

Hauptseminar KDD

SS 2002

Prof. Dr. Hans-Peter Kriegel

Eshref Januzaj

Karin Kailing

Peer Kröger

Matthias Schubert

Session: Clustering

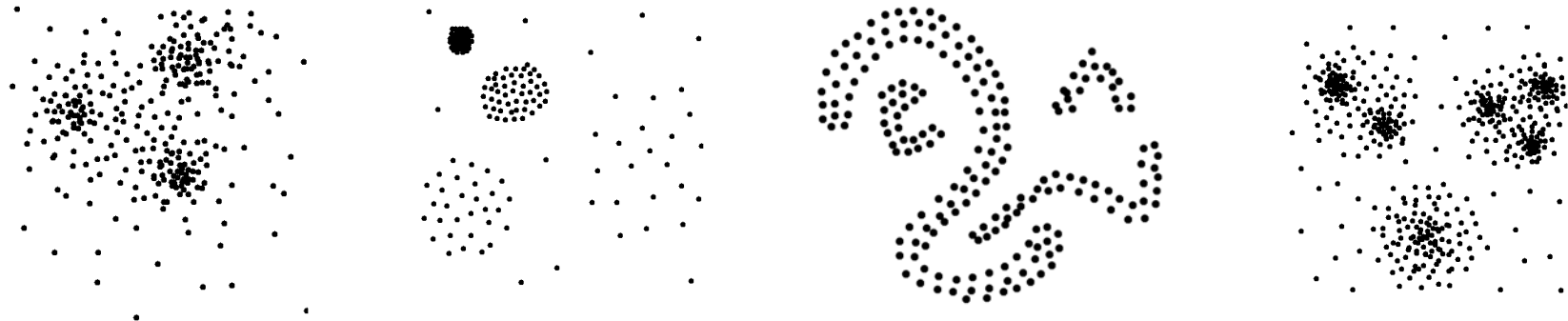
Inhalt

- Einleitung
- Klassifikation von Clustering-Verfahren
- Dichte-basierte Verfahren

Einleitung

Ziel des Clustering

- Identifikation einer endlichen Menge von Kategorien, Klassen oder Gruppen (*Cluster*) in den Daten
- Objekte im *gleichen* Cluster sollen möglichst ähnlich sein
- Objekte aus *verschiedenen* Clustern sollen möglichst unähnlich zueinander sein



Cluster unterschiedlicher Größe, Form und Dichte
hierarchische Cluster

Einleitung

Anwendungen (Überblick)

- Kundensegmentierung
Clustering der Kundentransaktionen
- Bestimmung von Benutzergruppen auf dem Web
Clustering der Web-Logs
- Strukturierung von großen Mengen von Textdokumenten
Hierarchisches Clustering der Textdokumente
- Erstellung von thematischen Karten aus Satellitenbildern
Clustering der aus den Rasterbildern gewonnenen Featurevektoren

Klassifikation von Clustering-Verfahren

Partitionierende Verfahren

- Parameter: Anzahl k der Cluster, Distanzfunktion
- sucht ein „flaches“ Clustering in k Cluster mit minimalen Kosten

Hierarchische Verfahren

- Parameter: Distanzfunktion für Punkte und für Cluster
- bestimmt Hierarchie von Clustern, mischt jeweils die ähnlichsten Cluster

Dichtebasierte Verfahren

- Parameter: minimale Dichte in einem Cluster, Distanzfunktion
- erweitert Punkte um ihre Nachbarn solange Dichte groß genug

Andere Clustering-Verfahren

- Fuzzy Clustering
- Graph-theoretische Verfahren
- neuronale Netze

Partitionierende Verfahren

Grundlagen

Ziel

eine Partitionierung in k Cluster mit minimalen Kosten

Lokal optimierendes Verfahren

- wähle k initiale Cluster-Repräsentanten
- optimiere diese Repräsentanten iterativ
- ordne jedes Objekt seinem ähnlichsten Repräsentanten zu

