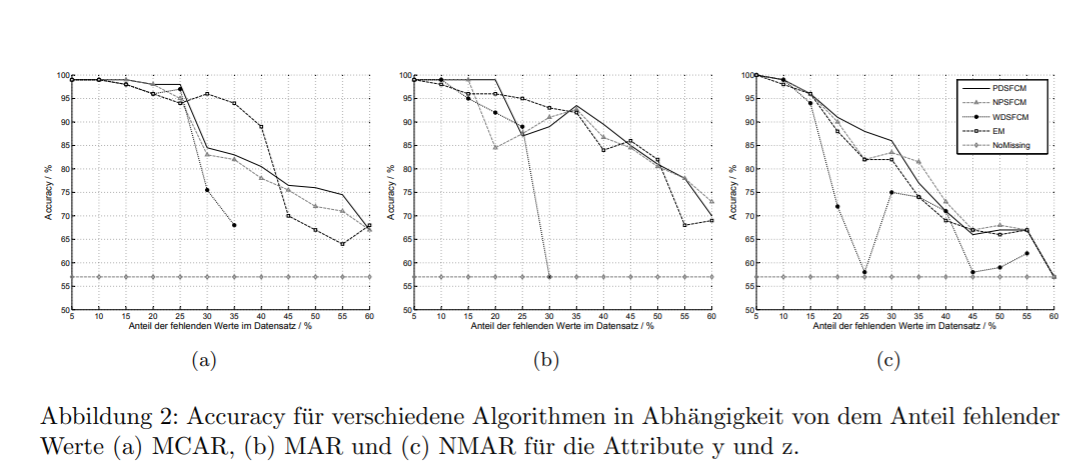
**Recherche Cluster-Algorithmen**

* Datenaufbereitung:
  + Daten skalieren
  + Kategorisch ordinal -> numerisch
  + Nominal kategorisch -> one-hot encoding -> numerisch
* Umgang mit fehlenden Werten
  + warum fehlen die Werte?
    - Systematisch fehlend? NMAR
    - Unsystematisch fehlend -> zufällig? MAR
  + Eliminieren (Münzen oder Merkmale) >>macht vllt in Einzelfällen Sinn
  + Imputation: man könnte z.B. Einzelwerte vom Typ auf die Münze übertragen (finde ich aber kritisch)
  + Algorithmus anpassen
    - nur vorhandene Werte einbeziehen
      * **whole-data-strategy:** erst alle vollständig vorhandenen Datensätze clustern, dann unvollständige (partielle Distanz)
      * **partial distance strategy**: partielle Distanzen verwenden
    - Werte der Clusterzentren einbeziehen
      * **nearest prototype strategy:** fehlenden Attributwerte durch Werte des Centroiden ersetzen, in jedem Iterationsschritt
* verschiedene testen und aus Domainenperspektive anhand konkreter bspe schauen, was im Ergebnis mehr Sinn macht.
* Library?



*Abbildung 1* [*Quelle*](http://rosdok.uni-rostock.de/file/rosdok_document_0000000157/rosdok_derivate_0000004060/gvd2009_2.H.21_Himmelspach.pdf#:~:text=Die%20fehlenden%20Werte%20k%C3%B6nnen%20einer,vollst%C3%A4ndigen%20Daten%2D%20matrizen%20Analysen%20durchzuf%C3%BChren.)

* Cluster festlegen?
* Min. Distanz festlegen?
* Iterativer Prozess: Anpassung des Modells bis Cluster auch aus domainenpespektive passen

|  |  |
| --- | --- |
| **kMeans (partitioning clustering)**  *Gower distance: geeignet für mixed data*  *Varianten: PAM, CLARA* | |
| + für große Datensets: CLARA statt PAM (effizienter)  + scikit learn | - ggf. sensitive to outliers (auch abhängig von Ähnlichkeitsmaß) >> kModes/ PAM besser  - #Cluster vorher festlegen  - Wahl Ähnlichkeitsmaß  - Daten numerisch  - ineffektiv für hohe #Features -> sparse  - Schwierigkeiten bei nicht-konvexen Clustern |

|  |  |
| --- | --- |
| **Fuzzy kMeans**  Für unvollständige Daten: [Paper](https://www.researchgate.net/publication/5607056_Fuzzy_C-means_clustering_of_incomplete_data) | |
| + s. kMeans | - s kMeans |

