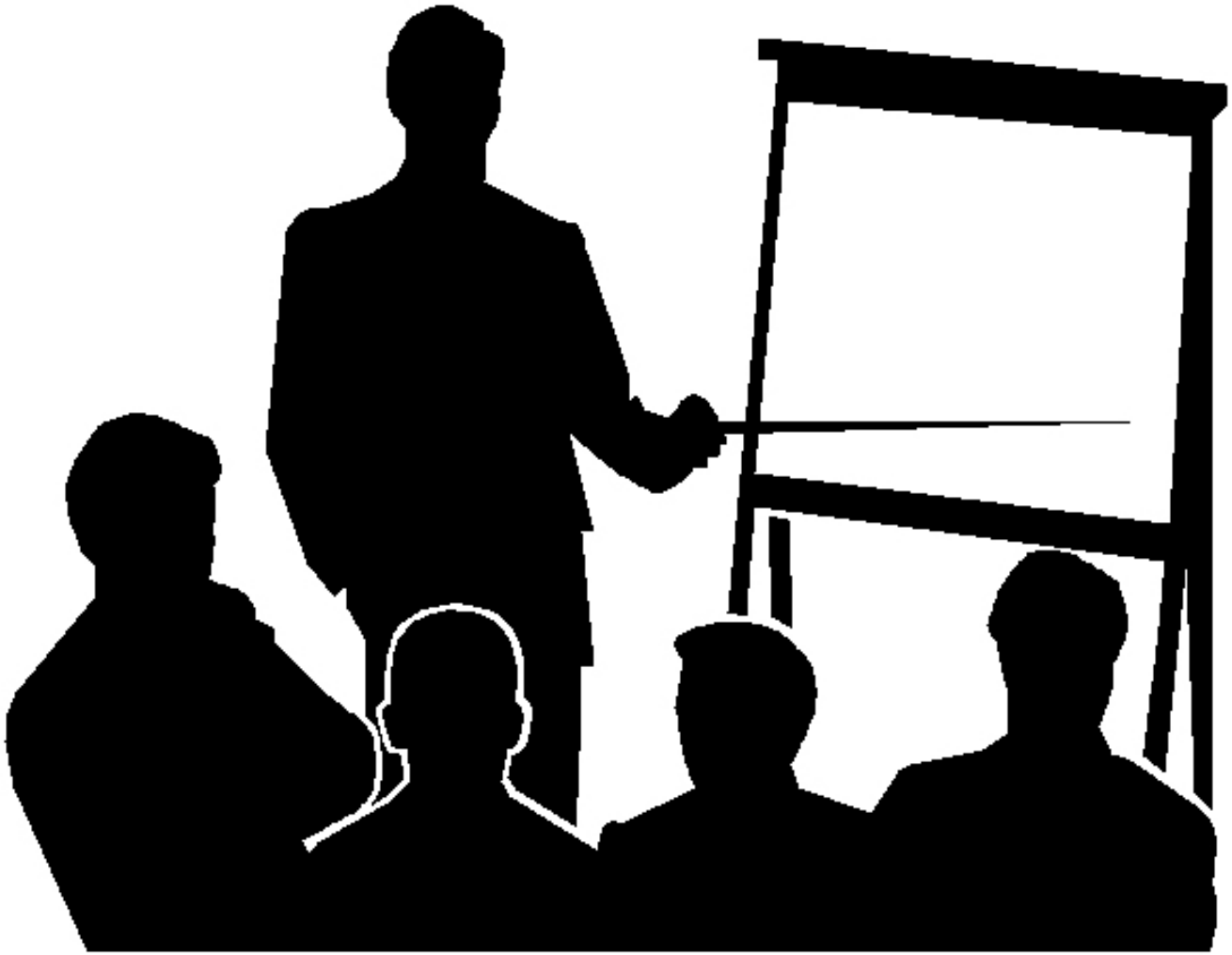
Clustering

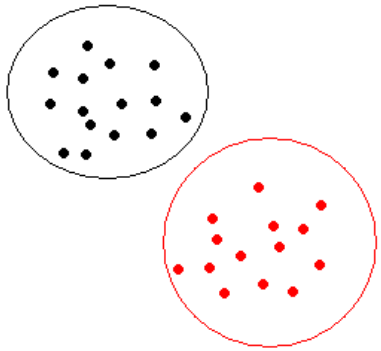
Motivation



Quelle: http://www.ha-w.de/media/schulung01.jpg

Was ist Clustering

- Idee: Gruppierung von Objekten so, dass:
  - Innerhalb einer Gruppe sollen die Objekte möglichst ähnlich sein
  - Zwischen verschiedenen Gruppen sollen möglichst große Unterschiede bestehen



