenhang eher als experimentell anzusehen.

1.3 Kernfrage und Ablauf

Parallel zu der beschriebenen Recherche ist ebenfalls ein Blick in typische PIM-System wie z.B. Akeneo-PIM geworfen worden. Ziel war es zu verstehen, wie die Produktdaten in solchen System typischerweise vorliegen. Dies ist nötig, um einzuschätzen, ob und wie das Clustering der Produktdaten in solchen Systemen möglich ist.

Zum einen weisen die Produktdaten einen hohen Grad an Struktur auf. Die Attribute sind fest definiert. Umfangreiche Constraints sorgen für die Einhaltung des definierten Schemas. Zum anderen kommen sehr verschiedenartige Datentypen vor wie z.B. Textfelder, Einfach- und Mehrfachauswahl, numerische Daten mit verschiedenen Einheiten. Zudem sind viele Attribute lokalisierbar und weisen je nach Sprache andere Werte auf. Die klassische Clusteranalyse arbeitet allerdings nur mit numerischen Daten. Somit war abzusehen, dass das Clustern selbst kein einfacher Prozess sein wird.

Dadurch verschob sich der Fokus der Arbeit. Bevor mögliche Anwendungen der Clusteranalyse evaluiert werden, stellt sich zuerst die Frage, ob so vielfältige Produktdaten überhaupt geclustert werden können. Die Kernfrage der Arbeit lautet also:

Ist das effektive Clustern hoch-komplexer Produktdaten möglich und wenn ja, wie?

Diese Fragestellung war der Ausgangspunkt weiterer Recherchen, Experimente und Versuche. Sie werden in den folgenden Kapiteln erläutert.

Kapitel 2 gibt einen Überblick über die wissenschaftlichen Grundlagen und Erkenntnisse des Themenfeldes.

Auf dieser Basis wird in *Kapitel 3* eine Konzeption dargelegt, wie das Clustern durchgeführt werden kann. Ebenso werden Versuche geplant, die zur Überprüfung des Konzeptes dienen.

Kapitel 4 beschreibt die Umsetzung des dargelegten Konzeptes. Thema dabei sind die Implementierung nötiger Software-Komponenten, Tools und Techniken, die zur Überprüfung der Fragestellung angewendet worden sind.

Die Auswertung und Evaluation der Versuche erfolgt in *Kapitel 5*.

Im finalen *Kapitel 6* wird das Fazit gezogen, ob die Kernfrage beantwortet werden kann. Ebenso wird ein Ausblick für weitere Versuche und Fragestellungen in der Zukunft gegeben.

2 Clusteranalyse

2.1 Begriff und Einordnung

King definiert die *Clusteranalyse* als die "[...] Generierung eines Klassifizierungsschemas, welches Individuen in eine feste Anzahl an Gruppen einteilt, so dass sich die Individuen innerhalb einer Gruppe auf eine Art und Weise ähnlich sind und unähnlich denen in anderen Gruppen" [King15, Kap. 1.1 What Is a Cluster?]. Diese Gruppen werden als auch Cluster bezeichnet.

Dieser Prozess des Clustering ist eine Methode des *unüberwachten Lernens (unsupervised learning)* – einem Teilgebiet des maschinellen Lernens. Papp et al. schreiben dazu: "Machine Learning beschäftigt sich mit der Entwicklung von Methoden, die Modelle von Zusammenhängen aus Daten erlernen, anstatt sie *per Hand* zu implementieren" [PWMO19, Kap. 5 Statistik-Grundlagen]. Ferner geben sie an, dass die Unterschiede zur Statistik fließend sind [PWMO19, Kap. 5 Statistik-Grundlagen]. Unüberwachtes Lernen bedeutet dabei, dass die verwendeten Daten nicht im Vorhinein gekennzeichnet sind (Unlabeled Data). Stattdessen werden Ähnlichkeiten und Muster in den Daten selbst gesucht ohne eine vorgegebene Zielgröße. Häufig dienen diese Analysen als erste Schritte der *Data Exploration* aus denen im Anschluss neue Erkenntnisse und Anwendungen abgeleitet werden. [PWMO19, Kap. 5.2.3 Unüberwachtes Lernen]

Allgemein wird die Clusteranalyse als eine Form des Data Minings gesehen. Laut Bissantz und Hagedorn "beschreibt [Data Mining] die Extraktion implizit vorhandenen, nicht trivialen und nützlichen Wissens aus großen, dynamischen, relativ komplex strukturierten Datenbeständen" [BiHa09]. Mit "Wissen" meinen sie dabei in den Daten implizit enthaltene Muster, welche für den Anwender interessant sind und mit einer hinreichenden Wahrscheinlichkeit tatsächlich in den Daten existieren. [BiHa09]

2.2 Notation – TODO

Datenpunkte und -set

- Objekte oder Datenpunkte sind Vektoren aus numerischen und/oder kategorischen Attributen: $x = (a|a \text{ is numerical or categorical attribute})$
- Menge aus Datenpunkten x (= Datenset) mit Großbuchstaben angegeben $x \in X$
- i.d.R. $n = |X|$

Cluster

- Cluster sind spezielle Teilgruppen aus dem Datenset $C \subset X$
- Cluster können Datenpunkten oder andere Cluster enthalten: $C = e|e = C_i \vee x \in X$

- leere Cluster gibt es nicht $C \neq \emptyset$
- die Menge aller Cluster ist $K = C_1, C_2, ...$ und $k = |K|$
- Cluster haben häufig einen Mittelpunkt $c = centroid(C)$

Cluster-Zuordnungen

- Vektor aus Labels für jeden Datenpunkt in X: $Y = (l|l_i = \text{ label for } x_i \in X)$

2.3 Distanz- und Ähnlichkeitsmaße – TODO

2.3.1 Definition

Clustering erfolgt über Bestimmung der "Nähe" (engl. proximity) der Objekte zueinander. [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

Bestimmung der Nähe mittels Abstands- bzw. Distanzmaßen. Distanzmaß wird über Funktion d dargestellt [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

Tabelle 2.1: Eigenschaften der Abstandfunktion d

1.	$d(x_1, x_2) \geq 0$	Distanzen sind stets positiv
2.	$d(x_1, x_1) = 0$	zwei gleiche Objekte haben immer einen Abstand von 0
3.	$d(x_1, x_2) = d(x_2, x_1)$	die Distanzfunktion ist kommutativ bzw. symmetrisch
4.	$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$	Distanzen geben stets den kürzesten Weg an

=> [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

Statt Distanzmaße auch Verwendung von Ähnlichkeitsmaßen $s(x, y)$ (engl. similarity) möglich. Ähnlichkeit i.d.R. im Interval $[0, 1]$ angegeben, wobei $s(x, x) = 1$. Wenn Distanz z.B. durch Normalisierung ebenfalls im Interval $[0, 1]$ liegt, dann gilt: $d(x, y) = 1 - s(x, y)$ [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

Distanz und Ähnlichkeit dadurch beliebig austauschbar, deshalb diese Verwendung empfehlenswert.

