Wissensentdeckung – Entdeckungswissenschaft

Datensegmentierung mittels Clusteranalyse

[GM98]

Martin Kleefeld

1 Einleitung

- *Data Mining:*
  - Extraktion von „verborgenem Wissen“ aus Datensammlungen
  - wichtig: Unterteilung der Daten in Teilmengen/Klassen (Segmentierung) $\rightarrow$ automatisch: Clusteranalyse
  - Lernverfahren
    - überwacht: Klassenzugehörigkeit bekannt
    - unüberwacht: Klassenzugehörigkeit unbekannt, Anzahl Klassen unbekannt
- *Clusteranalyse:*
  - statistische Verfahren
  - objektive und automatische Einteilung einer Objektmenge in kleinere Teilmengen

1.1 Zielsetzung der Clusteranalyse

- Finden von Strukturen in hochdimensionalen Räumen

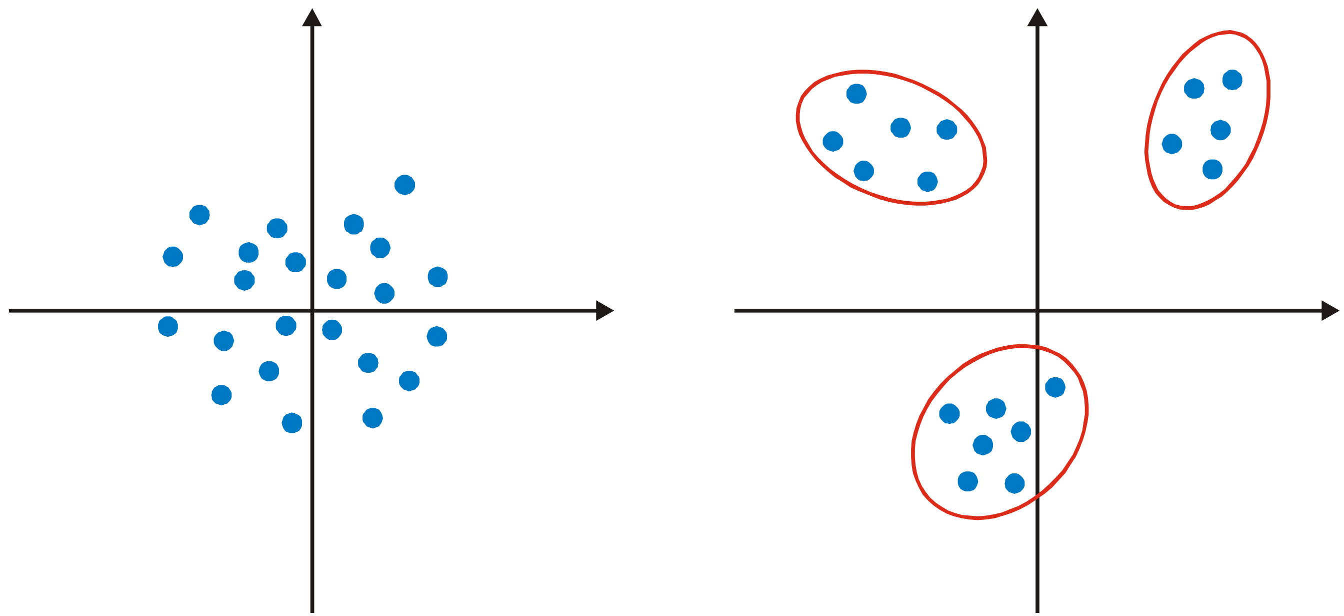
♦ Anwendungsgebiete:

- traditionell: Biologie, Chemie, Soziologie, Psychologie, BWL
- neu: Data Mining

♦ Voraussetzungen:

- Beobachtungen (Objekte) liegen in M-dimensionalem Merkmalsraum
- Merkmalsraum kann mit Distanzmaß ausgestattet werden
- Struktur von Klassen existiert

- Ziel:
  - automatisches Finden einer Gruppierung
  - Widerspiegelung der Struktur durch Gruppierung
  - *Homogenität* innerhalb einer Gruppe
  - *Heterogenität* zwischen Gruppen



2 Ausgangsdaten

- ♦ Format:
  - Datenmatrix:
    - ∗ Zeilen: Beobachtungen
    - ∗ Spalten: Variablen (Merkmale, Eigenschaften)

