FACHHOCHSCHULE ERFURT UNIVERSITY OF APPLIED SCIENCES
Angewandte Informatik

SKILL 2022
Bisecting K-Prototypes: Effizientes hierarchisches Clustering gemischter Datensets

Hannes Dröse

Fachhochschule Erfurt
adesso SE, Jena

Motivation

FACHHOCHSCHULE ERFURT UNIVERSITY OF APPLIED SCIENCES
Angewandte Informatik

- Masterarbeit in Kooperation mit adesso SE Standort Jena: E-Commerce

adesso

- Anwendungen von Clusteranalyse auf Produktdaten
  - bessere Empfehlungsalgorithmen ([Cui21], [KRRT01], [OhKi19])
  - bessere Klick-Raten in Suchmaschinen ([KoLo12])
  - Anomalie- und Duplikaterkennung etc.

- Problem: Produktdaten in PIM-Systemen sehr komplex

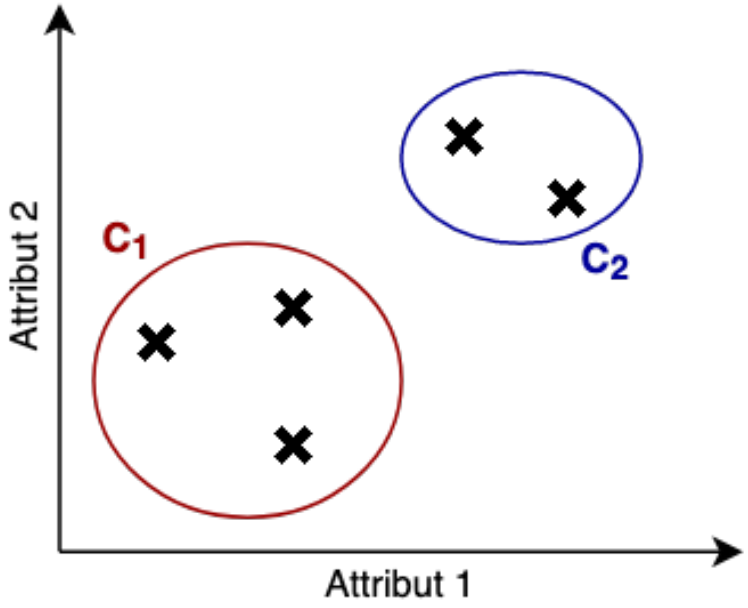
Produktdaten in PIM-Systemen

id	Title	Color	Height	5G	OS	Material	…
1	Samsung Galaxy S20 128GB red	red	152 mm	*null*	Android 10	*null*	…
2	Samsung Galaxy S20 128GB black	black	152 mm	*null*	*null*	plastic, silicone	…
3	Samsung Galaxy S21	grey	151 mm	false	Android 11	*null*	
4	Samsung Galaxy S21 5G	grey	151 mm	true	Android 11	*null*	…
5	Samsung Galaxy S22	black	164 mm	true	Android 12	plastic, aluminium	…

- durchzogen von **fehlenden Werten**
- **vielfältige Datentypen:** numerisch, kategorial, multi-kategorial, Strings, Dateien etc.