(Frage: Erwähnung der Proximity Matrix? ist eigentlich nicht wichtig für diese Arbeit)

2.3.2 Numerische Attribute

durch (rationale) Zahlen dargestellt [KaRo09, Kap. 1.2 Types of Data and How to Handle Them], mit stetigen (engl. continuous) Werten [Huan98]. Umfasst damit sowohl Daten in Intervall- und Verhältnisskalen (engl. interval and ratio data) [Bosl12, Kap. 1 Basic Concepts of Measurement]

Minkowski-Familie

Datenpunkte damit numerischen Vektoren also Punkte. Bestimmung des Abstand der Punkte => verschiedene Maße, gehören alle zur Minkowski-Familie [King15]:

Tabelle 2.2: Übersicht gängiger Maße aus der Minkowski-Familie

Name	p-Norm	Formell
Minkowski	allgemein	$d(x,y) = \sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$
Manhattan	$p = 1$	$d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
euklidisch	$p = 2$	$d(x,y) = \sqrt{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^2}$
Chebyshev	$p = \infty$	$d(x,y) = \max \lvert x_i - y_i \rvert$

höhere p-Norm bedeutet i.d.R. robustere Bestimmung des Abstands. Manhattan => Winkel zwischen zwei Punkten, euklidisch => Länger der Geraden durch die Punkte [King15, Kap. 12.3 Which Proximity Measure Should Be Used?]

verschiedenste Versionen und Abwandlungen dieser Maße [siehe Cha07]

- aber die meisten basieren auf Minkowski-Familie
- oder lassen sich auf Minkowski abbilden (z.B. Maße basierend auf der Pearson Korrelation nutzen im Kern euklidischen Abstand zur Berechnung)
- liefern ähnliche Ergebnisse wie klassische Minkowski
- [siehe Cha07] Clustering mit verschiedenen Maßen und dann Ergebnisse geclustert. (eigene Erkenntnis aus dem Paper) im Kern jeder Cluster Gruppe war auch ein Vertreter der Minkowski-Familie

Normalisierung

manche numerischen haben logarithmischen Zusammenhang (10 zu 20 ist gleichbedeutend mit 100 zu 200), dann $x'_i = \log x_i$ für gleichmäßige Abstände

Projektion auf $[0, 1]$ => $x' = \frac{x - \text{avg } X}{\max X - \min X}$; empfehlenswert, wenn Attribute gleichgewichtet sein sollen

manchmal ist Normalisierung nicht sinnvoll, da dadurch verschiedene “Gewichtung” der Attribute bestimmtes Wissen dargestellt wird => erfordert Domänenwissen

Eventuell sogar arbeiten mit Gewichtsvektor für die Attribute => eher spezielle Anwendungsfälle

gesamter Abschnitt [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

2.3.3 Kategorische Attribute

statt Zahlen bestehen Datenpunkte aus ihnen zugeordneten Kategorien oder Labels [Bosl12, Kap. 5 Categorical Data]

Manchmal gibt es eine sinnvolle Reihenfolge der Labels (z.B. xs, s, m, l, xl), dennoch keine Aussage über Verhältnis oder Intervall zueinander möglich. mit Reihenfolge als ordinal bezeichnet, ohne als nominal [Bosl12, Kap. 1 Basic Concepts of Measurement, Kap. 5 Categorical Data]

Ordinale Attribute

zwei Möglichkeiten:

Reihenfolge ignorieren und als nominal verarbeiten => siehe folgende Abschnitte

aber eher empfohlen: Umwandlung und Verarbeitung als numerische Attribute:

- Nummerierung der Labels nach ihrer Reihenfolge von 1 bis n
- dann: $x' = \frac{x-1}{n-1}$
- dadurch im Interval $[0, 1]$ in $n - 1$ gleichmäßige Abschnitte eingeteilt

ganzer Abschnitt [KaRo09, Kap. 1.2 Types of Data and How to Handle Them]

Nominale Attribute

dieser Teil aus [KaRo09, Kap. 1.2.5 Nominal, Ordinal, and Ratio Variables]

Umwandlung in sog. binäre Attribute:

1. bei 2 Ausprägungen => 0 = erste Ausprägung, 1 = zweite Ausprägung z.B. yes/no
2. ab 3 =>

- entweder in 2 Ausprägungen komprimieren
- oder ein neues Attribut pro Kategorie definieren: 0 = gehört nicht zur Kategorie, 1 = gehört zur Kategorie

folgende aus [KaRo09, Kap. 1.2.5 Binary Variables]

für 1. => sog. “symmetrische” binäre Attribute, 2. => “asymmetrische”; Verarbeitung mit speziellen Ähnlichkeitsmaßen:

Simple matching $s(x, y) = \frac{|x \cap y| + |\bar{x} \cap \bar{y}|}{|x \cup \bar{x} \cup y \cup \bar{y}|}$

- jeder Match (beide Datenpunkte haben Label und beide Punkte haben ein Label **nicht**) wird gezählt
- nur für symmetrische geeignet
- gibt Varianten davon mit unterschiedlichen Gewichten für Matches usw. => lassen sich alle auf Simple matching abbilden

Jaccard-Koeffizent $s(x,y) = \frac{|x \cap y|}{|x \cup y|}$

- nur die tatsächlich vorhandenen Attribute (mit 1 codiert) werden miteinander verglichen
- nicht-vorhandene Attribute bei beiden Datenpunkten werden ignoriert
- gleiches Bsp:
  - 3 Farben (rot, grün, blau)
  - x ist rot; y ist blau
  - Simple matching würde $\frac{1}{3}$ ergeben, weil ja beide gemeinsam haben **nicht** grün zu sein
  - Jaccard hingegen sagt Ähnlichkeit $\frac{0}{2}$, weil keines von zwei vorhandenen Attributen matched

2.3.4 String-Attribute

freie Textfelder

Tokenization

!!!QUELLEN!!! [CRF03]

Bei nominalen Attributen mit vielen Ausprägungen (bsp Produkttitel => jeder ist anders)

- Zerlegen in einzelne Wörter => Tokenization
- Rückführung der Wörter auf ihren Stamm (z.B. Porter-Stemming)
- Entfernen von Füllwörtern => Stop-Word Removal

Am Ende erhalten wir eine feste Menge an Tokens oder Keywords => Menge kategorischen Daten => asymmetrisch binär verarbeiten

oder auch Überführung in Vektor-Space ...

String-Metrics (eventuell weglassen??)

!!!QUELLEN!!! [CRF03] [RaRa11]

Alternativ kann Abstand auch mittels String-Metrics ermittelt werden.