Anwendungen

- Datenreduktion
- Mustererkennung
- Medizin
- Marketing
- Bildverarbeitung
- Biologie
- Dokumentengruppierung
- …

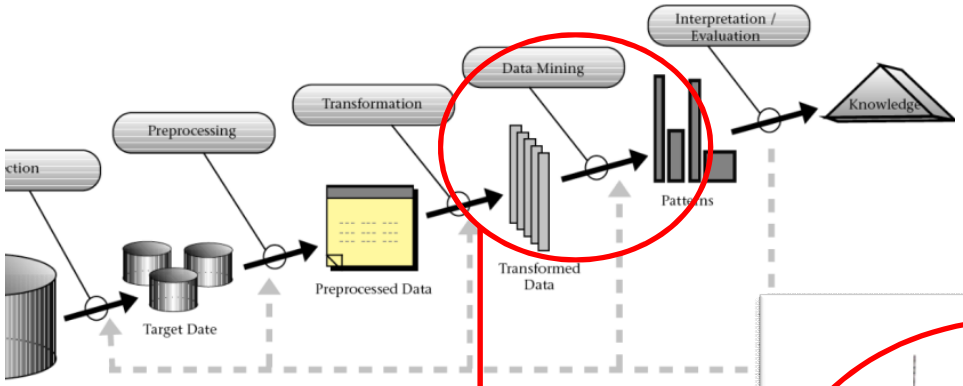
Ziele des Clustering

- Entwicklung eines Klassifikators
- Entwicklung von Schemen für das Gruppieren von Daten.
- Aufstellen von Hypothesen
- Testen von Hypothesen

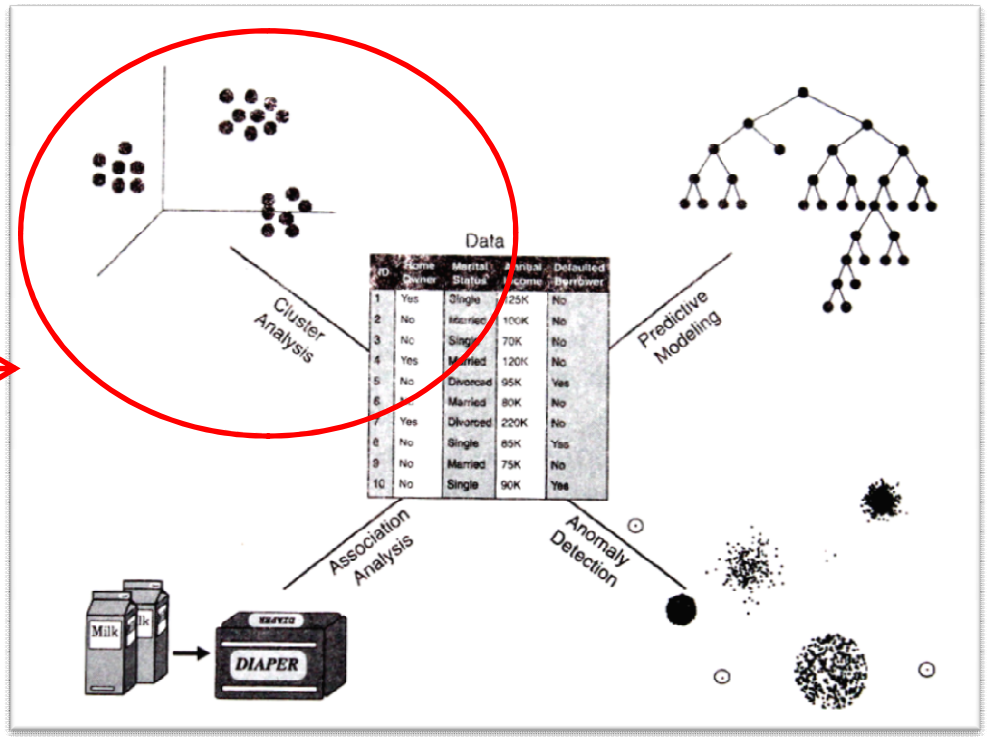
Anforderungen an Clusterverfahren

- Skalierbarkeit
- Datentypenunabhängigkeit
- Formunabhängigkeit
- möglichst wenige Eingabeparameter
- Ausreißer –und Rauschbehandlung
- Unabhängig von der Reihenfolge der Daten
- Arbeiten mit Daten höherer Dimension
- Interpretierbar/Nutzbarkeit
- Informationen über mögliche Clusteranzahl