2.1 Datentypen

- ♦ *qualitativ (symbolisch, nominal):*
  - endliche Anzahl von verschiedenen Merkmalsausprägungen
  - Ausprägungen sind gleichwertig
    $\Rightarrow$ nur Aussagen über Gleichheit

- ♦ *quantitativ:*
    - ∗ (mindestens) Ordnungsrelation sinnvoll
- • ordinalskaliert:
    - ∗ Rangfolge $\Rightarrow$ Vergleichsoperatoren
- • metrisch:
    - ∗ intervallskaliert:
        - – Wertebereich: $\mathbb{R}$
        - – kein Bezugs-/Nullpunkt
    - ∗ verhältnisskaliert:
        - – fester Bezugspunkt $\Rightarrow$ Verhältnisbildung

♦ *binär:*

- genau zwei Merkmalsausprägungen
- Kodierung:
  - qualitative Daten: beliebig
  - quantitative Daten: einheitlich

2.2 Datenqualität

♦ *Fehlwerte:*

- ungültige Werte in einigen Objekten
- können auch Informationen enthalten (z.B. „unter der Nachweisgrenze“)
- möglicher Ansatz: betreffende Objekte bzw. Merkmale weglassen
- besser: geeignete Berücksichtigung

3 Distanz- und Ähnlichkeitsmaße

- gegensätzliche Beziehung
- große Bedeutung für Qualität des Ergebnisses

♦ Definition:

- *Distanz:* Abbildung d: $X \times X \to \mathbb{R}_+$, so daß $\forall\ \mathbf{x}_i, \mathbf{x}_j \in X$ gilt:
  - ∗ $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ (Positivität)
  - ∗ $d(\mathbf{x}_i, \mathbf{x}_j) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
  - ∗ $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$ (Symmetrie)
- *Metrik:* zusätzlich gilt $\forall\ \mathbf{x}_k \in X$:
  - ∗ $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$ (Dreiecksungleichung)

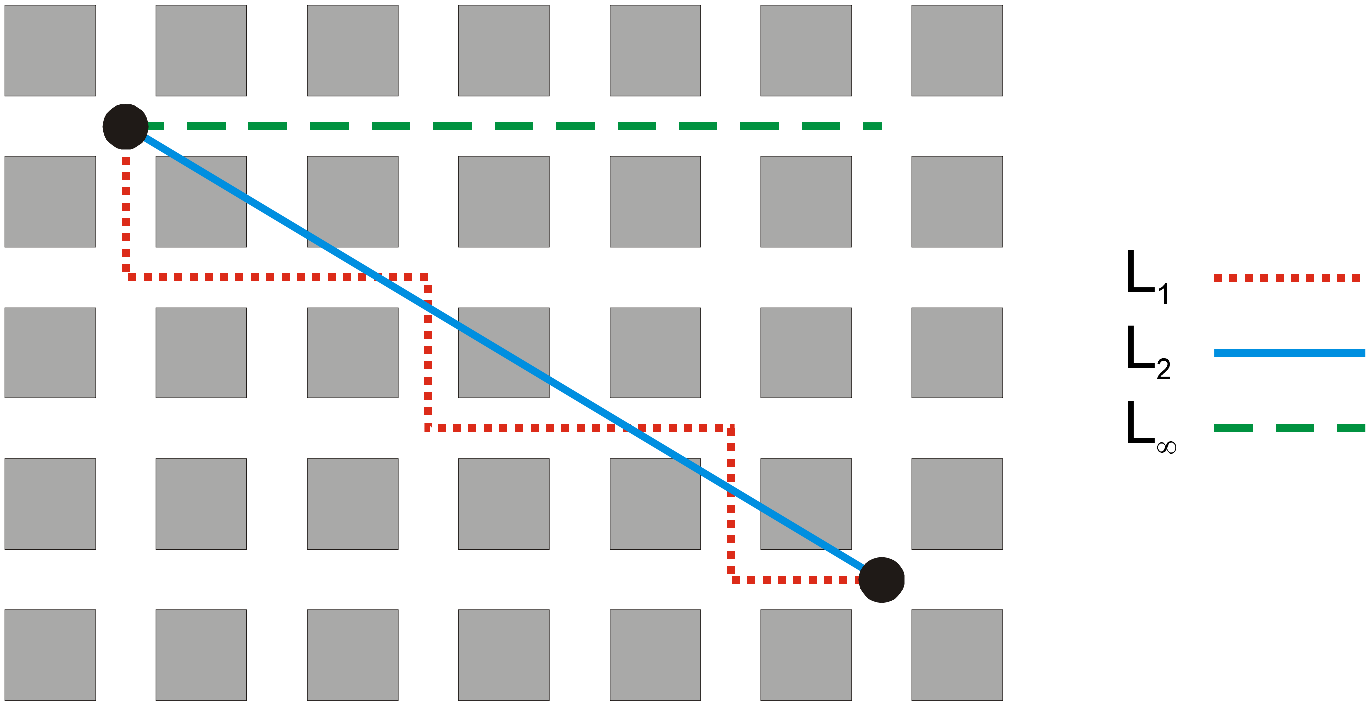
- *Normierung*
  - notwendig bei verschiedenen Wertebereichen
  - gewährleistet Vergleichbarkeit der Merkmale
- *Gewichtung*
  - vorgegeben oder automatisch ermittelt
  - Veränderung des Einflusses einzelner
    - Objekte *(Zeilengewichte)* bzw.
    - Merkmale *(Spaltengewichte)*

