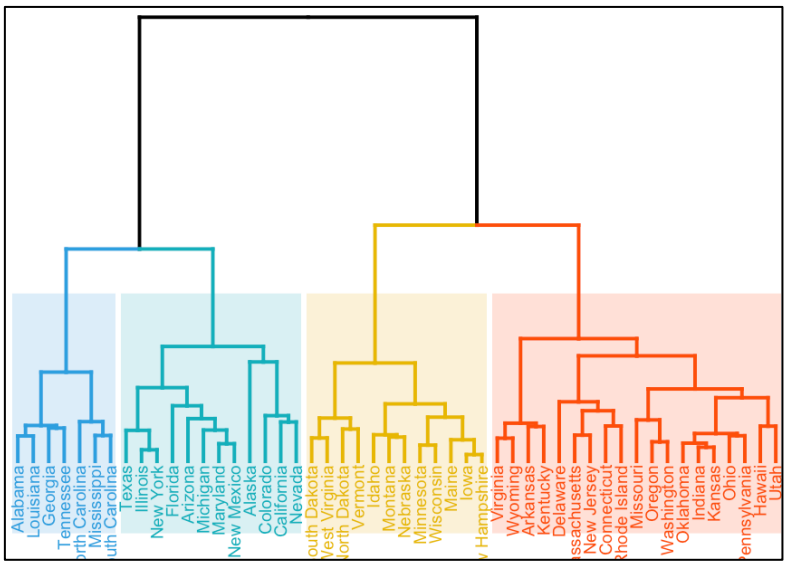
Prof. Dr.-Ing.habil.
Dirk Joachim Lehmann
Data Science in IoT
Fakultät für Informatik
di.lehmann@ostfalia.de
www.dirk-lehmann.de

Dendrogramm-Explorer

Hintergrund

Daten begegnen uns in der beruflichen alltäglichen Praxis sehr häufig, beispielsweise in Form einer Excel-Tabelle bzw. als Excel-Sheet. Eine Excel-Tabelle mit einer Anzahl an unterschiedlichen Spalten stellt aus wiss. Sicht einen hoch-dimensionalen Datensatz dar. Jede Spalte repräsentiert ein anderes Attribut bzw. eine weitere Dimension. Bei der Analyse von hoch-dimensionalen Daten interessieren sich Analysten für die Struktur der Anordnung von Daten im Datenraum, auch Muster genannt. Insbesondere erlaubt es das Wissen um diese Strukturen relevante Rückschlüsse zu ziehen, auf Eigenschaften der - den Daten zugrundeliegenden - Domäne. Die Analyse von hoch-dimensionalen Daten stellt noch immer eine große Herausforderung dar, auch weil es nicht einfach ist, diese Strukturen für den Analysten sinnvoll zu repräsentieren und darzustellen.

Eine Möglichkeit der Darstellung von Struktureigenschaften hochdimensionaler Daten sind *Dendrogramme*. Ein Dendrogramm ist eine 2D Baumvisualisierung der Clustereigenschaften /Gruppierungseigenschaften hochdimensionaler Daten. Sie werden in der Literatur auch den Begriff des *Hierarchisches Clustering* in diesem Zusammenhang finden. Details wie ein Dendrogramm prinzipiell konstruiert wird entnehmen Sie bitte dem Video [1] entnehmen.



https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/003-hierarchical-clustering-in-r/figures/005-visualizing-dendrograms-cutree-1.png, Stand: 14.08.2022

Abb.: Beispiel eines Dendrogramms - 2D Baumvisualisierung von Struktureigenschaften hochdimensionaler Daten