String-Metrics messen die Ähnlichkeit zweier Strings

Bsp:

- Levenshtein: Anzahl an Buchstaben eingefügt/geändert/gelöscht um x auf y umzuwandeln
- Hamming: Anzahl an **ungleichen** Buchstaben => beide Strings müssen gleich lang sein
- Jaro: Verhältnis gleicher zu verschiedenen Buchstaben, performant mit guten Ergebnissen

- Jaro-Winkler: Erweiterung von Jaro mit höherem Gewicht auf den Anfang (Prefix) der Wörter

Verwendung:

- entweder String-Metrics selbst als Distanzmaß nutzen
- oder "ordinale" Reihenfolge mittels String-Metrics bestimmen und dann als ordinal betrachten

2.3.5 Gemischte Attribute

[KaRo09, Kap. 1.2.6 Mixed Variables]

verschiedene Ansätze:

Umwandlung in numerische Werte

- siehe ordinal
- siehe nominal => String-Metrics

Umwandlung in kategorische Werte

- ordinal als nominal betrachten (siehe Ordinale Attribute)
- numerisch Umwandeln durch Diskretisierung
  - Bildung von festen Bändern/Gruppen z.B. Alter => $10-19$, $20-29$ etc.

Separates Clustern je Attribut-Klasse und Subspace-Clustering

mehrmaliges Clustern mit jeweils nur Attributen eines Skalenniveaus

anschließend Vergleich der verschiedenen gebildeten Cluster

noch weitere komplexere Ansätze, wo verschiedenste Kombinationen an Attributen für das Clustern ausgewählt werden und verglichen werden => "Subspace-Clustering" [siehe JiCh17]

Kombinierte Distanzfunktion

Bewertung jedes Attributs mit geeignetem Distanzmaß. Anschließend zusammenrechnen mit Gleichgewichtung

ab hier aus [Huan98]

z.B. bei k-Prototype werden numerische und kategorische mit jeweils geeigneten Distanzmaßen verrechnet:

$d(x,y) = wd_numerical(x,y) + (1-w)d_categorical(x,y)$, also z.B. euklidisch und simple matching oder Manhattan und Jaccard etc.

w soll eine Gleichgewichtung erzeugen also z.B. $\frac{Anzahl numerischer Attribute}{Anzahl aller Attribute}$

2.4 Clustering-Verfahren – TODO

2.4.1 Partitionierendes Clustering

Überblick

Minimierungsproblem: initiale Cluster-Zuordnungen (Partitionen), dann Veränderungen der Cluster-Zuordnungen bis ein lokales Minimum gefunden worden ist. [King15, Kap. 4.1 Introduction]

Verschiedene Algorithmen und Varianten: initiale Selection der Cluster, Wahl der Distanzmaße, Vorgehen während der Minimierung [King15, Kap. 4.1 Introduction] und für welche Arten von Attribute geeignet, dazu später mehr [siehe Huan98]

Effizienz:

- neigt zu lokalen Minima => mehrmaliges Wiederholen des Clustering mit verschiedenen zufälligen Startpunkten
- dennoch sehr effizient lösbar: $\mathcal{O}(n \cdot k \cdot l)$
- n Datenpunkte, k Cluster, l Wiederholungen des Clusterings
- k und l kleine Werte, viel kleiner als n
- also: $\mathcal{O}(n)$
- [Huan98]

Aber: k muss vorher bekannt sein => verschiedene k probieren oder z.B. mit hierarchischem Verfahren k abschätzen [King15, Kap. 4.1 Introduction]

k-Means

klassischster Vertreter des Clusterings und weit verbreitet und genutzt [Huan98]

Jedes Cluster wird über einen Schwerpunkt (Mittelpunkt) repräsentiert [StKaKu00]. Es gilt die Datenpunkte so den Clustern zuzuordnen, dass die Summe der Abstände der Datenpunkte zum Mittelpunkt so klein wie möglich sind. Schwerpunkt wird über Mittelwert (engl. mean) der Clustermitglieder bestimmt. [King15, Kap. 4.5 K-Means Algorithm]

Funktioniert daher nur mit numerischen Werten [Huan98]

Ablauf [King15, Kap. 4.5 K-Means Algorithm]:

1. Wahl der initialen k Startpunkte
2. Zuordnen aller Datenpunkte zum nächstliegenden Schwerpunkt
3. Neuberechnung der Schwerpunkte mittels Durchschnittswert (engl. mean) der Datenpunkte
4. ab 2. wiederholen, solange bis keine/kaum noch Änderungen der Schwerpunkte

Für Wahl der initialen Schwerpunkte [King15, Kap. 4.3 The Initial Partition]:

- ersten k Datenpunkte
- oder gleichmäßig aus der gesamten Liste
- oder zufällig über gesamte Liste
- und weitere

z.B. mehrmals mit verschiedenen random Startpunkten durchführen und bestes Ergebnis nehmen => verteilte Berechnung möglich

Statt die Schwerpunkte nur einmal am Ende eines Durchlaufes neuberechnen (Forgy's Method) auch permanente Neuberechnung (mit jedem neu zugeordnetem Datenpunkt) möglich (MacQueen's Method) und weitere Abwandlungen [King15, Kap. 4.5 K-Means Algorithm]

k-Medoids

[ArVa16]

- Variante des k-Means
- median statt mean als Clusterschwerpunkte
- median Berechnung komplexer, da Abstand aller Punkte eines Clusters zueinander berechnet werden muss
- aber laut Autor bessere Ergebnisse, schnellere Konvergenz und dadurch trotzdem kürzere Rechenzeit

k-Modes

[Huan98]

Variante des k-Means für kategorische (Attribute)

Änderungen zum k-Means:

- Simple matching als Distanzmaß (allerdings auch andere z.B. Jaccard möglich)
- statt Durchschnittswertes (engl. mean) wird Clusterschwerpunkt aus dem Modus (engl. mode) bestimmt
- Update der modes eines Clusters während des Clustering anhand ihrer Häufigkeit (Frequenz)

nach einem Durchlauf, Set erneut durchgehen und checken, ob einige Punkte nicht einem anderen Cluster zugeordnet werden müssen, da sich Modus des Datensets während des Clusterns ja häufig ändert. nach Reassign => Modus Frequenz in beiden Clustern anpassen

k-Prototypes

[Huan98]

Kombi aus k-means und k-modes für gemischte Datensets

nutzt ein Distanzmaß für numerische und eins für kategorische Werte + Gewicht zur gleichgewichtung der Attribute siehe gemischte Attribute

Schwerpunkt der numerische Attribute ist der Durchschnitt (mean) und der nominalen ist der Modus (mode) => also Kombi aus beiden

2.4.2 Hierarchisches Clustering

Überblick

!!!QUELLEN!!!

schrittweise Zuordnung zu Clustern => top-down (divisive) oder bottom-up (agglomerative)

Visualisierung: Dendogramm

Vorteil: k muss vorher nicht bekannt sein, sondern Hierarchie mit beliebiger Feinheit, Erkenntnisse aus Kind- und Vater-Clustern

Nachteile: einmal kombiniert/geteilt kein Reassignment der Cluster mehr, teilweise seltsame Ergebnisse, Fehler ziehen sich immer weiter durch

rechenintensiv: Vergleich von jedem Datenpunkt mit jedem anderen Datenpunkt, also mindestens $\mathcal{O}(n^2)$ meistens noch aufwendiger, da Datenset k mal durchsucht werden muss. Da hierarchisch ist $k \approx n$, also meistens $\mathcal{O}(n^2 \log n)$ (mit Optimierungen) bzw. $\mathcal{O}(n^3)$

Agglomeratives Clustering (Bottom-Up)

!!!QUELLEN!!!