3.1 Abstandsmaße für metrische Daten

(zwischen Beobachtungsvektoren $\mathbf{x}_i$ und $\mathbf{x}_j$)

- *Minkowski-Metrik (L_r-Abstand):*
  - *euklidischer Abstand* (r = 2): $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{M} (\mathbf{x}_{il} - \mathbf{x}_{jl})^2} = ||\mathbf{x}_i - \mathbf{x}_j||$
    - * invariant bzgl.: Verschiebung, Drehung, Spiegelung
    - * nicht invariant bzgl.: Skalierung
  - *Absolutabstand / City-Block-Metrik* (r = 1): $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{M} |\mathbf{x}_{il} - \mathbf{x}_{jl}|$
    - * invariant bzgl.: Verschiebung
    - * nicht invariant bzgl.: Drehung, Spiegelung, Skalierung

- *Maximumabstand / Supremum-Metrik* (r = ∞):
  $d(\mathbf{x}_i, \mathbf{x}_j) = \max_{l=1,\dots,M} |\mathbf{x}_{il} - \mathbf{x}_{jl}|$
  - ∗ invariant bzgl.: Verschiebung
  - ∗ nicht invariant bzgl.: Drehung, Spiegelung, Skalierung



- *Canberra-Metrik:* $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{M} \frac{|\mathbf{x}_{il} - \mathbf{x}_{jl}|}{|\mathbf{x}_{il}| + |\mathbf{x}_{jl}|}$
  - invariant bzgl.: Skalierung
  - nicht invariant bzgl.: Verschiebung, Drehung, Spiegelung
- weitere Abstandsmaße:
  - Mahalanobis-Abstand
  - χ^2-Abstand
  - nichtmetrischer Koeffizient
  - Distanzmaß von Kendall

3.2 Abstandsmaße für binäre Daten

- *Kontingenztafel* binärer Merkmale der zwei Beobachtungsvektoren $\mathbf{x}_i$ und $\mathbf{x}_j$

Beobachtung i	Beobachtung j		
	0	1	
0	A_{00}	A_{01}	$A_{0\#}$
1	A_{10}	A_{11}	$A_{1\#}$
	$A_{\#0}$	$A_{\#1}$	$M = A_{\#\#}$

  - A_{00}: Anzahl übereinstimmender Nullen
  - A_{01}: Anzahl der Wertepaare mit $\mathbf{x}_i = 0$ und $\mathbf{x}_j = 1$
  - $A_{1\#}$: Anzahl der Einsen in $\mathbf{x}_i$
  - $A_{\#0}$: Anzahl der Nullen in $\mathbf{x}_j$

- *M-Koeffizient:*
  - Ähnlichkeitsmaß: $s = \frac{A_{00} + A_{11}}{M}$
  - relativer Anteil übereinstimmender Komponenten
  - Notation als Abstandsmaß: $d = 1 - s = \frac{A_{01} + A_{10}}{M}$
    (Hamming-Distanz)
  - vertauschungsinvariant: „0“ und „1“ werden gleichbehandelt
  - verallgemeinerbar durch zusätzlichen Gewichtsparameter

- *S-Koeffizient:*
  - Ähnlichkeitsmaß: $s = \dfrac{A_{11}}{M - A_{00}}$
  - relativer Anteil übereinstimmender Einsen
  - „1“ (Eigenschaft vorhanden) wird als wichtiger als „0“ (Eigenschaft nicht vorhanden) angenommen $\Rightarrow$ nicht vertauschungsinvariant
  - verallgemeinerbar durch zusätzlichen Gewichtsparameter