Typen von Cluster-Repräsentanten

- Mittelwert des Clusters (*Konstruktion zentraler Punkte $\Rightarrow$ Centroide*)
- Element des Clusters (*Auswahl repräsentativer Punkte $\Rightarrow$ Medoide*)
- Wahrscheinlichkeitsverteilung des Clusters (*Erwartungsmaximierung*)

Partitionierende Verfahren

Übersicht

- Clustering durch Varianz-Minimierung
- k-means [MacQueen 67]
- PAM [Kaufman & Rousseeuw 1990]
- CLARANS [Ng & Han 1994]
- EM-Algorithmus [Dempster, Laird & Rubin 1977]

Hierarchische Verfahren

Grundlagen

Ziel

Konstruktion einer Hierarchie von Clustern (*Dendrogramm*), so daß immer die Cluster mit minimaler Distanz verschmolzen werden

Dendrogramm

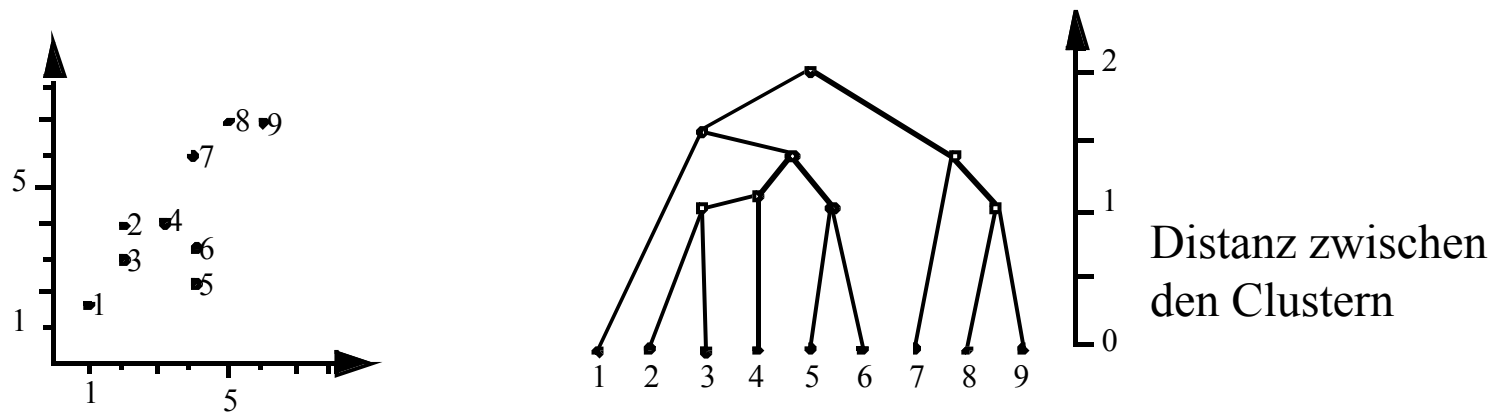
ein Baum, dessen Knoten jeweils ein Cluster repräsentieren, mit folgenden Eigenschaften:

- die Wurzel repräsentiert die ganze DB
- die Blätter repräsentieren einzelne Objekte
- ein innerer Knoten repräsentiert die Vereinigung aller Objekte, die im darunterliegenden Teilbaum repräsentiert werden

Hierarchische Verfahren

Grundlagen

Beispiel eines Dendrogramms



Typen von hierarchischen Verfahren

- Bottom-Up Konstruktion des Dendrogramms (*agglomerative*)
⇒ Algorithmus *Single-Link*
- Top-Down Konstruktion des Dendrogramms (*divisive*)

Dichtebasiertes Clustering

Grundlagen

Idee

- Cluster als Gebiete im d -dimensionalen Raum, in denen die Objekte dicht beieinander liegen
- getrennt durch Gebiete, in denen die Objekte weniger dicht liegen