Einordnung in den KDD-Prozess

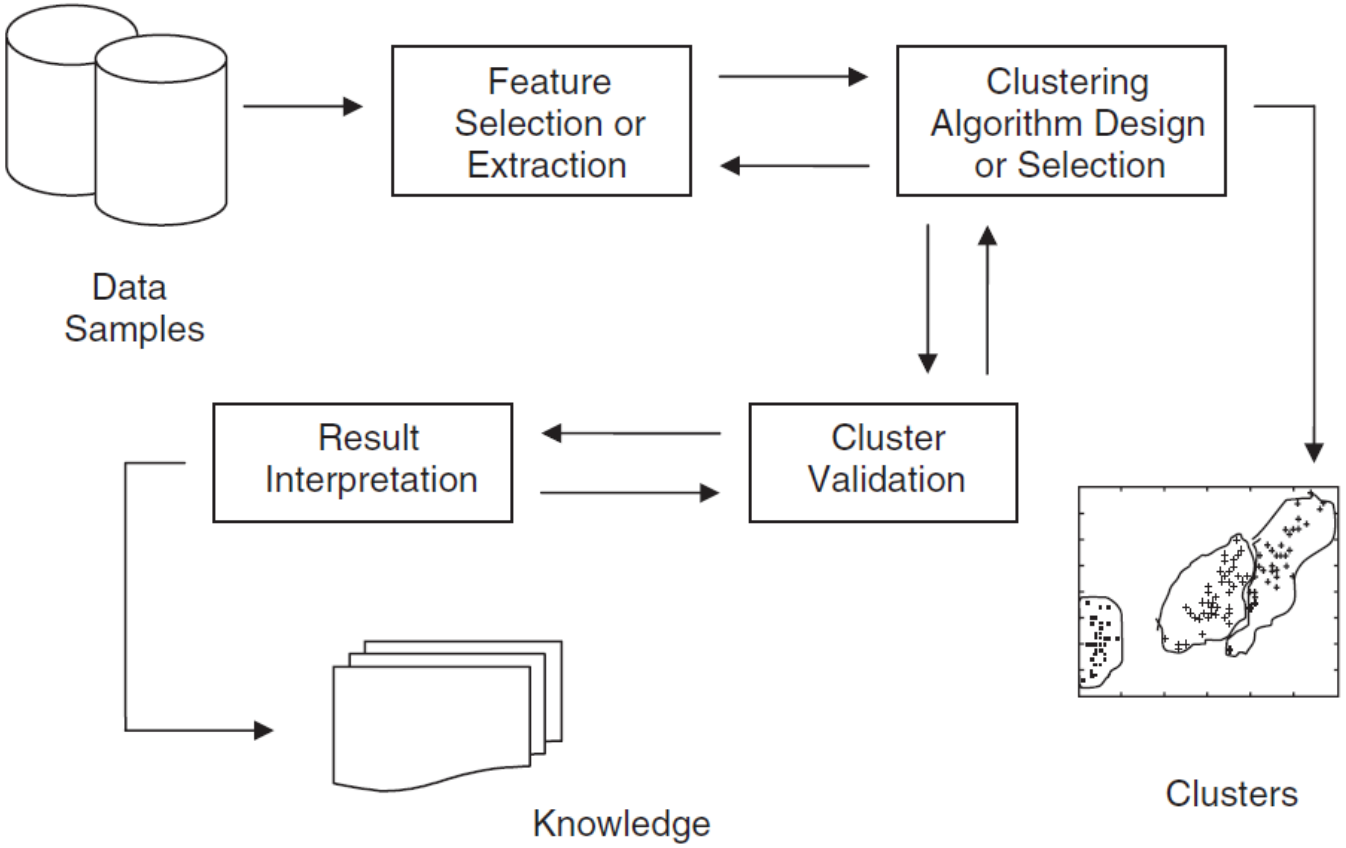


[Fayyad]



[Tan 2006]

Clusteranalyse



[Xu 2009]

Ähnlichkeit –und Distanzfunktionen

Datengrundlage

- Menge von *n* Datenmustern $D = \{x_1 \dots x_n\}$
- Jedes Musterdatum x_j hat *m* Merkmale $x_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jm} \end{pmatrix}$
  - kategorische Merkmale (nominale,ordinale)
  - numerische Merkmale

Erinnerung: Fehlende und falsche Attribute in einzelnen Mustern!