Ablauf:

1. alle Datenpunkte in einem eigenen Cluster
2. Ermittlung der beiden am nächsten gelegenen Cluster
3. Fusion dieser beiden Cluster zu einem größeren
4. ab 2. wiederholen bis alle Punkte in einem Cluster sind

Theoretisch könnte auch vorher abgebrochen werden, aber für den Anwender sind i.d.R. die obersten "größten" Cluster-Hierarchien am interessantesten

zentrale Frage: wie werden zwei Cluster mit mehreren Datenpunkten miteinander verglichen => Linkage

Bsp immer die beiden dichtesten Datenpunkt beider Cluster => Single-Link

Name	Formell	Beschreibung	Eigenschaften
Single-Link	$d(X,Y) = \min d(x_i, y_j)$	Abstand durch die beiden dichtesten Punkte definiert	neigt zur Bildung von langen Ketten (chaining effect), anfällig gegen Outlier
Complete-Link	$d(X,Y) = \max d(x_i, y_j)$	Abstand durch die beiden am entferntesten liegenden Punkte definiert	neigt zur Bildung von vielen sehr kleinen Clustern, anfällig gegen Outlier

Name	Formell	Beschreibung	Eigenschaften
Average-Link	$d(X,Y) =$ avg $d(x_i, y_j)$	durchschnittlicher Abstand aller Punkte der beiden Cluster zueinander	ähnlich dem Mittelpunkt von k-Means, aber viel aufwendiger, da immerzu jeder Punkt eines Clusters mit jedem anderen Punkt der anderen Clustern verglichen werden muss
Mediod-Link	$d(X,Y) =$ ...	Median Punkt eines Cluster repräsentiert das Cluster	ähnliches Verhalten wie Average-Link, aber effizienter zu berechnen, da der Median eines neu gebildeten Clusters nur einmal bestimmt werden muss
k-Centroid-Link ???	$d(X,Y) =$ ...	vergleicht den Durchschnitt der k zentralsten Punkte der Cluster miteinander; allgemeine Form von Mediod-Link ($k = 1$) und Average-Link ($k = \|C\|$)	liegt je nach k irgendwo zwischen Average- und Mediod-Link; k ist ein weiterer Parameter, der festgelegt werden muss
Ward-Method	$d(X,Y) = \frac{d(\bar{x},\bar{y})^2}{\frac{1}{\|X\|}+\frac{1}{\|Y\|}}$	berechnet die Varianz für jedem potenziellen Merge, Merge mit höchster Reduktion der Varianz wird ausgewählt	sehr aufwendig zu berechnen, da immer wieder jeder Punkt mit jedem anderen verrechnet werden muss

Insgesamt alle sehr rechenintensiv. single- und complete-link bei guter Implementierung in $\mathcal{O}(n^2)$, alle anderen mindestens in $\mathcal{O}(n^2 \log n)$ oder $\mathcal{O}(n^3)$ (vor allem bei Average-Link und Ward-Method)

Diversives Clustering (Top-Down)

Überblick ergibt für Menschen intuitivere Cluster, da wir ebenfalls mental so vorgehen [King15, Kap. 3.3 Agglomerative versus Divisive Clustering]

theoretisch weniger aufwendig als agglomerativ, wenn nicht komplette Hierarchie generiert wird => weniger Durchläufe da i.d.R. $k < n - k$ (von 1 Cluster zu k weniger Schritte als von n zu k Clustern)

Problem: erstes Cluster riesig groß mit theoretisch $2^{n-1} - 1$ möglichen Arten des Splitterns (also $\mathcal{O}(2^n)$)

in der Praxis wird dieses Problem durch geschicktes Vorgehen umgangen:

DIANA [KaRo09, Kap. Divisive Analysis (Program DIANA)]

klassisches, erstes Verfahren, wie Abspaltung bei einer Partei

Ablauf:

1. Start mit allen Punkten in einem großen Cluster
2. Berechnung der durchschnittlichen Abweichung aller Punkte voneinander
3. Punkt mit größter Abweichung => Startpunkt einer "Splittergruppe"
4. Berechnung der durchschnittlichen Abweichung ohne Punkt mit größter Abweichung
5. Differenz der beiden Abweichungen größer geworden und am größten => Punkt der "Splittergruppe" zuordnen
6. wiederholen 4. bis Differenzen nur noch negativ
7. Durchmesser der Cluster berechnen $\emptyset = \max d(x, y)$
8. Cluster mit größtem Durchmesser verwenden und ab 2. wiederholen, bis gewünschte Abbruchbedingung erreicht ist

Abbruchbedingung kann bestimmter Durchmesser sein oder ein festes k, spätestens Schluss, wenn $k = n$

Laufzeit (eigene Überlegung; kleiner Kommentar dazu in [StKaKu00])

- Berechnung der durchschnittlichen Abweichung: jeder Punkt mit jedem anderen vergleichen => $\mathcal{O}(n^2)$
- da Cluster immer kleiner werden, wird Berechnung mit jedem Schritt einfacher

[RaRa11] Versuche mit agglomerativ und divisive nach DIANA => DIANA immer doppelt so schnell fertig für das gesamte Datenset

Bisecting k-Means [StKaKu00]

zur Teilung des größten Clusters wird k-Means mit $k = 2$ genutzt

- dadurch viel effizienter da k-Means in $\mathcal{O}(n)$
- k-Means wird zwar k mal aufgerufen, aber gleichzeitig wird die Anzahl der Datenpunkte je Cluster mit jeder Teilung kleiner

2.5 Cluster-Validität

2.5.1 Überblick

Die Clusteranalyse selbst ist ein Verfahren des unüberwachten Lernens und findet interne Muster in Daten ohne Referenz zu externen Zuweisungen (Labels). Dennoch finden verschiedene Clusteringverfahren unterschiedliche Gruppenzuteilungen. Auch die verschiedenen Parameter, die für ein jeweiliges Clusteringverfahren gesetzt werden, beeinflussen das Ergebnis. Daher bedarf es Methoden zur Evaluation und Vergleich der Clusterings miteinander. [RAAQ11]