Anforderungen an dichtebasierte Cluster

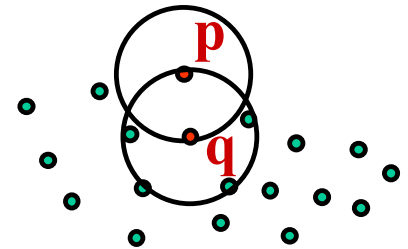
- für jedes Objekt eines Clusters überschreitet die lokale Punktdichte einen gegebenen Grenzwert
- die Menge von Objekten, die den Cluster ausmacht, ist räumlich zusammenhängend

Dichtebasiertes Clustering

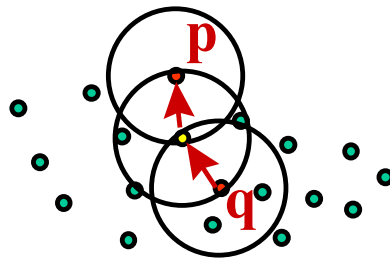
Grundbegriffe [Ester, Kriegel, Sander & Xu 1996]

- Ein Objekt $o \in O$ heißt *Kernobjekt*, wenn gilt:

$$|N_\varepsilon(o)| \geq \text{MinPts}, \text{ wobei } N_\varepsilon(o) = \{o' \in O \mid \text{dist}(o, o') \leq \varepsilon\}.$$



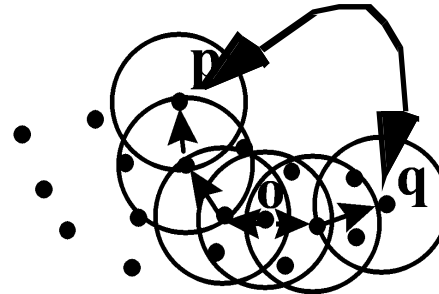
- Ein Objekt $p \in O$ ist *direkt dichte-erreichbar* von $q \in O$ bzgl. ε und MinPts , wenn gilt: $p \in N_\varepsilon(q)$ und q ist ein Kernobjekt in O .
- Ein Objekt p ist *dichte-erreichbar* von q , wenn es eine Kette von direkt erreichbaren Objekten zwischen q und p gibt.



Dichtebasiertes Clustering

Grundbegriffe

- Zwei Objekte p und q *dichte-verbunden*, wenn sie beide von einem dritten Objekt o aus dichte-erreichbar sind.



- Ein *Cluster* C bzgl. ε und $MinPts$ ist eine nicht-leere Teilmenge von O mit für die die folgenden Bedingungen erfüllt sind:

Maximalität: $\forall p, q \in O$: wenn $p \in C$ und q dichte-erreichbar von p ist, dann ist auch $q \in C$.

Verbundenheit: $\forall p, q \in C$: p ist dichte-verbunden mit q .

Dichtebasiertes Clustering

Grundbegriffe

- Grundlegende Eigenschaft

Sei C ein dichte-basierter Cluster und sei $p \in C$ ein Kernobjekt.

Dann gilt:

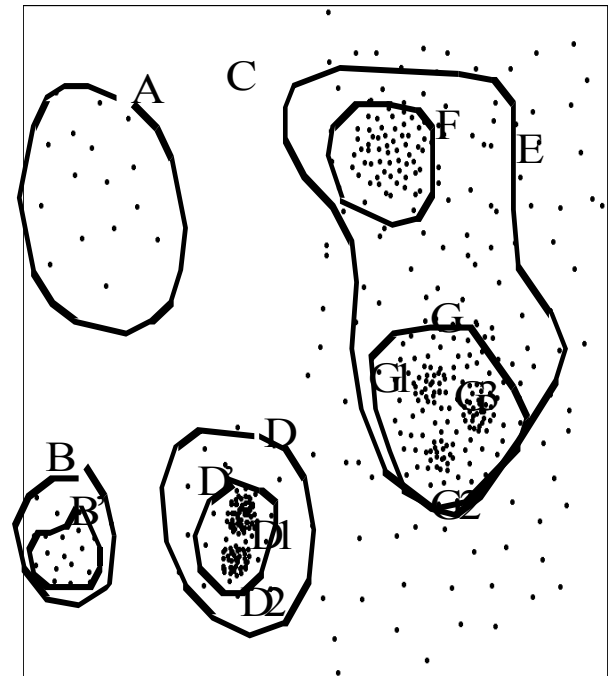
$$C = \{o \in O \mid o \text{ dichte-erreichbar von } p \text{ bzgl. } \varepsilon \text{ und } MinPts\}.$$