Ähnlichkeitsfunktion

Definition einer Ähnlichkeitsmetrik

$$s(x_i, x_j) = s(x_j, x_i)$$

(Symmetrie)

$$s(x_i, x_i) = 1$$

(Reflexivität)

$$0 \leq s(x_i, x_j) \leq 1$$

(Positiv)

Wertebereich [0..1]
0: keine Ähnlichkeit
1: hohe Ähnlichkeit

Distanzfunktion

Definition einer Distanzmetrik

$d(x_i, x_j) = d(x_j, x_i)$ (Symmetrie)

$d(x_i, x_i) = 0$ (Reflexivität)

$d(x_i, x_j) \geq 0$ (Positiv)

$d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$ (Dreiecksungl.)

Wertebereich [0..unendl.]

0: Identität

Beispiele

Minkowski-Distanz

(r=2,euklidischer Abstand)

$$d_r(x_i, x_j) = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}$$

Simple Matching Koeffizient

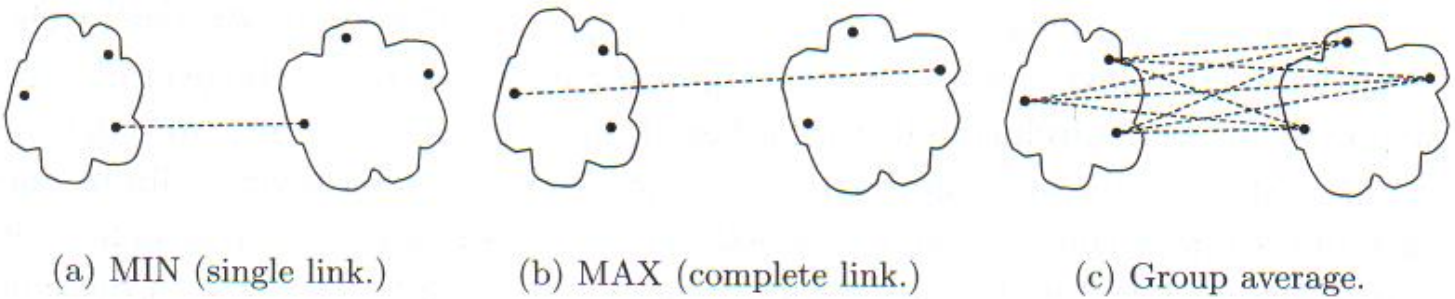
$$s(x, y) = \frac{f_{11} + f_{00}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

Kosinus-Ähnlichkeit

$$s(x, y) = \cos(\varphi) = \frac{< x, y >}{||x|| \, ||y||}$$

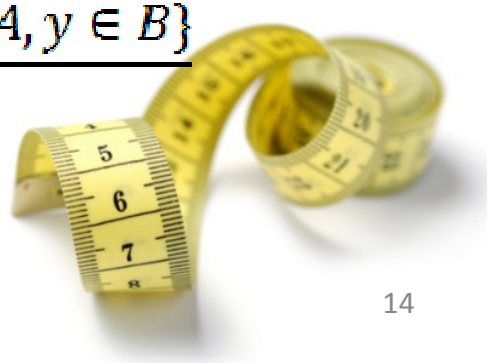
Transformation von Ähnlichkeiten in Distanzen möglich: *d= 1-s*

Abstände zwischen Clustern



[Tan 2006]

- Single link : $D(A,B) = \min\{d(x,y) \mid x \in A, y \in B\}$
- Complete link : $D(A,B) = \max\{d(x,y) \mid x \in A, y \in B\}$
- Average link : $D(A,B) = \frac{\sum\{d(x,y) \mid x \in A, y \in B\}}{|A||B|}$

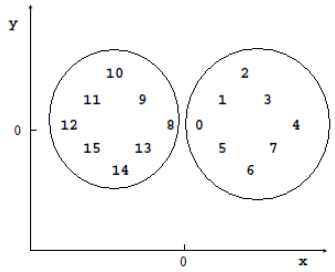


Clusteringverfahren

Partitionierendes Clustering

– Gesamtdatenmenge wird in Teilmengen zerlegt

– Jedes Objekt wird genau einem Cluster (Gruppe) zugeordnet

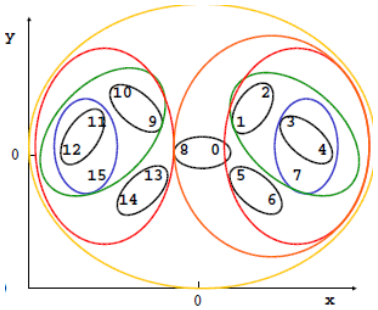


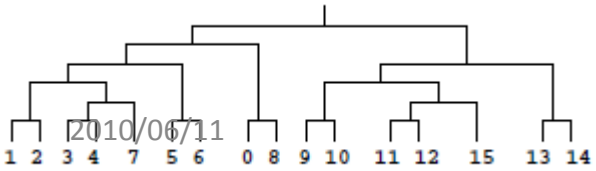
Hierarchisches Clustering

– Cluster in Clustern

– Anordnung in einem Baum

- Wurzel entspricht kompletten Datensatz
- Blätter: einzelnes Objekt
- Knoten: Zusammenfassung von Clustern