Externe Indizes (engl. External Indices) sind Metriken, welche die berechnete Gruppenzuteilung mit einer extern vorgegeben Zuteilung vergleicht. Das heißt, es gibt eine erwartete Art der Gruppierung, welche nicht Teil des Datensets selbst ist. Diese Indizes messen nun den Grad der Übereinstimmung zwischen berechnetem und gewünschten Clustering-Ergebnis. [RAAQ11]

Interne Indizes (engl. Internal Indices) messen die Qualität des Clusterings ohne externe Informationen. Die Cluster sollten möglichst "kompakt" gruppiert und die verschiedenen Gruppen gut von einander getrennt sein. Diese Indizes versuchen diese Anforderungen zu quantifizieren. [King15]

Die **Stabilität** ist ebenfalls ein wichtiger Faktor. Hierbei wird das Datenset mehrmals mit verschiedenen Modifikationen geclustert. Solche Modifikationen können die Änderung einzelner Werte oder das Weglassen ganzer Spalten sein. Ein "stabiles" Clustering erzeugt auch mit diesen Veränderungen ähnliche Ergebnisse. Für die konkrete Bewertung der Stabilität werden externe oder interne Indizes verwendet und ihre Ergebnisse über die verschiedenen Clusterings miteinander verglichen. [King15, Kap. 8.1. Cluster Validity – Introduction]

Manche Autoren (z.B. [RAAQ11] und auch [King15]) führen formal noch sog. **Relative Inidzes** an. Hiermit wird die Performance verschiedener Clustering-Verfahren und verschiedener Meta-Parameter der Clusterings über das gleiche Datenset verglichen. In der Praxis erfolgt dies aber stets über den Vergleich der externen oder internen Indizes der Clustering-Ergebnisse miteinander.

2.5.2 External Indices

Überblick

Zur Bewertung der Übereinstimmung zweier Clustering-Ergebnisse können grundsätzlich Metriken aus dem Bereich der Klassifikation verwendet werden [HuAr85]. Beispiele dafür wären Maße wie die Entropie oder ihr Gegensatz die Reinheit (engl. Purity), welche analysieren, wie viele falsche zu richtigen Zuordnungen innerhalb einer Klasse aufgetreten sind [RAAQ11]. Ebenso das F-Maß (engl. F-Measure), welches aus dem Bereich des Document Retrievals bekannt ist. Es simuliert durchgeführte "Suchen" und vergleicht die gefundenen mit den erwarteten Suchergebnissen [StKaKu00].

Das Problem mit diesen Maßen ist, dass die Benennung der Labels im Ergebnis von entscheidender Bedeutung in der Bewertung ist. Nehmen wir an, wir haben ein Datenset X mit vier Datenpunkten. Die berechnete Clusterzuordnung ist $Y = (0, 0, 1, 1)$ und die erwartete Zuordnung ist $Y' = (1, 1, 0, 0)$. Metriken der Klassifikation würden eine Übereinstimmung von null Prozent feststellen, da keines der Labels in Y und Y' den gleichen Klassen zugeordnet worden ist. Im Kontext der Clusteranalyse weisen Y und Y' aber eine perfekte Übereinstimmung auf, denn es geht alleine um die Zuordnung der Datenpunkte in die gleichen Gruppen. Wie diese Gruppen benannt werden (in diesem Beispiel 0 oder 1), spielt dabei keine Rolle. Daher sind für die Clusteranalyse eigene Metriken entwickelt worden, welche unabhängig von der Benennung der Labels die Übereinstimmung zwischen den Zuordnungen berechnen können. [HuAr85]

Rand-Index

Der Rand-Index war einer der ersten Metriken, die speziell für die Clusteranalyse entwickelt worden sind [King15, Kap. 8.4 Indices of Cluster Validity]. William M. Rand veröffentlichte dieses Maß 1971 [Rand71] und es ist immer noch eines der populärsten und weitverbreitetsten (bzw. vor allem die Verbesserung dieses Indexes siehe nächster Abschnitt). [Stei04]

Um unabhängig von der Bennung der Labels zu werden, vergleicht man nicht wie in der Klassifikation die gefundenen Labels eins zu eins miteinander. Stattdessen werden alle möglichen Paarungen der Datenpunkte in X betrachtet und die Zuordnung der Paare in Y und Y' miteinander verglichen. [Rand71]

Tabelle 2.4: Kontingenz der Datenpunkt-Paare in Y und Y' [Stei04]

$Y \setminus Y'$	Paar im gleichen Cluster	Paar in anderem Cluster
Paar im gleichen Cluster	a	b
Paar in anderem Cluster	c	d

Die Tabelle zeigt die möglichen Fälle: Ein Paar aus Datenpunkten kann entweder dem gleichen Cluster oder zwei unterschiedlichen Clustern zugeordnet worden sein. Haben sowohl Y als auch Y' das Paar jeweils in das gleiche Cluster (mit dem gleichen Label wie auch immer dieses benannt ist) zugordnet, so wird dieses Paar zu a gezählt. Haben beide jeweils ein unterschiedliches Cluster zugeordnet, so wird das Paar in d gezählt usw. [Stei04]

$$rand = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}}$$

Der Rand-Index teilt nun die Menge an Paaren, welche die gleiche Zuordnung erhalten haben (a und d) durch die Anzahl aller möglichen Paarungen. Das Ergebnis ist ein Wert zwischen 0 (keine Übereinstimmung) und 1 (perfekte Übereinstimmung). [Rand71]

Der Rand-Index weist eine Reihe von Problemen auf, welche teilweise von Rand selbst erkannt oder später ermittelt worden sind:

- Je höher die Anzahl an Clustern ist, desto höher (näher an der 1) liegt der Rand-Index standardmäßig. Das liegt daran, dass es viele Paare in Y und Y' gibt, die unterschiedlichen Clustern zugeordnet worden sind. Das ist bei einer hohen Anzahl an Clustern ein häufig auftretendes Phänomen. [Rand71]
- Zum Vergleich verschiedener Clustering-Verfahren werden häufig Monte-Carlo-Simulationen mit großen Mengen an zufällig generierten Datensätzen und Clusterings verwendet. Werden zwei Zufalls-Zuordnung miteinander verglichen, so gibt der Rand-Index keine Werte nahe der 0, was aber wünschenswert wäre. Die Anzahl an zufälligen Übereinstimmungen wird also nicht adequat herausgerechnet. [King15]