Clusteranalyse

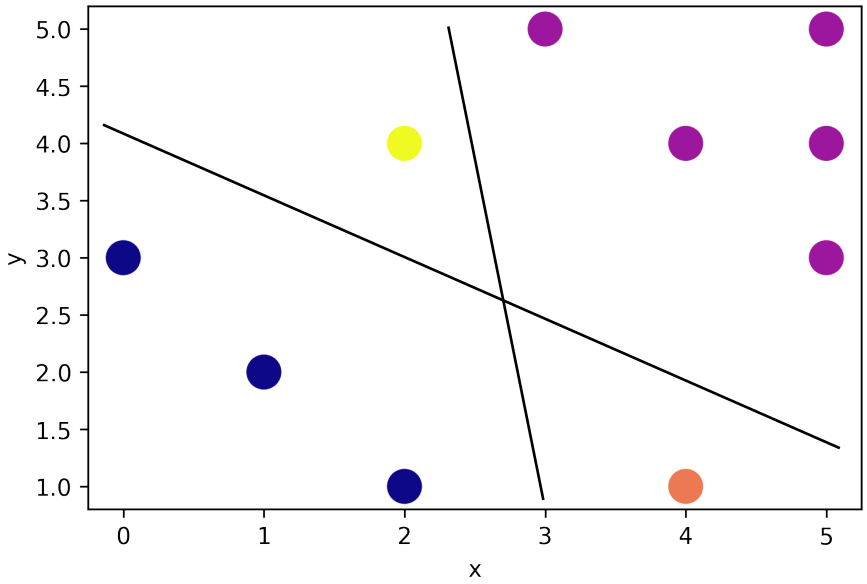
Die Einteilung von Datenpunkten in Gruppen, “[. . .] sodass sich die Individuen innerhalb einer Gruppe auf eine Art und Weise ähnlich sind und unähnlich denen in anderen Gruppen” [King15]



- Cluster entstehen aus „sich naheliegenden Punkten“
- Berechnung mittels Distanzfunktion $d(x_1, x_2)$

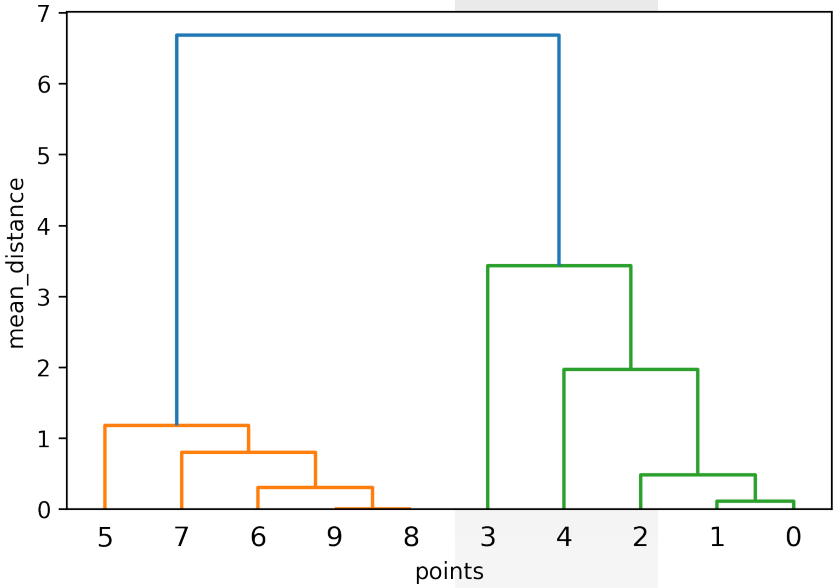
Clustering-Verfahren

partitionierend



- Einteilung in k eindeutige Cluster
- Minimierungsverfahren
- z.B.: K-Means
- $\mathcal{O}(n)$

hierarchisch



- verschachtelte Cluster für alle möglichen k
- Top-down oder Bottom-up
- keine späteren Korrekturen
- $\mathcal{O}(n^2)$ bis $\mathcal{O}(n^3)$

Clustering-Verfahren 2

Bisecting K-Means [StKaKu00]

- Top-down-Clustering-Verfahren
- nutzt **K-Means** für die **Zweier-Splits**
- Laufzeit: $\mathcal{O}(n)$ – evtl. $\mathcal{O}(n \log n)$

K-Prototypes [Huan98]

- K-Means-Variante für gemischte Datensets (numerisch und kategorial)
- Mittelpunkte aus Durchschnitt (numerisch) bzw. Modus (kategorial)
- kombinierte Distanzfunktion:

$$d(x_1, x_2) = d_{num}(x_1^{num}, x_2^{num}) + w \cdot d_{cat}(x_1^{cat}, x_2^{cat})$$