Vorgestellte Verfahren

- Partitionierendes Clustering
  - K-Means Verfahren
  - EM-Verfahren
  - (DBSCAN)
- Hierarchisches Clustering
  - Agglomerierendes Verfahren

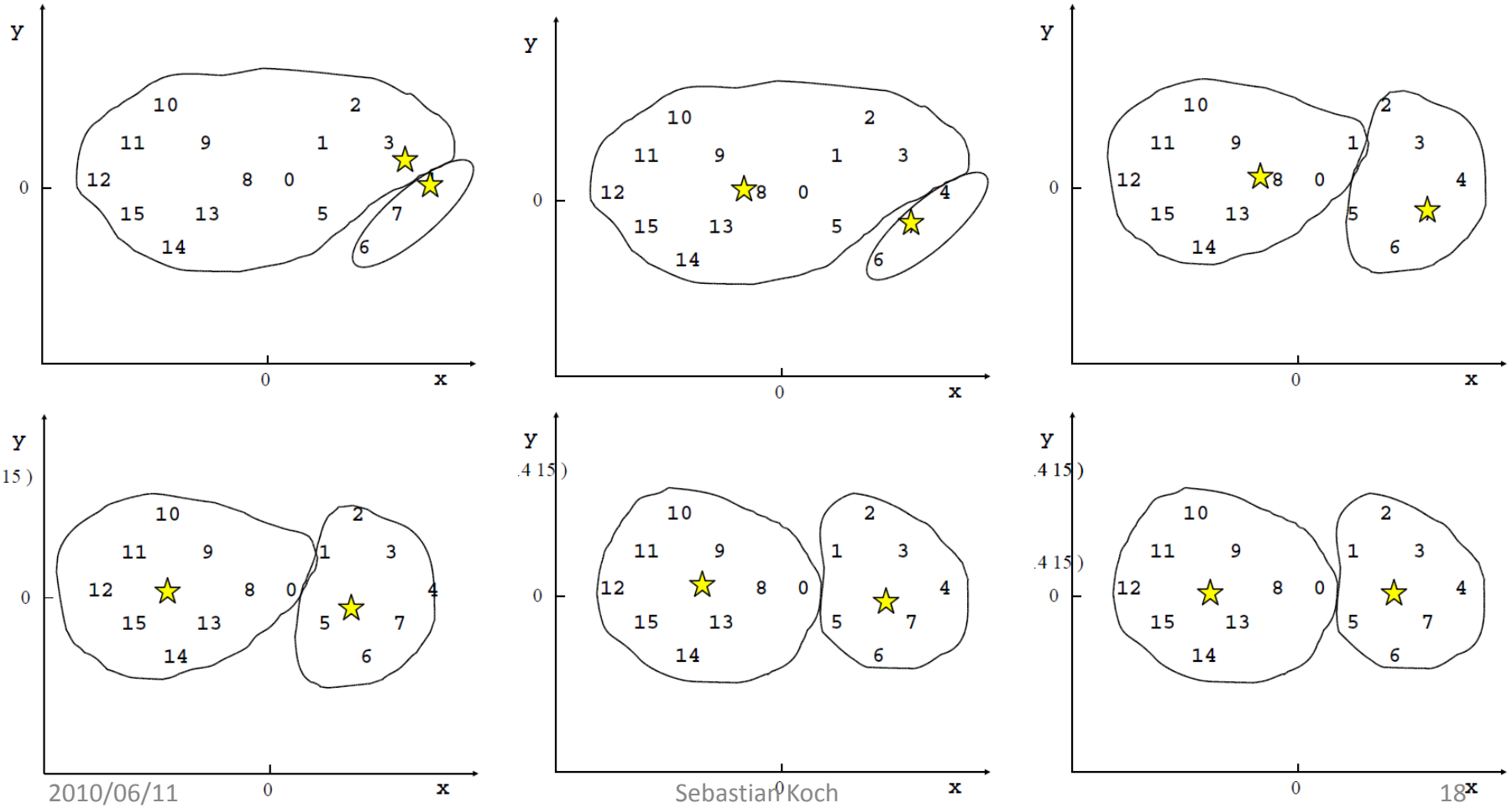
k-means-Algorithmus

- partitionierendes Verfahren
- einfach und schnell
- bekanntester Algorithmus
- prototypbasierend (Zentren)
- teilt die Daten in k Cluster ein, k vorgegeben
- benutzt Optimalitätskriterium