Adjusted Rand-Index

Aus den genannten Problemen haben eine Vielzahl von Autoren versucht, eine bessere Variante des Rand-Indexes zu finden. Vor allem das "Herausrechnen" von angeblichen hohen Übereinstimmungen bei zufälligen Zuteilungen (engl. correction for chance) ist eines der Hauptanliegen gewesen [HuAr85]. Von allen vorgestellten Lösungen scheint die Variante von Hubert und Arabie [HuAr85] diejenige mit den wünschenwertesten Eigenschaften zu sein. (siehe die Versuche z.B. von Steinley [Stei04])

$$arand = \frac{rand - rand_{expected}}{rand_{max} - rand_{expected}}$$

Im Allgemeinen geht es darum, den Rand-Index zu berechnen und anschließend den "erwarteten Rand-Index" für zufällige Zuordnungen davon abzuziehen. Verschiedene Autoren haben nun unterschiedliche Methoden hergeleitet, um diesen "erwarteten Rand-Index" zu berechnen. [HuAr85]

$$arand = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

Hubert und Arabie berechnen den "erwarteten Rand-Index" aus der Menge aller möglichen Permutationen der Paarungen und ihrer Übereinstimmung $\frac{(a+b)(a+c)+(c+d)(b+d)}{\binom{n}{2}}$. Die gegebene Formell zeigt eine breits vereinfachte Umstellung von Steinley. [Stei04]

Dieser "Adjusted Rand-Index" liefert Werte zwischen -1 und 1. Negative Ergebnisse oder Werte um die 0 zeigen, dass die Übereinstimmung der beiden Clusterings der zu erwartenden Übereinstimmung aufgrund von Zufall entspricht (oder sogar deutlich darunter liegt) und damit nicht signifikant ist. Höhere Zahlen nahe der 1 stehen nachwievor für ein perfekte Übereinstimmung, welche zusätzlich statistische Signifikanz aufweist. Auch der Bias zu hohen Werten bei einer hohen Anzahl an Clustern wird daurch ausgeglichen. [Stei04]

2.5.3 Internal Indices

In der Literatur werden eine Vielzahl von Indizes beschrieben, welche die Qualität der gefundenen Cluster messen können. Einen Überblick dazu geben Rendón et al. [RAAQ11]. An dieser Stellen sollen nur ein paar Vertreter vorgestellt werden, welche für diese Arbeit von besonderer Relevanz sind.

Silhouettenkoeffizient

Der Silhouettenkoeffizient wurde von Peter J. Rousseeuw entwickelt [Rous87] und ist ein häufig verwendetes Maß für die Qualität von Clusterings.

$$s(x) = \frac{b(x) - a(x)}{\max a(x), b(x)}$$
$$a(x) = d(x, C_x)$$
$$b(x) = \min_{C_x \neq C_i} d(x, C_i)$$

Für jeden Datenpunkt x im Datenset wird die Silhouetten-Weite $s(x)$ berechnet. Diese ergibt sich aus der Differenz zwischen dem durchschnittlichen Abstand zu allen Datenpunkten im selben Cluster ($a(x)$) und dem durchschnittlichen Abstand zu allen Datenpunkten im direkt benachbarten Cluster ($b(x)$). Anschließend wird der Wert zwischen -1 und 1 normiert. [Rous87]

- Eine negative Silhouetten-Weite gibt an, dass der Datenpunkt näher am benachbarten als an seinem eigenen Cluster liegt und somit falsch zugeordnet worden ist.
- Werte um die 0 zeigen, dass der Datenpunkt fast mittig zwischen beiden Clustern platziert ist.
- Befindet sich die Weite nahe der 1, so liegt der Datenpunkt mittig in seinem eigenen Cluster und das benachbarte Cluster ist ordentlich weit entfernt.

$$sil = \frac{1}{n} \sum_{i=1}^{n} s(x_i)$$

Der Koeffizient ergibt sich schließlich aus dem Durchschnitt aller Silhouetten-Weiten aller Datenpunkte. Werte nahe der 1 deuten auf kompakte und wohl-separierte Cluster hin. [Rous87]

In den Versuchen von Rendón et al [RAAQ11] erwies sich dieser Index als einer der besten Indikatoren für ein gutes Clustering-Ergebnis. Er ist die direkte mathematische Definition für die Anforderung, dass Datenpunkte im gleichen Cluster möglichst ähnlich und zu den Punkten der andere Cluster möglichst unähnlich sein sollen. Ein Nachteil dieses Indizes ist die quadratische Laufzeit, da jeder Punkt mit jedem anderen verglichen werden muss.

Davies-Bouldin Index

Der Davies-Bouldin Index wurde von seinen Namesgebern David L. Davies und Donald W. Bouldin [DaBo79] entwickelt. Ziel war es, eine Metrik zu konstruieren, welche die durchschnittliche Ähnlichkeit zwischen benachbarten Clustern berechnet.

$$dbi = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{d(c_i, C_i) + d(c_j, C_j)}{d(c_i, c_j)}$$

Die Ähnlichkeit zwischen den Clustern K (K steht hier für die Menge an gefundenen Clustern C_i) berechnet sich aus dem durchschnittlichen Abstand der Punkte eines Clusters zu ihrem Cluster-Mittelpunkt ($d(c_i, C_i)$ und $d(c_j, C_j)$) geteilt durch den Abstand der beiden Cluster-Mittelpunkte zueinander ($d(c_i, c_j)$). [DaBo79]

Je kleiner der Wert des Indexes, desto enger liegen die Datenpunkte um ihren Cluster-Mittelpunkt im Verhältnis zum benachbarten Cluster. Möglichst niedrige Werte nahe der 0 sind also als optimal anzusehen. [DaBo79]

Der große Vorteil von diesem Verfahren ist die geringere Laufzeit, da die Punkte der Cluster nur mit ihrem Mittelpunkt verrechnet werden. Nachteilig ist, dass die Be- und Verechnung der Mittelpunkte primär nur für numerische Vektoren definiert ist. In den Versuchen von Rendón et al. [RAAQ11] schnitt dieser Index genauso gut ab wie der Silhouettenkoeffizient.

3 Konzeption

3.1 Überblick

- möglichst realistische Datenquellen => Akeneo-PIM + Icecat Datenkatalog
- Clustering mittels möglichst effizienten Verfahren
  - dabei verschiedene Attribute, Datentypen, Gewichtungen probieren
  - möglichst ohne "externe" Informationen clustern können
- Evaluation => ist clustering sinnvoll => Was ist sinnvoll?
- Überblick über das Vorgehen

3.2 Datenquellen und -sets

3.2.1 Akeneo-PIM

Überblick

- open source PIM-System
- weite Verbreitung
- typischer Funktionsumfang
- große Menge an Addons/Plugins (u.a. Icecat Importer)

Systemübersicht

- Microservice Cluster
- Frontend UI
- Rest-Api

Datenstrukturen

- viele Verschiedene: AUFZÄHLUNG...
- nur einige relevant => genauerer Blick

Product & Product Values

- Meta-Values
- "values" Eintrag ist Mapping: attribute.code => product_value

Category

- Baumstruktur für Kategorisierung

Family

- definiert Attribute pro Family und required/optional

Attribute & Attribute Group

- verschiedene Typen + Constraints

Channel, Currency, Locale

- alles lokalisierbar (Sprache und Währung)
- zusätzlich noch scopes => verschiedene Werte für gleiche Locale z.B. Web und Mobile

Measurement Family

- Maßeinheiten mit Umrechnung

3.2.2 Icecat Produkt-Katalog

- offener Katalog mit XXXXXXXX Produktdaten
- einigermaßen standardisiert nach UNSPEC-Taxonomy

3.2.3 Datenset

Anforderungen

- Produkte mit wenigen Attributen
- Produkte mit vielen Attributen
- unterschiedliche aber naheliegende Kategorien
- Produkte mit Variationen und verschiedenen Generationen
- Duplikate

Umsetzung

- Smartphone-Hüllen speziell für Samsung Galaxy Reihe
- Smartphones aus Samsung Galaxy Reihe S20 und S21 + Versionen und Farben
- evtl. noch weitere Smartphones
- evtl. ein paar Tablets
- evtl. noch was ganz anderes wie Ladegeräte o.ä.