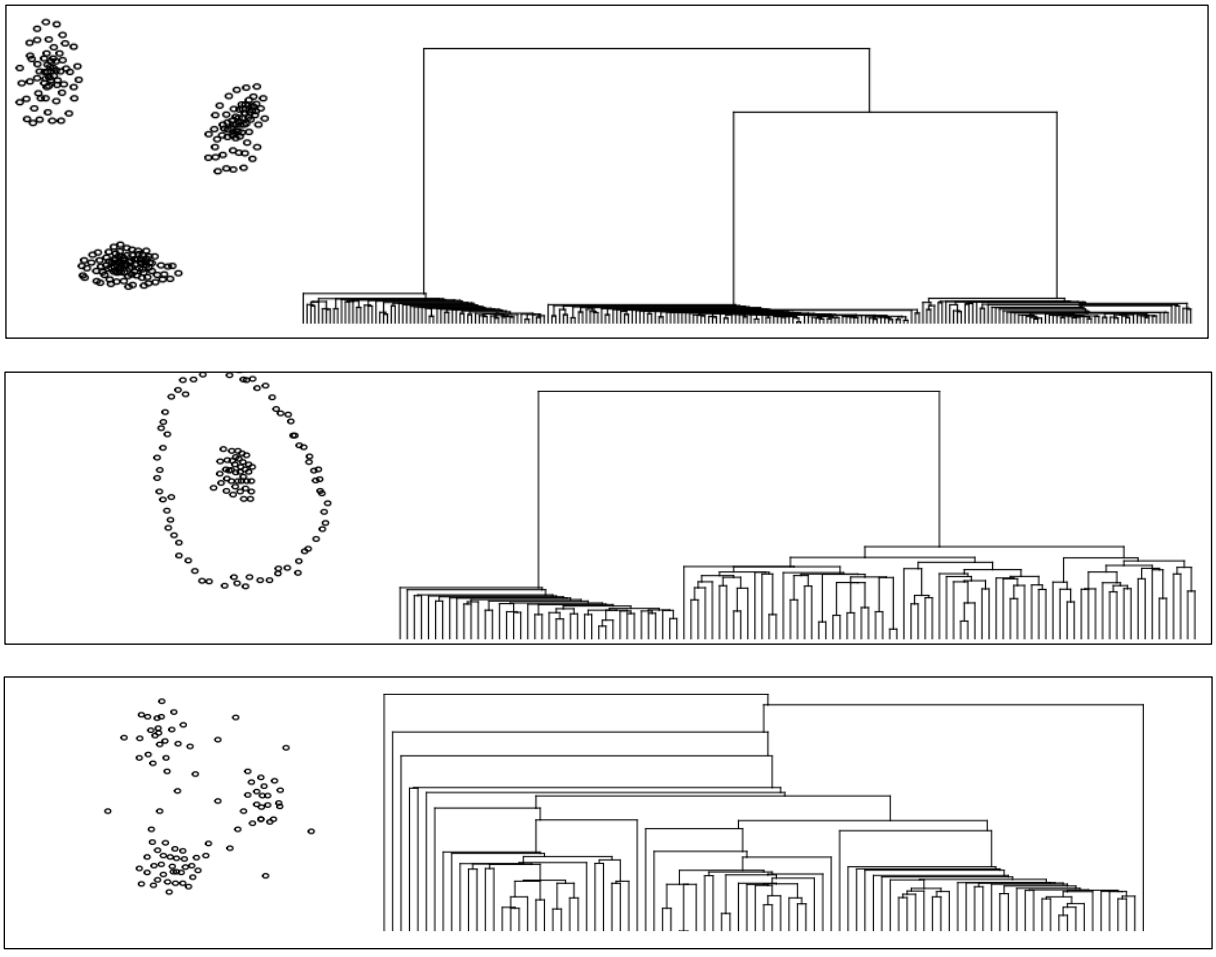
- -

Vorteile von Dendrogrammen – eine Kurzbetrachtung

Dendrogramme besitzen mehrere Vorteile für die visuelle Analyse von hochdimensionalen Daten für den Nutzer. Zum einen bieten sie die eine 2D Repräsentation der Daten, wodurch sie sowohl für den Monitor als auch direkt für den Print geeignet sind, zum anderen lassen sich strukturelle Charakteristika der Daten in der Topologie der Dendrogramme selbst ablesen und explorativ erkennen, was dem Nutzer eine explorative Analyse ermöglicht.

Dazu ein erläuterndes Beispiel. Die folgenden drei Abbildungen zeigen auf der rechten Seite je einen zwei-dimensionalen Datensatz und links ein - zu den Daten zugehöriges - Dendrogramm.