Wähle k Clusterzentren
repeat
-weise jedes Objekt dem Zentrum zu, zu dem es am nächsten liegt
-berechne neue Zentren
until Zentren ändern sich nicht mehr

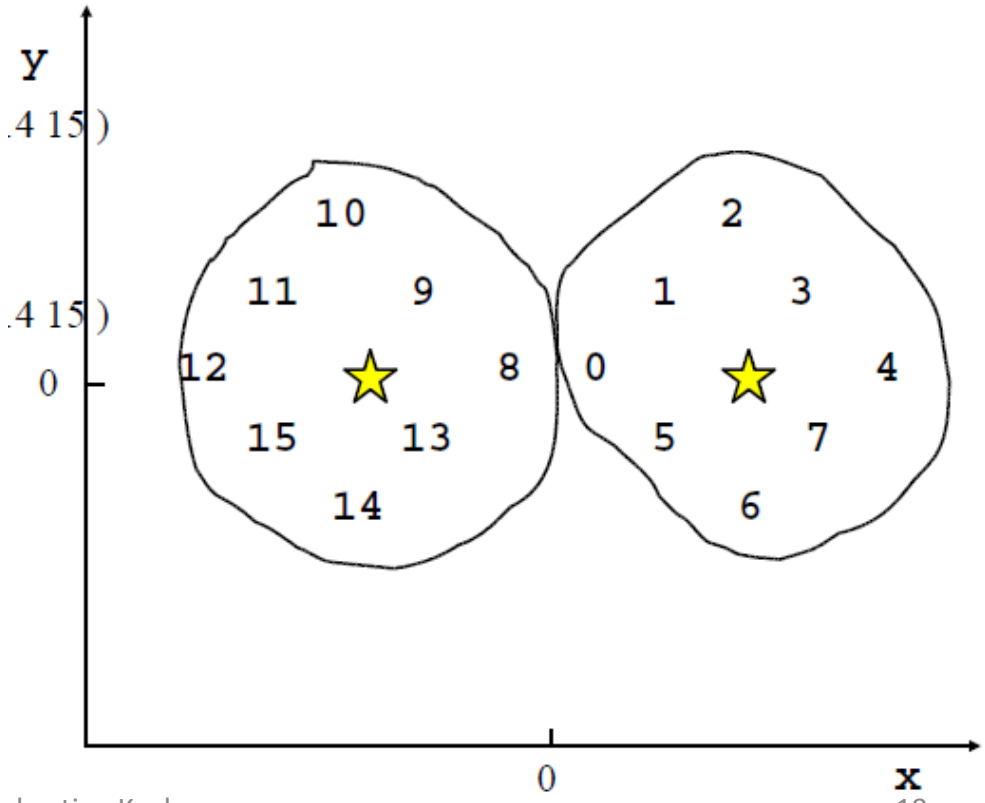
Beispiel

nach [J. Fürnkranz]



Beispiel – Ergebnis und Bemerkungen

- schlechte Wahl der Initialisierungszentren
- leere Cluster
- Ausreißer
- Clustertypen/ Clusterformen



EM-Verfahren

- Verallgemeinerung des k-Means
- Keine exklusive Zuordnung von Objekten zu einem Cluster
- Clustering anhand eines statistischen Modells
- Annahme: Daten sind Zufallsgrößen
- Gesucht: Parameter des Modells

EM-Algorithmus

- Bestimme Anfangsmodellparameter
- repeat
  - E-Schritt
    - Bestimme für jedes Objekt die Wahrscheinlichkeit, dass es zu jeder Verteilung gehört.
  - M-Schritt
    - Schätze anhand der Wahrscheinlichkeiten aus dem E-Schritt die Parameter des Modells (Maximum Likeliehood)
- until Modellparameter ändern sich nicht mehr

KM: Ordne Objekte einem Cluster zu

KM: Berechne neue Zentren für die Cluster

Bemerkung

- Zugrundeliegende Verteilungen muss bekannt sein.
- Große Menge an Musterdaten nötig
- Anzahl der Cluster muss bestimmt werden
- Cluster mit unterschiedlichen Dichten werden berücksichtigt
- Modell vereinfacht die Struktur der Daten
- Viele Prozesse haben eine Verteilung