Idee: Kombination beider Verfahren => Bisecting K-Prototypes

Konzeption der Distanzfunktion

- **offen:** Umgang mit *fehlenden Werten*
- **Ansatz:** Inspiration durch Jaccard-Koeffizienten $\frac{|x_1 \cap x_2|}{|x_1 \cup x_2|}$

$$d(x_1, x_2) = \frac{\sum d'(x_1^i, x_2^i)}{|x_1^{non-null} \cup x_2^{non-null}|}$$

$$d'(x_1^i, x_2^i) = \begin{cases} 0 & , x_1^i \text{ is } null \wedge x_2^i \text{ is } null \\ 1 & , x_1^i \text{ is } null \vee x_2^i \text{ is } null \\ |x_1^i - x_2^i| & , i \text{ is } numerical \\ 0 & , i \text{ is } categorical \wedge x_1^i = x_2^i \\ 1 & , i \text{ is } categorical \wedge x_1^i \neq x_2^i \end{cases}$$

- **wichtig:** numerische Attribute vorher auf Intervall [0;1] normalisieren

Konzeption der Distanzfunktion 2

- **offen:** Umgang mit *multi-kategorialen* Werten

id	Title	Color	Height	5G	OS	Material	…
1	Samsung Galaxy S20 128GB red	red	152 mm	*null*	Android 10	*null*	…
2	Samsung Galaxy S20 128GB black	black	152 mm	*null*	*null*	**plastic, silicone**	…
3	Samsung Galaxy S21	grey	151 mm	false	Android 11	*null*	
4	Samsung Galaxy S21 5G	grey	151 mm	true	Android 11	*null*	…
5	Samsung Galaxy S22	black	164 mm	true	Android 12	**plastic, aluminium**	…

- **Ansatz:** Jaccard-Koeffizient auf Attribut-Ebene

$$d'(x_1^i, x_2^i) = \begin{cases} \dots \\ 1 - \dfrac{|x_1^i \cap x_2^i|}{|x_1^i \cup x_2^i|} & , i \text{ is } multi-categorical \end{cases}$$