Daten Dendrogramm



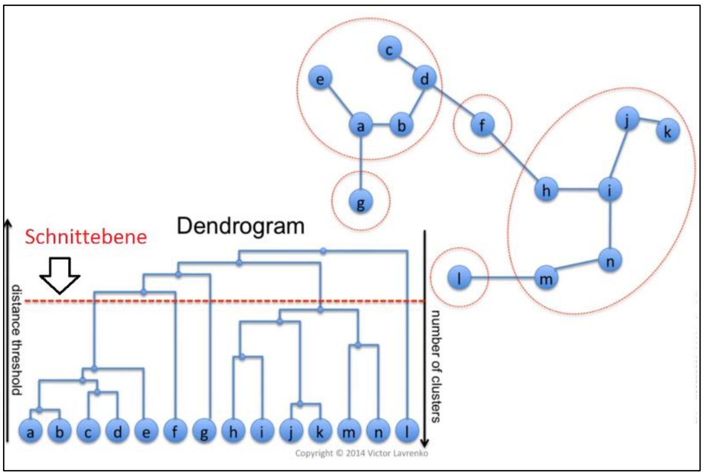
Es ist zu erkennen, dass sich die Struktureigenschaften der Dendrogramme ändern, wenn sich die Verteilung der Daten ändert. Diese Eigenschaft nutzen wir aus, um aus den Dendrogrammen Rückschlüsse über die Verteilungs- und Struktureigenschaften der Daten schließen zu können.

Der „Clou“ ist nun, dass dieses Vorgehen ebenfalls für Daten mit mehr als zwei Dimensionen sehr gut funktioniert: Sprich, für Daten mit 3,4, 5 oder mehr Dimensionen. Solche Daten lassen sich selbst nicht mehr sehr gut visuell darstellen, aber die Dendrogramme dieser Daten höherer Dimensionen bleiben 2D Repräsentationen, wodurch es möglich wird über die Dendrogramme Rückschlüsse auf hoch-dimensionale Daten zu ziehen.

- -

Schnittebenen von Dendrogrammen – eine Kurzbetrachtung

Betrachten wir einen vertikalen Schnitt durch das Dendrogramm, entspricht dies einer konkreten Gruppierung der Daten (auch Clusterisierung genannt):



https://www.youtube.com/watch?v=1jW9xlEtQao, **Stand: 14.08.2022**

Abb.: Schnitteben eines Dendrogramms

Durch Variation dieses Schnitts lassen sich unterschiedliche Gruppierungen und Partitionen der Daten generieren.

- -

Head Maps von Dendrogrammen – eine Kurzbetrachtung

Überlegen wir, wir hätten ein Dendrogramm der Datenelemente eines hoch-dimensionalen Datensatzes berechnet. Nun transponieren wir die Datenelemente und Dimensionen – drehen die Excel-Tabelle also um 90 Grad bzw. vertauschen die Zeilen und Spalten. Dann können wir jetzt ein Dendrogramm der Dimensionen in analoger Weise berechnen. Somit lässt sich ein Dendrogramm für die Datenelemente und für die Dimensionen ermitteln. Beide Dendrogramme lassen sich darstellen, farbcodiert, mit weiteren Eigenschaften über die Daten-Tabelle. Solche Darstellungen bezeichnen wir als Head Maps von Dendrogrammen, welche weitere Dateneigenschaften visuell codieren. Folgende Abbildung stellt eine solche Repräsentation dar:

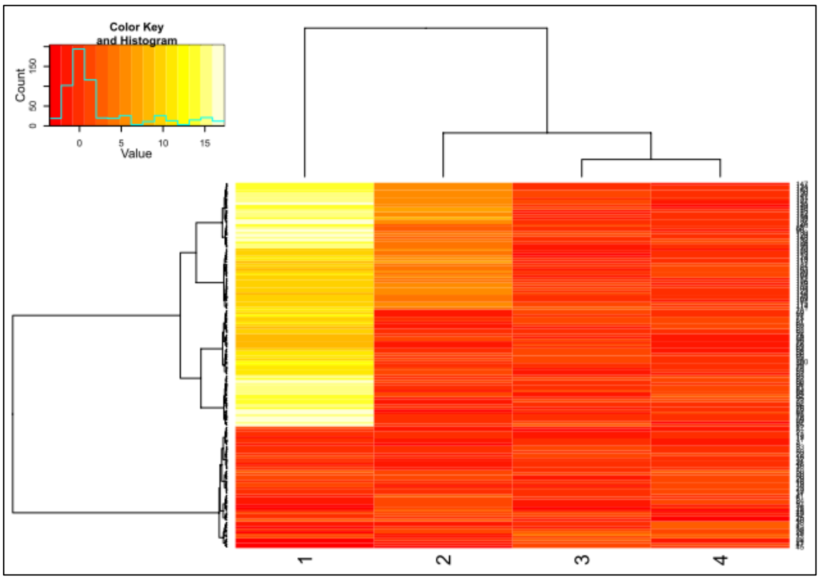


Abb.: Head Maps von Dendrogrammen

- -

Ziel des Projektes

Im Projekt wird ein System in Python entwickelt, um Dendrogramme zu erstellen, interaktiv zu manipulieren und zu visualisieren, mit dem Ziel dem Nutzer ein Werkzeug in die Hand zu geben um Struktureigenschaften hoch-dimensionaler Daten explorieren und analysieren zu können.

Referenzen

[1] https://www.youtube.com/watch?v=1jW9xlEtQao

[2] Lehrvideo: Emily Fox et al., Machine Learning: Clustering & Retrieval, University of Washington, 2016, https://www.coursera.org/lecture/ml-clustering-and-retrieval/the-dendrogram-MfcBU

[3] https://en.wikipedia.org/wiki/Dendrogram

[4] https://online.visual-paradigm.com/knowledge/business-design/what-is-dendrogram/

[5] NCSS Statistical Software, Hierarchical Clustering / Dendrograms, https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

[6] Constructing Overview + Detail Dendrogram-Matrix Views, Jin Chen, Alan M. MacEachren, and Donna J. Peuquet, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2010 https://www.researchgate.net/profile/Donna-Peuquet/publication/38015409_Constructing_Overview_Detail_Dendrogram-Matrix_Views/links/545f9a370cf27487b450a881/Constructing-Overview-Detail-Dendrogram-Matrix-Views.pdf

---

Angebot

Im Rahmen ihres *Praxisprojektes*, *Masterseminars*, *Masterprojektes*, ihrer *Bachelorarbeit*, *Masterarbeit* oder ähnlichen Studienleistungen - wie z. B. einem *interdisziplinären Digitalisierungsprojekt* - können sie gerne diese Aufgabe bearbeiten.

Melden Sie sich gerne bei mir unter: di.lehmenn@ostfalia.de

Aufgaben/Arbeitspakete

Das Projekt besteht aus klar voneinander abgrenzbaren Arbeitspaketen (AP):

- **AP1) Dashboard / User-Interface**
  - Eingabe: Datenadressen/Datenpfade/URLs für hoch-dimensionale Datensätze
  - Ausgabe: Visualisierung der Dendrogramme, Schnittebenen, Head Maps, Analyseinformationen, und Widgets für entsprechende Aktionen

  Beispiele für in diesem Paket umzusetzende Features:
    - Darstellen und Visualisierungen von Dendrogrammen
    - Editieren von Parametrisierungen
    - Editieren von Schnittebenen
    - Darstellen von Heat Maps
    - Zoom und „Detail on Demand" Funktionen
    - Darstellung von Metainformationen, wie Änderungsverläufe für eingegebenen Analyseergebnisse oder Nutzerinformationen etc.
    - Darstellen von Widgets für Actions etc.
    - …

- **AP2) Generieren von Dendrogrammen**
  - Eingabe: Datenstrukturen hoch-dimensionaler Datensätze
  - Ausgabe: Dendrogramme in geeigneten Datenstrukturen

  Beispiele für in diesem Paket umzusetzende Features:

        - Dendrogramme metrischer Daten durch Hierarchisches Clustering mit Single Linkage
        - Dendrogramme metrischer Daten durch Hierarchisches Clustering mit Centroid-basiertes Linkage
        - Dendrogramme metrischer Daten durch Hierarchisches Clustering mit Complete Linkage
        - Dendrogramme nicht-metrischer Daten
        - Dendrogramme hybrider Daten (metrischer und nicht-metrischer Daten)
        - Bereitstellen unterschiedlicher Distanzmaße für die Erzeugung von Dendrogrammen
        - …