Hierarchisches Clustering

- Kann unterteilt werden in agglomerierendes (bottom-up) und unterteilendes (top-down) Clustering
- Bottom-up
  - Jeder Datensatz ist ein eigener Cluster
  - In jedem Schritt werden zwei Cluster verschmolzen bis nur noch ein Cluster vorhanden ist
- Top-down
  - Alle Datensätze in einem Cluster
  - Sukzessives teilen eines Clusters, bis jeder Datensatz in einem eigenen Cluster steht

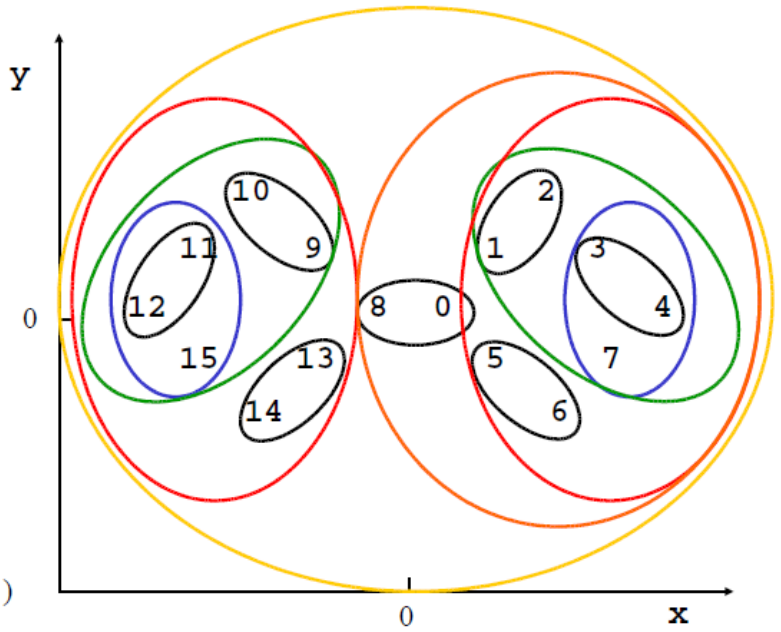
Agglomerativer Algorithmus

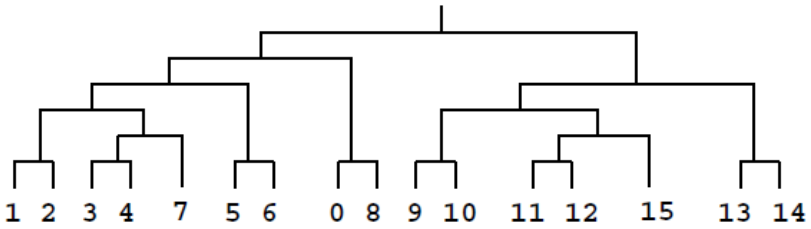
- Jedes Objekt wird einem eigenen Cluster zugeordnet
- Bestimme die Distanzmatrix aller Objekte zueinander mit D(A,B)
- repeat
  - Vereine die beiden nächstliegenden Cluster
  - Berechne die neue Distanzmatrix
- until nur noch ein Cluster übrig

Beispiel

Bottom-up clustering (average-link):

min distance = 2.00000 (8) (0)
min distance = 2.82843 (2) (1)
min distance = 2.82843 (4) (3)
min distance = 2.82843 (6) (5)
min distance = 2.82843 (10) (9)
min distance = 2.82843 (12) (11)
min distance = 2.82843 (14) (13)
min distance = 3.16228 (7) (3 4)
min distance = 3.16228 (15) (11 12)
min distance = 4.73756 (3 4 7) (1 2)
min distance = 4.73756 (11 12 15) (9 10)
min distance = 4.74131 (1 2 3 4 7) (5 6)
min distance = 4.74131 (9 10 11 12 15) (13 14)
min distance = 5.57143 (0 8) (5 6 1 2 3 4 7)
min distance = 9.90476 (13 14 9 10 11 12 15) (5 6 1 2 3 4 7 0 8)