|  |  |
| --- | --- |
| **DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ->** OPTICS  Gebiete mit hoher Datenpunktdichte werden zu Clustern zusammengefasst   1. Benenne Datenpunkte als Kern-, Rand- oder Rauschpunkte. 2. Lösche alle Rauschpunkte. 3. Verbinde Kernpunkte, die innerhalb einer ε-Kugel liegen, durch eine Kante. 4. Eine Menge verbundener Kernpunkte bilden ein separates Cluster. 5. Weise jeden Randpunkt dem Cluster eines benachbarten Kernpunkts zu. | |
| + erkennt Cluster beliebiger Form  + filtert Rauschen gut  >> gute Ergänzung zu kMeans  + scikit learn | - Parameter festlegen: Radius für Dichte und MinPts (wie viele Punkte in Radius liegen müssen)  - Cluster unterschiedlicher Dichte erkennen schwierig  - Dichte bei hochdimensionalen Daten schwierig zu definieren |

|  |  |
| --- | --- |
| **Gaussian Mixture Models with Expectation-Maximation -> model based clustering**  data are considered as coming from a mixture of underlying probability distributions  <https://cran.r-project.org/web/packages/mclust/index.html> | |
| + identify observations with high or low cluster uncertainty | - Ergebnisse stark abhängig von angenommener Verteilung  - computational performance für hochdimensionale Daten schlecht |

|  |  |
| --- | --- |
| **(Agglomerative) Hierarchical Clustering (AGNES)**  Jede Beobachtung zunächst eigenes Cluster -> verschmelzen auf Basis von Ähnlichkeitsfunktion(en) bis 1 großes Cluster (gegengesetztes Vorgehen: divise hierarchical clustering).  Output: Dendogram  [Konzept](https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec#:~:text=Space%20complexity%3A%20The%20space%20required,the%20number%20of%20data%20points.) | |
| + Cluster vorher nicht festzulegen  + Darstellung (Dendogramm)  + scikit learn | - Zeitkomplexität O(n^3), Speicherkomplexität O(n^2) – nicht geeignet für große Datenmengen  - Wahl a) Ähnlichkeitsmaß zw. Punkten b) Ähnlichkeit zw. Clustern -> Vor/Nachteile verschiedener Methoden haben starken Einfluss auf Ergebnisse |

|  |  |
| --- | --- |
| **Auto Encoder, z.B. Boltzman machines**  Clustering is difficult to do in high dimensions because the distance between most pairs of points is similar. Using an autoencoder lets you re-represent high dimensional points in a lower-dimensional space. It doesn't do clustering per se - but it is a useful *preprocessing* step for a secondary clustering step. | |
| + | - |

|  |  |
| --- | --- |
| **Mean Shift Clustering** | |
| + scikit learn | - |

Anomalie Detection

**Recherche Cluster Evaluation**

Source**:** [**https://arxiv.org/pdf/1905.05667.pdf**](https://arxiv.org/pdf/1905.05667.pdf)

**PRE CLUSTERING**.: Does the data exhibits random behavior or not?

* Monte Carlo analysis
* Bootstrapping

**POST CLUSTERING**. / partitional algorithms / internal validation

* metrics based on the proximity matrix  
  Unfortunately, the mere construction of the whole proximity matrix is computationally expensive, O(n 2 ), and this validation method cannot be used without sampling for large data sets.
* metrics of cohesion (intra cluster), e.g. cluster SSE
* metrics of separation (inter cluster), e.g. SSB
  + Calisnki-Harabasz coefficient (variance ratio criterion) -> SSE, SSB
  + silhouette coefficient -> Positive values indicate a high separation between clusters. Negative values are an indication that the clusters are mixed with each other, zero, it is an an indication that the data are uniformly distributed throughout the Euclidean space
  + Dunn index
  + Xie-Beni (fuzzy clustering)
  + Ball-Hall index
  + Hartigan index
  + Xu coefficient t

YES: kmeans  
NO: density based algorithms (DBSCAN,..)

-

**POST CLUSTERING**. / hierarchical algorithms / internal validation

* Cophenetic Correlation Coefficient (CPCC)
* Hubert Statistic

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Sonstiges

* Scaling Up Schema Discovery for RDF Datasets: <https://hal.archives-ouvertes.fr/hal-01757028/document>

<https://openreview.net/pdf?id=f42oCMVdpOj>

* DBSCAN für RDF? <https://github.com/BOUHAMOUM/SC-DBSCAN>