3.3 Abstandsmaße für qualitative Daten

- einfaches Ähnlichkeitsmaß:
  - $s(\mathbf{x}_i, \mathbf{x}_j) = \dfrac{\sum_{l=1}^{M} t(\mathbf{x}_{il}, \mathbf{x}_{jl})}{M}$ mit $t(\mathbf{x}_{il}, \mathbf{x}_{jl}) = \begin{cases} 1 & \text{für } \mathbf{x}_{il} = \mathbf{x}_{jl} \\ 0 & \text{für } \mathbf{x}_{il} \neq \mathbf{x}_{jl} \end{cases}$
  - relativer Anteil übereinstimmender Merkmale
- Ähnlichkeitsmaß von *Hyvärinen:*
  - $s(\mathbf{x}_i, \mathbf{x}_j) = \dfrac{\sum_{l=1}^{M} m_l \cdot t(\mathbf{x}_{il}, \mathbf{x}_{jl})}{\sum_{l=1}^{M} m_l}$ mit $t(\mathbf{x}_{il}, \mathbf{x}_{jl})$ wie oben
  - zusätzliche Berücksichtigung der Anzahl m_l beobachteter Ausprägungen je Merkmal

4 Clusterverfahren

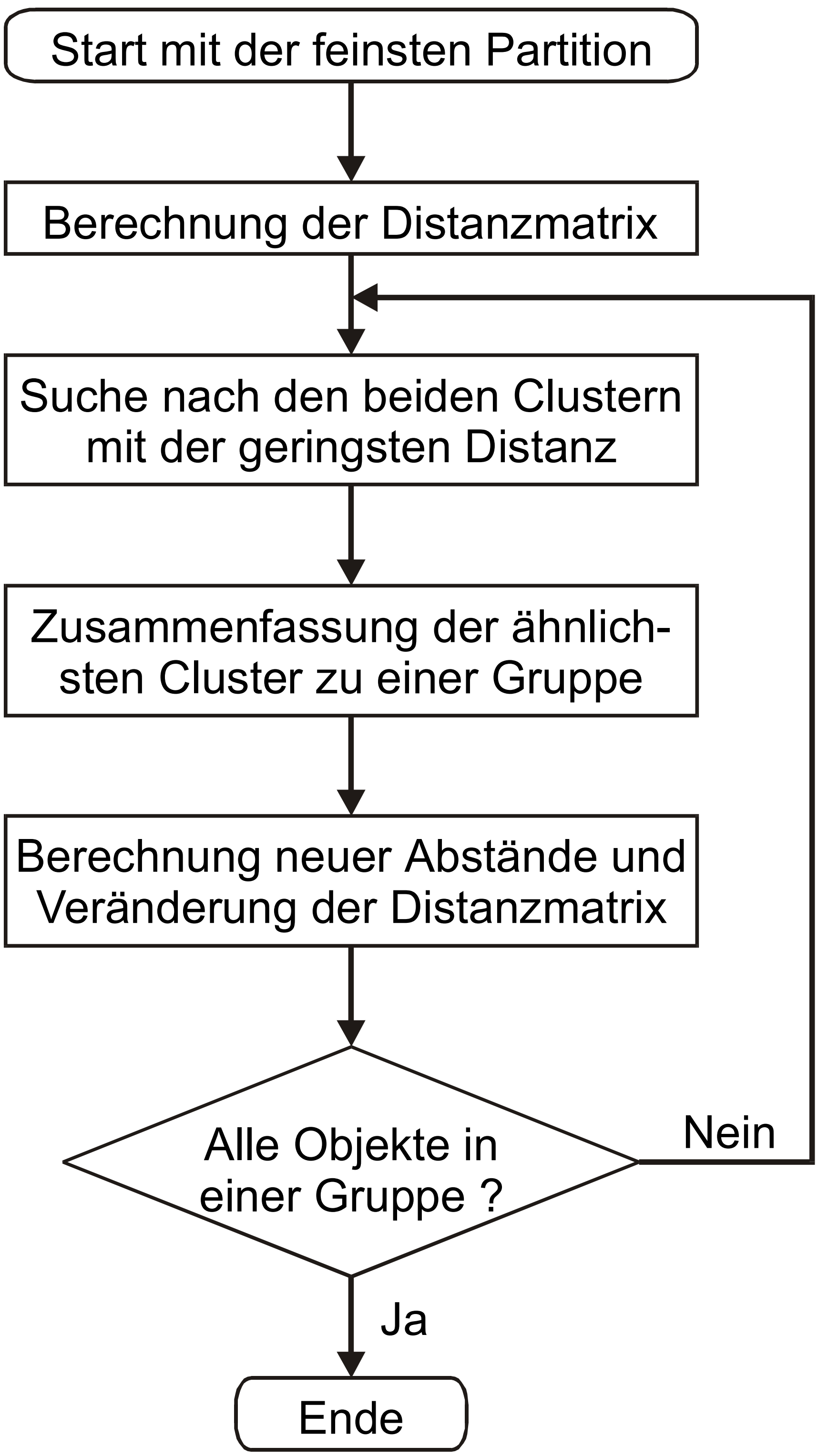
4.1 Einteilung

- *partitionierend:* Optimierung einer vorgegebenen Gruppierung durch Austausch von Objekten zwischen den Klassen
- *hierarchisch:* Fusion bzw. Unterteilung von Klassen → endgültige Partition kann ausgewählt werden
  - *agglomerativ:* sukzessive Fusion der beiden ähnlichsten Objekte bis nur noch eine Klasse existiert