Konzeption der Distanzfunktion 3

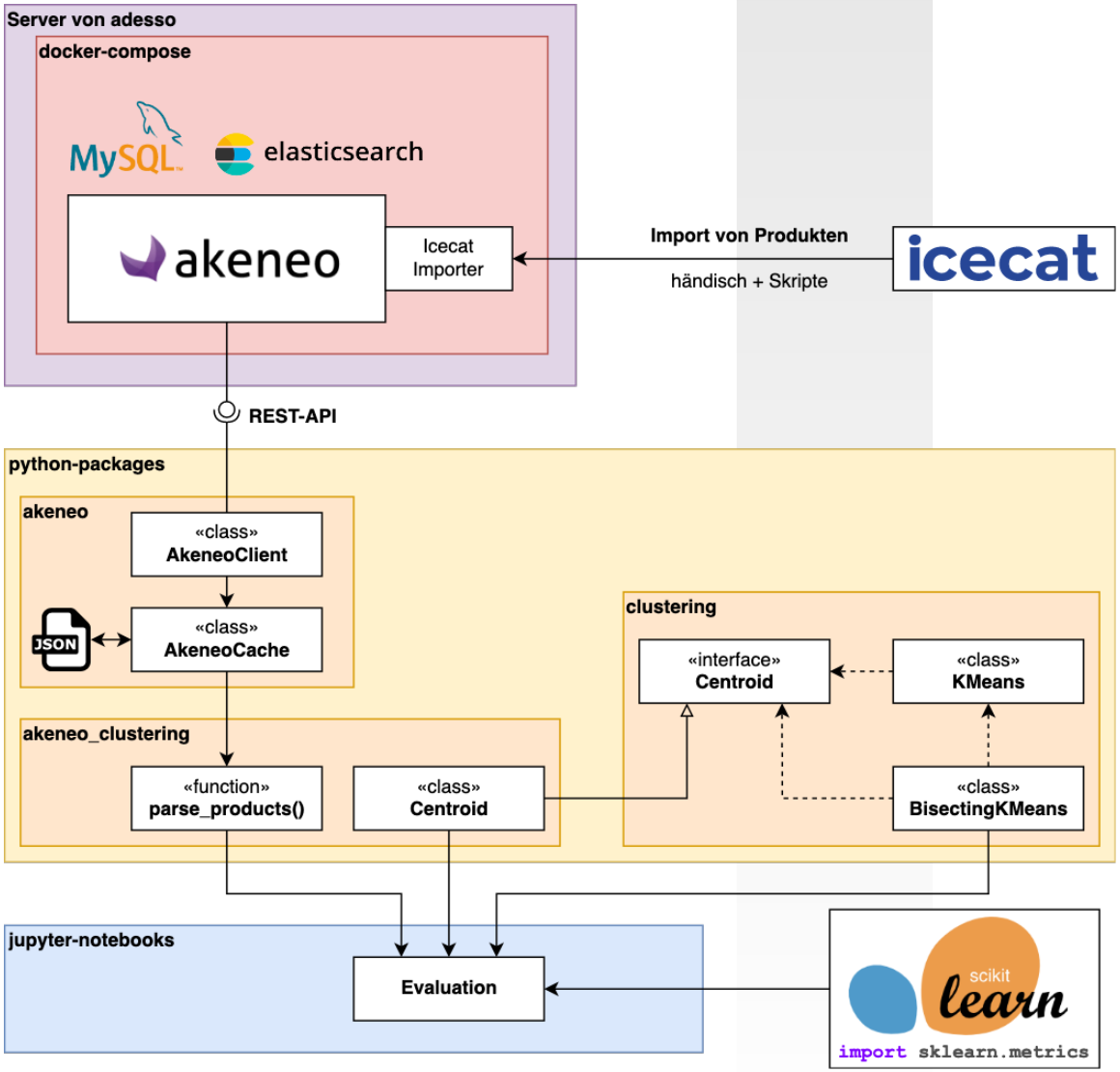
FACHHOCHSCHULE ERFURT UNIVERSITY OF APPLIED SCIENCES
Angewandte Informatik

- **offen:** Umgang mit *String*-Werten
- **Ansatz:** Umwandlung multi-kategoriale Attribute durch Tokenization, Stemming, Stop-Word-Removal

„Samsung Galaxy S20 128GB“
=> {samsung, galaxi, s20, 128gb}

Praktische Evaluation: Überblick

- Import von Produkten in ein PIM-System
- Implementierung des Clustering-Verfahrens
- Clustering des Datensets
- Evaluation mit Metriken für Stabilität, Qualität (Silhouetten-Koeffizient) etc.