$\Rightarrow$ Algorithmus *DBSCAN* [Ester, Kriegel, Sander & Xu 1996]

Dichtebasiertes Clustering

Probleme

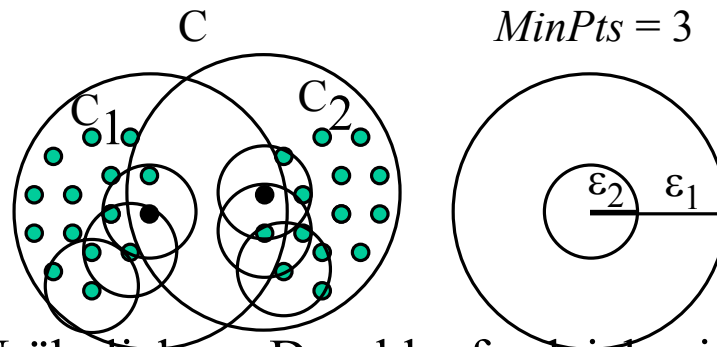
- hierarchische Cluster
- stark unterschiedliche Dichte in verschiedenen Bereichen des Raumes
- Cluster und Rauschen sind nicht gut getrennt



Dichte-basiertes hierarchisches Clustering

Grundlagen [Ankerst, Breunig, Kriegel & Sander 1999]

- für einen konstanten *MinPts*-Wert sind dichte-basierte Cluster bzgl. eines kleineren ε vollständig in Clustern bzgl. eines größeren ε enthalten



- in einem DBSCAN-ähnlichen Durchlauf gleichzeitig das Clustering für verschiedene Dichte-Parameter bestimmen

zuerst die dichteren Teil-Cluster, dann den dünneren Rest-Cluster

- kein Dendrogramm, sondern eine auch noch bei sehr großen Datenmengen übersichtliche Darstellung der Cluster-Hierarchie