  - *divisiv:* sukzessive Unterteilung der Klassen (ausgehend von einer Klasse) → sehr aufwendig

- *disjunkte* Klasseneinteilung: jedes Objekt gehört zu genau einer Klasse (z.B. Buch $\rightarrow$ Standort)
- *nicht-disjunkte* Klasseneinteilung: eindeutige Zuordnung eines Objektes zu einer Klasse ist nicht notwendig (z.B. Buch $\rightarrow$ Sachgebiet)
- *exhaustiv:* alle Objekte werden zugeordnet
- *nicht exhaustiv:* „nur die wichtigsten" Objekte werden zugeordnet
- *stabil:* zufällige Stichproben erzeugen sehr ähnliche Cluster

4.2 Hierarchische Clusterverfahren

- ♦ Algorithmus
  (für agglomerative Clusterverfahren):



- Beispiele:
  - *Single Linkage* Verfahren (Nearest Neighbor Method):
    - Distanz zwischen zwei Klassen = minimale Distanz zwischen Objektpaaren
    - Erkennung verzweigter, gekrümmter, linien- oder kreisförmiger Punktwolken
    - Verkettung von Klassen (Brücken)
    - instabil gegenüber Stichproben
  - *Complete Linkage* Verfahren (Furthest Neighbor Method):
    - Distanz zwischen zwei Klassen = maximale Distanz zwischen Objektpaaren
    - sehr homogene Klassen
    - instabil gegenüber Stichproben