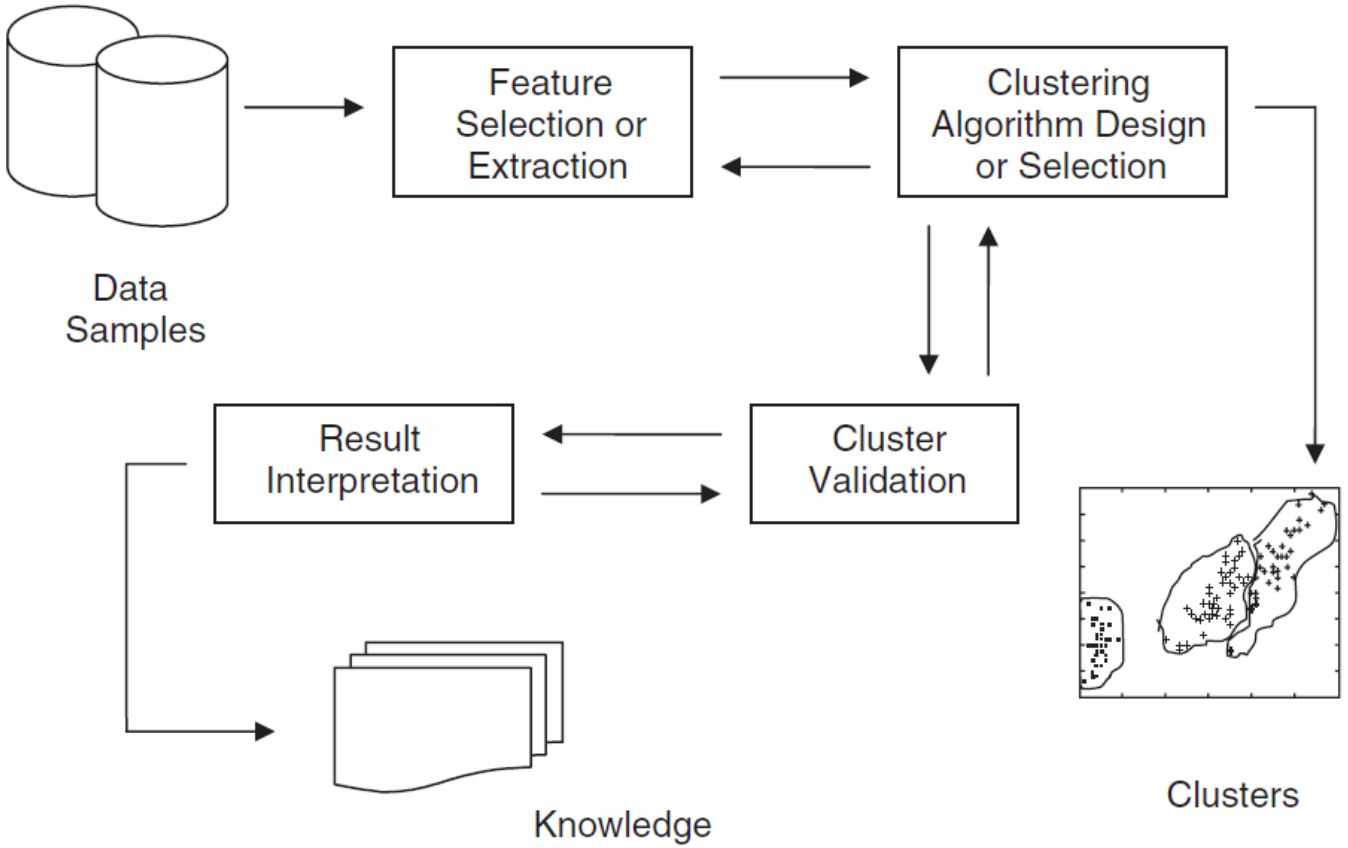


[J. Fürnkranz]

Bermerkungen

- Kein globales Optimalitätskriterium
- Zusammenfassung von Clustern ist endgültig
- Schlechte Laufzeit

Zusammenfassung



Quellen

- Fayyad (2007): kdd-process.png (PNG-Grafik, 1537x705 Pixel) - Skaliert (90%). Online verfügbar unter http://www.infovis-wiki.net/images/4/4d/Fayyad96kdd-process.png, zuletzt aktualisiert am 09.10.2007, zuletzt geprüft am 10.04.2010.
- J. Fürnkranz, G. Widmer (2005). Online verfügbar unter http://www.ke.tu-darmstadt.de/lehre/archiv/ws0405/mldm/clustering.pdf, zuletzt aktualisiert am 12.01.2005, zuletzt geprüft am 10.04.2010.
- Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2006): Introduction to data mining. Boston, Mass.: Pearson/Addison-Wesley.
- Xu, Rui; Wunsch, Donald C. (2009): Clustering. Oxford: IEEE Press; Wiley.