Github: https://github.com/hd-code/ma-product-clustering

Praktische Evaluation: Datenset

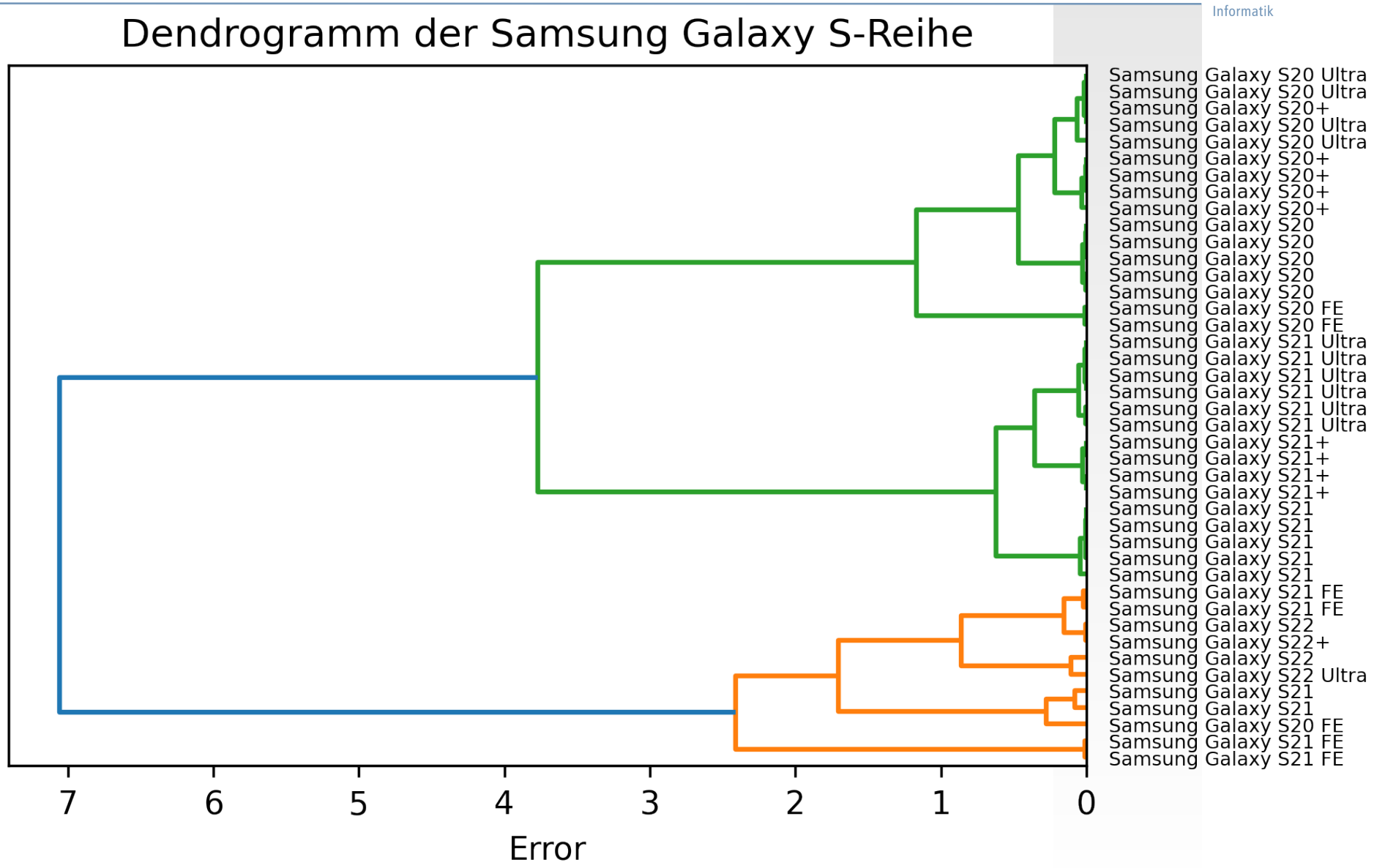
- 42 Samsung Galaxy Smartphones
- S20: 17 Stück, S21: 21 Stück, S22: 4 Stück
- jeweils mit den Varianten: Standard, Plus, Ultra und FE (Fan Edition)

Übersicht zu den Attributen der 42 Smartphones

Typ	Anzahl	gefüllt	Ø unique	Beispiele
numerisch	56	51,9 %	3,7	Weight, Width, Depth, Height
kategorial	106	67,1 %	1,3	OS installed, SIM Card Type
multi-kat.	22	60,0 %	3,5	Product Color, 3G standards
string	11	67,6 %	13,5	Title, Description
alle:	*195*	*61,7 %*	*2,9*	

Praktische Evaluation: Ergebnisse

FACHHOCHSCHULE ERFURT UNIVERSITY OF APPLIED SCIENCES
Angewandte Informatik



Fazit und Ausblick

- **Bisecting K-Prototypes** funktioniert grundsätzlich
- weitere Evaluation an anderen/größeren Datensets nötig
- nur **numerische** und **kategoriale** Attribute erzeugten **beste Cluster**
- nur **String-Attribute** erzeugten ebenfalls **adäquate Cluster** (Alternative für Clustering von semi-/unstrukturierten Datensets)
- Vergleich mit klassischen Verfahren wäre sinnvoll, aber aufwendig

Vielen Dank für die Aufmerksamkeit

Quellen (Ausschnitt)

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

[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

[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

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