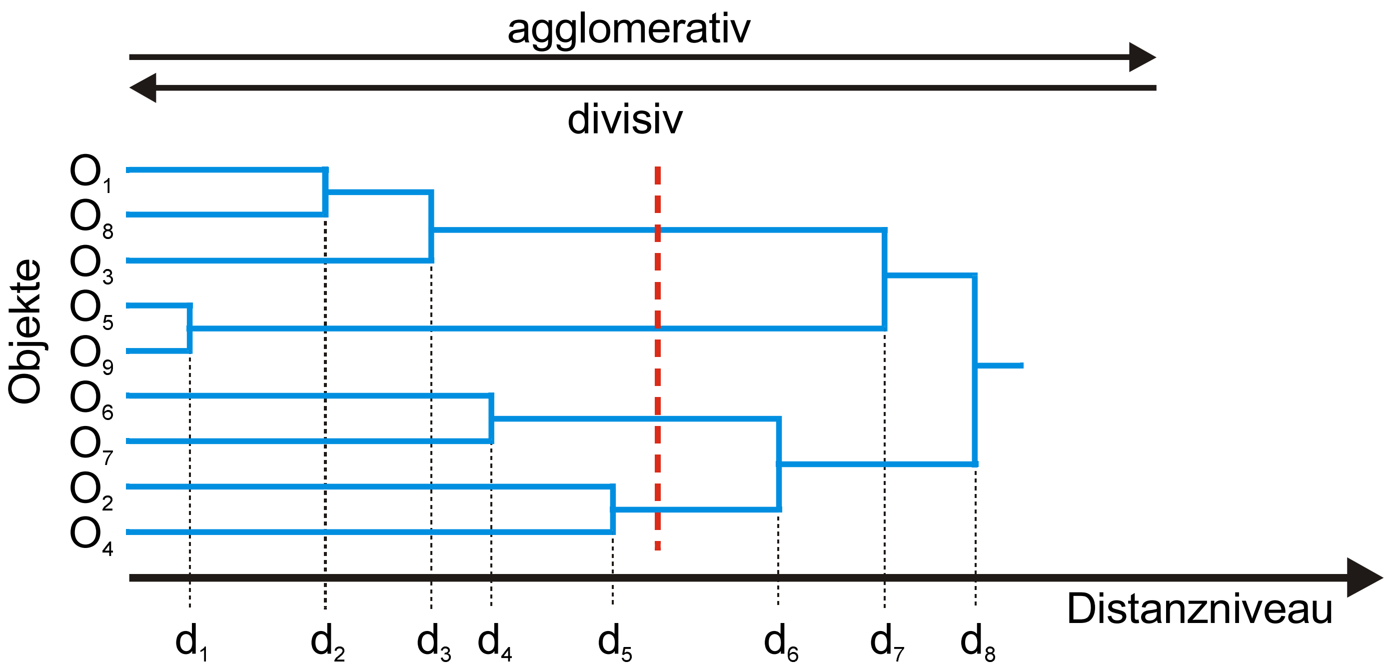
- *Average Linkage* Verfahren:
  - ∗ Kompromiß aus obigen Verfahren
  - ∗ Distanz zwischen zwei Klassen = arithmetisches Mittel aller paarweisen Distanzen
- *Zentroid*-Verfahren:
  - ∗ Distanz zwischen zwei Klassen = Distanz ihrer Klassenschwerpunkte (Zentroide)
- Minimalvarianz-Verfahren von *Ward:*
  - ∗ Distanz zwischen zwei Klassen = Varianzzuwachs der Klassenschwerpunkte bei paarweiser Verschmelzung
  - ∗ ähnlich große Klassen

- flexible Strategie nach *Lance & Williams:*
  - allgemeineres Verfahren: kann durch unterschiedliche Parameterwerte obige Verfahren annähern

♦ Eignung einiger hierarchischer Clusterverfahren für bestimmte Ziele:

	stabiles Verfahren	kompakte Cluster	langgestreckte Cluster	Ausreißer
Ward	+	+	-	-
Single Linkage	-	-	++	+
Complete Linkage	-	++	-	+
Average Linkage	++	+	-	+
Zentroid	+	-	-	-
Lance & Williams	+	+	+	+

♦ Dendrogramm:



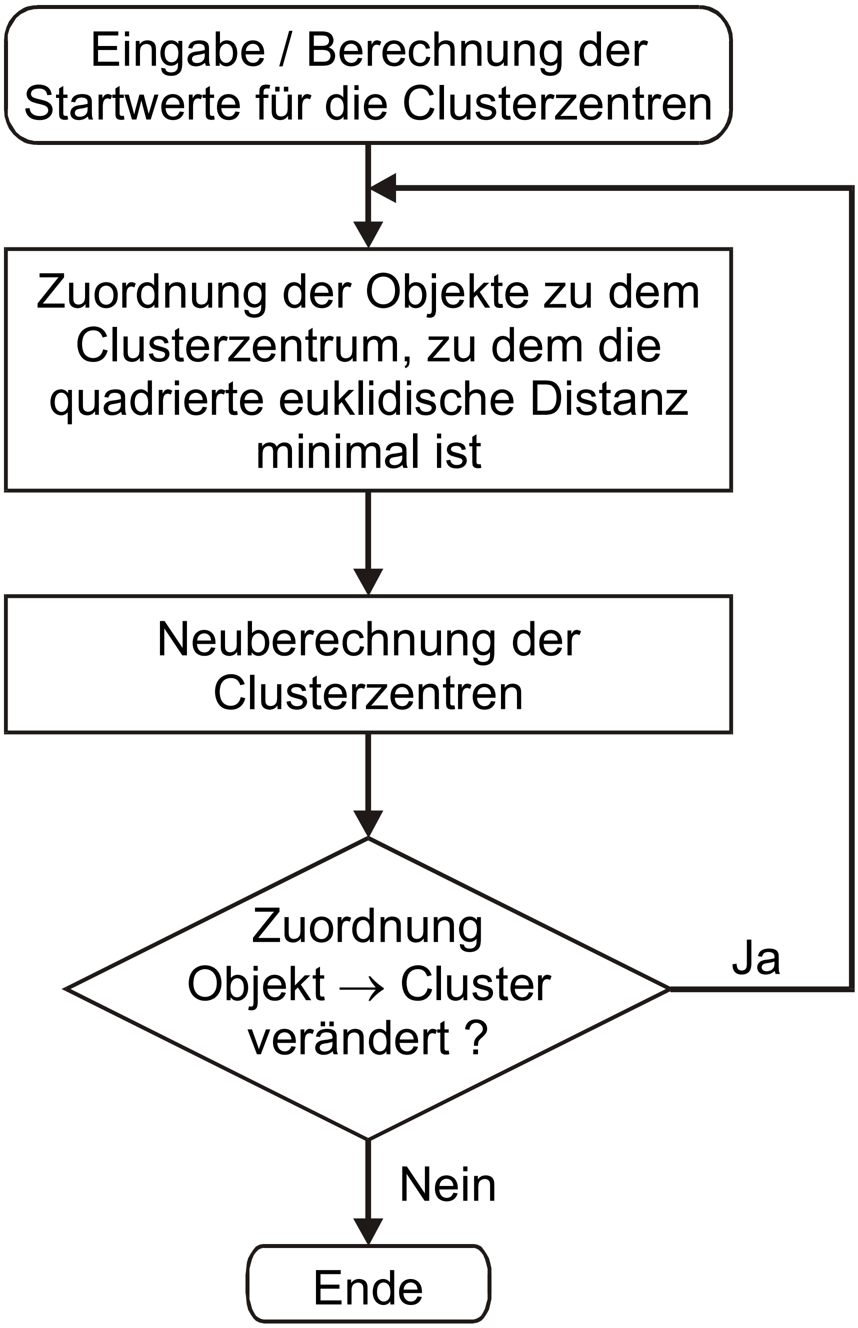
- Wahl einer Partition aufgrund
  - ∗ einer vorgegebenen Klassenzahl oder
  - ∗ eines großen Sprungs im Heterogenitätszuwachs

4.3 Partitionierende Clusterverfahren

- Ziel:
  - optimale Partition $\rightarrow$ Minimierung der klasseninternen Varianzen
    - Varianz innerhalb der Cluster $\triangleq$ Streuung der Daten, die nicht durch die Cluster erklärt wird *(Fehlerstreuung)*
- „optimale Lösung“:
  - Berechnung aller möglichen Partitionen und Auswahl der besten $\rightarrow$ nicht praktikabel (Kombinatorik) $\Rightarrow$ Näherungsverfahren

- Beispiele:
  - *K-means Verfahren* [Bac94]:
    - ∗ Konstruktion von Clusterzentren
    - ∗ Austausch von Objekten zwischen Clustern zur
      - – Minimierung der Varianzen innerhalb der Cluster
      - – Maximierung der Varianzen zwischen den Clustern
  - *Minimaldistanzverfahren:*
    - ∗ Variante des K-means Verfahrens
    - ∗ Austausch von Objekten zwischen Klassen aufgrund der euklidischen Distanz zum Klassenmittelwert
    - ∗ Möglichkeit der Bildung leerer Klassen

- Algorithmus [Bac94]:



5 Literatur

[GM98] U. Grimmer und H.-J. Mucha. Datensegmentierung mittels Clusteranalyse. In: G. Nakhaeizadeh (Hrsg.). Data Mining: Theoretische Aspekte und Anwendungen, S. 109-141, Physica-Verlag, Heidelberg, 1998. (Beiträge zur Wirtschaftsinformatik, Bd. 27) ISBN 3-7908-1053-3.

[Bac94] J. Bacher. Clusteranalyse: Anwendungsorientierte Einführung. R. Oldenbourg Verlag, München, 1994. ISBN 3-486-22832-3.

[Boc74] H. H. Bock. Automatische Klassifikation: Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten. Vandenhoeck & Ruprecht, Göttingen, 1974. (Studia Mathematica; Bd. 24) ISBN 3-525-40130-2.

[Sta85] H. Stahl. Clusteranalyse großer Objektmengen mit problemorientierten Distanzmaßen. Verlag Harri Deutsch, Frankfurt am Main, 1985. (Reihe Wirtschaftswissenschaften, Bd. 338) ISBN 3-87144-908-3.