- **AP 3) Generieren von 2D Einbettungen zur vergleichenden Darstellung nD Daten**
    - Eingabe: Datenstrukturen hoch-dimensionaler Datensätze
    - Ausgabe: 2D Einbettungen

    Beispiele für in diesem Paket umzusetzende Features:
        - Einfache Einbettungsansätze wie Star Koordinaten bzw. PCA-basierte Ansätze
        - „Linking & Brushing“ zw. Dendrogrammen und den Darstellungen in den Einbettungen
        - …

- **AP 4) Generieren von Head-Maps über Dendrogramme**
    - Eingabe: Datenstrukturen hoch-dimensionaler Datensätze
    - Ausgabe: interaktive Head Maps über Dendrogramme

    Beispiele für in diesem Paket umzusetzende Features:
        - Einfache Einbettungsansätze wie Star Koordinaten bzw. PCA-basierte Ansätze
        - „Linking & Brushing“ zw. Dendrogrammen und den Darstellungen in den Einbettungen
        - …

- **AP 5) Generieren von Schnittebenen über Dendrogramme**
    - Eingabe: Dendrogramme, Referenzen zu hoch-dimensionaler Datensätze
    - Ausgabe: Modell von Schnittebenen, resultierende Clusterisierung über Schnittebene

    Beispiele für in diesem Paket umzusetzende Features:
        - Erzeugen von vertikalen Schnittebenen
        - Erzeugen von linearen Schnittebenen (auch schräge also)
        - Erzeugen von kurven-esque Schnittebenen
        - Erzeugen von freihand-basierten Schnittebenen
        - Erzeugen einer Clusterisierung basieren auf einem Schnittebenenmodell
        - …

- **AP 6) Daten-Accessor**
    - Eingabe: Objekte und Datenadressen
    - Ausgabe: Speicher- und Lade-Operationen

    Beispiele für in diesem Paket umzusetzende Features:
        - Laden/Speichern von Dendrogrammen und HeatMaps konkreter Datensätze
        - Laden/Speichern von Datensätze in denen kontrafaktische Analysen durchgeführt werden sollen
        - Laden/Speichern von Clusterisierungen aus konkreten Schnittebenen-Modellen konkreter Datensätze
        - …

Zu den Arbeitspaketen hinzu kommen notwendige Recherchetätigkeiten, Make-Or-Buy-Entscheidungen, Beachtung von Lizenzfragen, Aspekte der Continuous Integration und des Code-Managements, Dokumentationsaufgaben, Fragen zum Aufsetzen/Deployment und der Migration von Entwicklersystemen (und zum Projektmanagement), wie sie in der Softwareentwicklung üblich sind.

Im Rahmen einer Projektbearbeitung wird nicht erwartet, dass unmittelbar alle APs bearbeitet werden können. Je nach Umfang Ihrer zu erbringenden Studierendenleistung können Teilaspekte einzelner APs bearbeitet werden. Den konkreten, jeweiligen Umfang stimmen wir im Vorfeld gerne gemeinsam ab.

Vorkenntnisse

Es ist hilfreich – aber keine Voraussetzung – wenn Sie Vorkenntnisse/Interesse mitbringen in

- Softwareentwicklung
- Python

- Grundkonzepte des Maschinellen Lernens
- Data Engineering

Vorarbeiten/Voraussetzungen

Kurzeinführung in Python:

https://www.youtube.com/watch?v=x_kYpwi1L1k

OnboardingProzess

Alle wichtigen Zugänge einrichten um die Arbeit am Projekt aufnehmen zu können:

http://46.38.235.241/webpage/dirkfiles/misc/onboarding/OnboardingProzess.pdf

---

Organisatorisches

Projektmanagement per Scrum in Trello:

https://trello.com/b/ftwh06og/dendrogramm-explorer

Projektcode-Management per GitHub:

https://github.com/DirkJLehmann/DendrogrammExplorer.git

Projektdokumentations-Management:

https://docs.google.com/document/d/1-gTOLpSzLoYRFqjNyUR7aimBSUqn3fGsjzjlgpi8sJ0/edit?usp=sharing

---

Bei Interesse melden Sie sich bitte unter:

Prof. Dr.-Ing.habil.
Dirk Joachim Lehmann
Data Science in IoT
Fakultät für Informatik
di.lehmann@ostfalia.de