3.3 Clustering

3.3.1 Clustering-Verfahren

Anforderungen:

- hierarchisch wünschenswert
  - mehr Informationen ableitbar
  - Anzahl an Clustern dynamisch ableitbar
- gute Laufzeit => potenziell hohe Menge an Produkten
- Verarbeitung verschiedenster Datentypen in den Attributen

Ansatz: Bisecting K-Prototypes mit Custom-Varianten von Distanzfunktionen

3.3.2 Distanzfunktion

Anforderungen

- Berechnung der Distanz + Berechnung des Mittelpunkts
- Kombi von verschiedenen Datentypen
- viele Null-Values

Tabelle 3.1: Verarbeitung der Akeneo-Attribut-Typen

Datentyp	Akeneo-Typ	Distanzmaß	Centroid
numerisch	Number, Metric, Price, Date	Minkowski	mean
kategorisch	Bool, SelectSingle, ReferenceDataSingle	Jaccard	mode
multi-kategorisch	SelectMulti, ReferenceDataMulti	Jaccard+	mode
strings	Text, Textarea	?	?
Datei	File, Image	-	-
Id	Identifier	-	-

Herleitung der Funktion

- $d(p_1, p_2) = \frac{v+n+c+m+?}{|p_1^{attr} \cup p_2^{attr}|}$
- $v = |p_1^{attr} \setminus p_2^{attr} \cup p_2^{attr} \setminus p_1^{attr}|$
- $n = \sum_{x \in p_1^{num_i},\ y \in p_2^{num_i}} |x - y|$
- $c = \sum_{x \in p_1^{cat_i},\ y \in p_2^{cat_i}} q(x, y)$
- $q(x, y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$
- $m = \sum_{x \in p_1^{mul_i},\ y \in p_2^{mul_i}} 1 - \frac{|x \cap y|}{|x \cup y|}$

... weitere Erklärungen etc...

Umgang mit verschiedenen Versionen der Funktion?

3.4 Evaluation

- Ziel: sinnvolle Cluster finden => Was heißt das?

3.4.1 Stabilität der Cluster

- Hintergrund:
  - Clustering ist wertlos, wenn Zuordnung zufällig erfolgt
  - Clustering sollte stabil, reproduzierbar & weitestgehend deterministisch sein
- mehrmaliges Clustern (gleiche Settings):
  - K-Means mit random init => evtl. unterschiedliche Ergebnisse bei verschiedenen Durchläufen
  - gute Distanzfunktion => klare Separierung der Produkte
  - je größer die Ähnlichkeit zwischen zwei Clusterings, desto stabiler
  - Bewertung der Ähnlichkeit mittels Jaccard-Koeffizient
- Analyse der Auswirkung Settings
  - Clustern mit verschiedenen Attributen/Gewichtungen
  - Bewertung der Ähnlichkeit der Clusterings
  - ebenfalls Jaccard-Koeffizient

3.4.2 Internal Indices

- allgemeine Stats zum Clustering:
  - durchschnittlicher Error
  - Entwicklung über Hierarchien
  - Anzahl an Produkten

=> eher untergeordnete Rolle, eventuell weglassen ???

3.4.3 External Indices

- Hintergrund:
  - bisher nur Aussage wie ähnliche/verschieden die Clusterings sind
  - Frage ist noch: Wenn Clusterings unähnlich,welches ist “besser”
  - Ansatz => Inspiration aus möglichen Anwendungen:
    - Auto-Kategorisierung
    - Einordnung neuer Produkte zu bestehenden (transitive Erkenntnisse)
    - Finden von Alternativ-Produkten mit gleichen Eigenschaften
    - Finden von Duplikaten
- Vergleich mit Akeneo Families und Akeneo Categories
  - Families: ähnliche Attribute, sollten sich ähnlich sein, eher technischer Check
  - Categories: menschengemachte Labels, spiegeln die intuitive Zuordnung wieder

  - Analyse der Ähnlichkeit zwischen Clustering und Families/Categories

- speziellere Checks:
  - bewusst eingefügte Duplikate => solange wie möglich im gleichen Cluster (top-down clustering)
  - Produkte unterschiedlicher Kategorien, aber mit Bezug (z.B. Smartphone plus Hülle) => bleiben sie länger im gleichen Cluster als andere Produkte dieser Kategorien?

3.5 Vorgehen / Versuche ?

3.5.1 Set "Samsung Galaxy Cases"

- Smartphone Hüllen für Samsung Galaxy Reihe sowohl
- überschaubar wenig Attribute

Versuch: Basis-Prozess, Fehler finden

- clustern nach einander mit numerisch + kategorisch (+ multi-kategorisch)
- Prozess und "Pipeline" aufsetzen

Versuch: gleiche Modelle

- Clustering soll Hüllen für gleiches Smartphone-Modell finden
- werden in Akeneo Categories vorher hinterlegt, Vergleich damit
- numerisch & kategorisch => einzeln und zusammen
- (multi-kategorisch => hier noch nicht möglich, da solche keine Attribute)
- gleich-gewichtet vs menschen-gewichtet

3.5.2 Set "Samsung Galaxy S 20er-Reihen"

- Smartphones Samsung Galaxy S 20 und 21 Modelle
- verschiedenen Versionen (Größen) und Farben
- sehr viel Attribute
- enthält Duplikate

Versuch: gleiche Modelle finden

- Clustering soll gleiche Modelle/Baureihen finden
- werden in Akeneo Categories vorher hinterlegt, Vergleich damit
- vor allem interessant: nur required Attribute vs optionale hinzunehmen => wird es dadurch besser oder schlechter?
- multi-kategorisch ebenfalls auswerten => macht es einen Unterschied?
- gleich-gewichtet vs menschen-gewichtet

Versuch: Duplikate finden

- Duplikate sollten sehr lange im gleichen Cluster sein

- eventuell zusammen mit dem vorherigen Versuch evaluieren ???
- Hauptfrage: gleiche Settings für beide Aufgaben nutzbar?