Dichte-basiertes hierarchisches Clustering

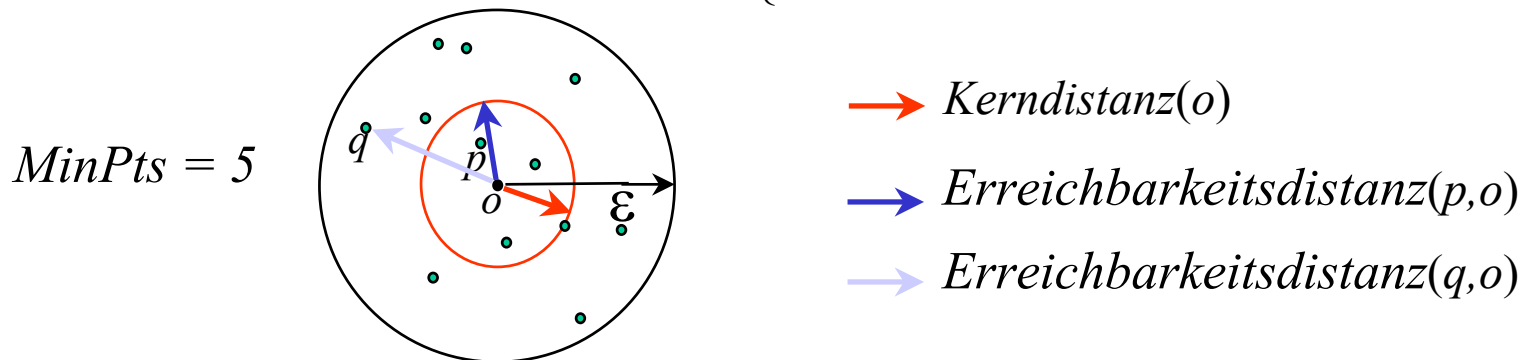
Grundbegriffe

Kerndistanz eines Objekts p bzgl. ε und $MinPts$

$$Kerndistanz_{\varepsilon, MinPts}(o) = \begin{cases} UNDEFINIERT, & \text{wenn } |N_{\varepsilon}(o)| < MinPts \\ MinPtsDistanz(o), & \text{sonst} \end{cases}$$

Erreichbarkeitsdistanz eines Objekts p relativ zu einem Objekt o

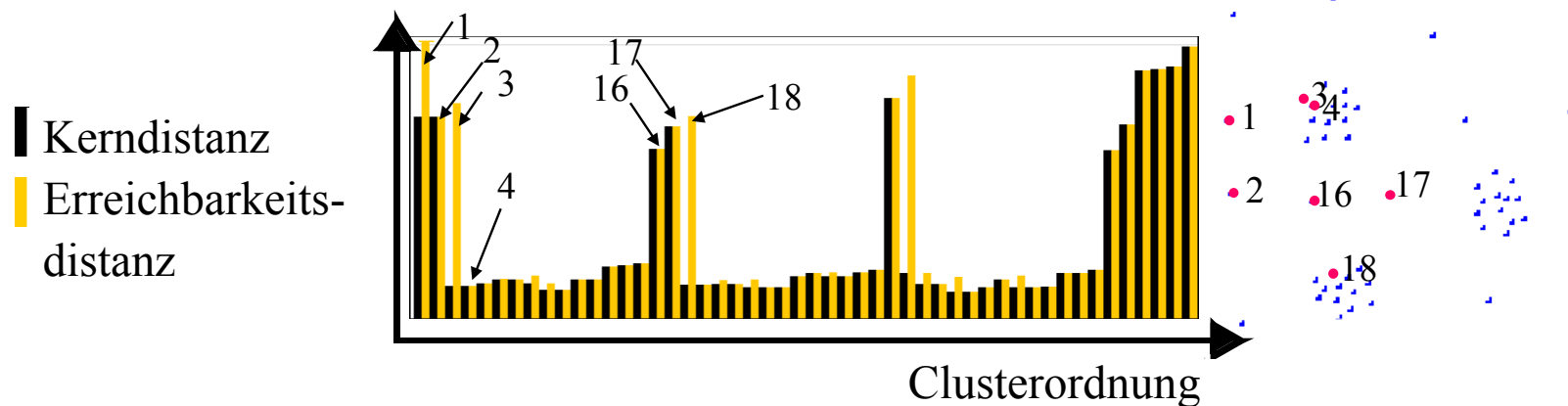
$$Erreichbarkeitsdistanz_{\varepsilon, MinPts}(p, o) = \begin{cases} UNDEFINIERT, & \text{wenn } |N_{\varepsilon}(o)| < MinPts \\ \max\{Kerndistanz(o), dist(o, p)\}, & \text{sonst} \end{cases}$$



Dichte-basiertes hierarchisches Clustering

Clusterordnung

- OPTICS liefert nicht direkt ein (hierarchisches) Clustering, sondern eine „Clusterordnung“ bzgl. ϵ und $MinPts$
- *Clusterordnung* bzgl. ϵ und $MinPts$
 - beginnt mit einem beliebigen Objekt
 - als nächstes wird das Objekt besucht, das zur Menge der bisher besuchten Objekte die minimale Erreichbarkeitsdistanz besitzt



Dichte-basiertes hierarchisches Clustering

Erreichbarkeits-Diagramm

- Zeigt die Erreichbarkeitsdistanzen (bzgl. ε und $MinPts$) der Objekte als senkrechte, nebeneinanderliegende Balken
- in der durch die Clusterordnung der Objekte gegebenen Reihenfolge