3.5.3 Set “Samsung Galaxy Smartphones und Hüllen”

- Kombi beider Sets
- Clustering mit gefundenen Konfig(s)
- Frage: Assoziation zwischen Modellen sichtbar?

...

3.5.4 weiteres ?

- weitere Versuche?
- eventuell durch Hinzunehmen von Smartphones und Hüllen anderer Hersteller ?
- weitere Produktgruppe – ähnlich aber schon anders z.B. Tablets?
- weitere Produktgruppe – ganz anders z.B. andere Elektroartikel?

4 Implementierung

- Akeneo-PIM:
  - auf Server von adesso
  - Akeneo Icecat Connector
  - Import der Daten
- Python Modules:
  - Akeneo Client => holt Daten aus Akeneo-PIM und dumpt in json Dateien => “raw data”
  - Akeneo Parser => parsing von json in Python-Strukturen (dataclasses) => “cleaned data”
  - Akeneo Clustering => Filtern der Attribute, Vorverarbeitung fürs Clustering
  - Clustering => eigene Implementierung des Bisecting K-Means mit generischen Datenstrukturen
  - Evaluation => Hilfsmethoden zur Auswertung der Ergebnisse
- Jupyter Notebooks:
  - Akeneo Rest Api Exploration
  - Clustering Basis Example (2d Vektoren)
  - verschiedene Iterationen des Clusterings und Distanzfunktionen

...

5 Auswertung

...

6 Fazit und Ausblick

Quellenverzeichnis

[ArVa16] Arora, Preeti ; Varshney, Shipra ; u. a.: Analysis of k-means and k-medoids algorithm for big data. In: *Procedia Computer Science* Bd. 78, Elsevier (2016), S. 507–512

[BiHa09] Bissantz, Nicolas ; Hagedorn, Jürgen: Data Mining (Datenmustererkennung). In: *Wirtschaftsinformatik* Bd. 51, Springer (2009), Nr. 1, S. 139–144

[Bosl12] Boslaugh, Sarah: *Statistics in a nutshell: A desktop quick reference* : O'Reilly Media, 2012

[Cha07] Cha, Sung-Hyuk: Comprehensive survey on distance/similarity measures between probability density functions. In: *City* Bd. 1 (2007), Nr. 2, S. 1

[ChZhYi19] Chen, Hongshen ; Zhao, Jiashu ; Yin, Dawei: Fine-grained product categorization in e-commerce. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, S. 2349–2352

[CRF03] Cohen, William W ; Ravikumar, Pradeep ; Fienberg, Stephen E ; u. a.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: *IIWeb*. Bd. 3 : Citeseer, 2003, S. 73–78

[Cui21] Cui, Yimin: Intelligent recommendation system based on mathematical modeling in personalized data mining. In: *Mathematical Problems in Engineering* Bd. 2021, Hindawi (2021)

[DaBo79] Davies, David L ; Bouldin, Donald W: A cluster separation measure. In: *IEEE transactions on pattern analysis and machine intelligence*, IEEE (1979), Nr. 2, S. 224–227

[Huan98] Huang, Zhexue: Extensions to the k-means algorithm for clustering large data sets with categorical values. In: *Data mining and knowledge discovery* Bd. 2, Springer (1998), Nr. 3, S. 283–304

[HuAr85] Hubert, Lawrence ; Arabie, Phipps: Comparing partitions. In: *Journal of classification* Bd. 2, Springer (1985), Nr. 1, S. 193–218

[JiCh17] Jia, Hong ; Cheung, Yiu-Ming: Subspace clustering of categorical and numerical data with an unknown number of clusters. In: *IEEE transactions on neural networks and learning systems* Bd. 29, IEEE (2017), Nr. 8, S. 3308–3325

[KaRo09] Kaufman, Leonard ; Rousseeuw, Peter J: *Finding groups in data: an introduction to cluster analysis*. Bd. 344 : John Wiley & Sons, 2009

[King15] King, Ronald S: *Cluster analysis and data mining: An introduction* : Stylus Publishing, LLC, 2015

[KoLo12] Kou, Gang ; Lou, Chunwei: Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data. In: *Annals of Operations Research* Bd. 197, Springer (2012), Nr. 1, S. 123–134

[KRRT01] Kumar, Ravi ; Raghavan, Prabhakar ; Rajagopalan, Sridhar ; Tomkins, Andrew: Recommendation systems: A probabilistic analysis. In: *Journal of Computer and System Sciences* Bd. 63, Elsevier (2001), Nr. 1, S. 42–61

[OhKi19] Oh, Yoori ; Kim, Yoonhee: A resource recommendation method based on dynamic cluster analysis of application characteristics. In: *Cluster Computing* Bd. 22, Springer (2019), Nr. 1, S. 175–184

[PWMO19] Papp, Stefan ; Weidinger, Wolfgang ; Meir-Huber, Mario ; Ortner, Bernhard ; Langs, Georg ; Wazir, Rania: *Handbuch Data Science: Mit Datenanalyse und Machine Learning Wert aus Daten generieren* : Carl Hanser Verlag GmbH Co KG, 2019

[RAAQ11] Rendón, Eréndira ; Abundez, Itzel ; Arizmendi, Alejandra ; Quiroz, Elvia M: Internal versus External cluster validation indexes. In: *International Journal of computers and communications* Bd. 5 (2011), Nr. 1, S. 27–34

[Rand71] Rand, William M: Objective criteria for the evaluation of clustering methods. In: *Journal of the American Statistical association* Bd. 66, Taylor & Francis (1971), Nr. 336, S. 846–850

[RaRa11] Rajalingam, N ; Ranjini, K: Hierarchical Clustering Algorithm - A Comparative Study. In: *International Journal of Computer Applications* Bd. 19, Citeseer (2011), Nr. 3, S. 42–46

[Rous87] Rousseeuw, Peter J: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In: *Journal of computational and applied mathematics* Bd. 20, Elsevier (1987), S. 53–65

[RoZa09] Robertson, Stephen ; Zaragoza, Hugo: *The probabilistic relevance framework: BM25 and beyond* : Now Publishers Inc, 2009

[Stei04] Steinley, Douglas: Properties of the Hubert-Arable Adjusted Rand Index. In: *Psychological methods* Bd. 9, American Psychological Association (2004), Nr. 3, S. 386

[StKaKu00] Steinbach, Michael ; Karypis, George ; Kumar, Vipin: A comparison of document clustering techniques (2000)

Anhang

- Link zum Repo
- Überblick zum Repo

Selbstständigkeitserklärung

Die Autoren versichern hiermit, dass die vorliegende Arbeit mit dem Thema

Generisches Clustern hoch-komplexer Produktdaten

selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmitteln angefertigt worden ist.

Erfurt, den 29.04.2022

.....
Hannes Dröse