Ludwig Maximilians Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung
Knowledge Discovery in Databases II
im Wintersemester 2013/2014

Kapitel 7: Data Mining in großen Datensammlungen

http://www.dbs.ifi.lmu.de/Lehre/KDD

Kapitelübersicht

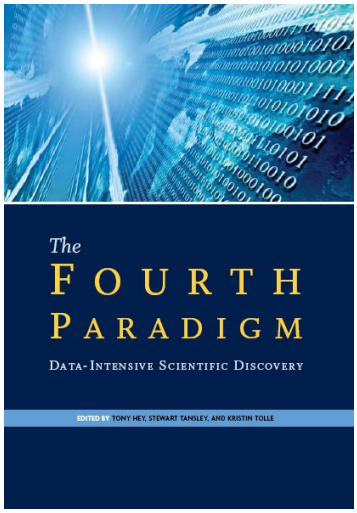
1. Big Data, Data Science und Dimensionen von Daten
2. Umgang mit großen Datenmengen
3. Sampling und Micro-Clustering
4. Online Data Mining und Inkrementelles Lernen

The Fourth Paradigm

1. Vor Eintausend Jahren: Experimentelle Wissenschaft
   - Beschreibung natürlicher Phänomene
2. In den letzten Jahrhunderten: Theoretische Wissenschaft
   - Newton'sche Gesetze, Maxwell-Gleichungen, …
3. Die letzten Jahrzehnte: Computergestützte Wissenschaft
   - Simulation komplexer Phänomene
4. Heute: Daten-Intensive Wissenschaft
   - Vereinigt Theorie, Experimente und Simulation

"Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

-The Fourth Paradigm - Microsoft



Big Data

Neue Herausforderung für Wissenschaftler aller Bereiche: Der Umgang mit riesigen Datenmengen

Auszug aus dem McKinsey Report *"Big data: The next frontier for innovation, competition, and productivity"*, Juni 2011:

Capturing the value of Big data:

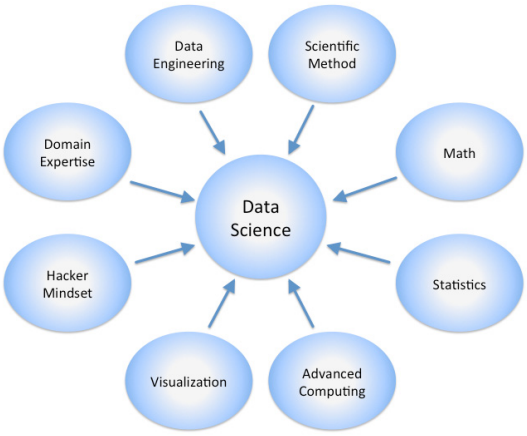
- 300Mrd USD potentieller Wert für das amerikanische Gesundheitswesen, jährlich.
- 250Mrd Euro potentieller Wert für den öffentlichen Dienst in Europa, jährlich.
- 600Mrd USD potentieller Wert durch die Verwendung von location based services.

"The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings."

-www.mckinsey.com/mgi/publications/big_data/

Data Science

- Beschreibt die Wissenschaft der Verwaltung und Analyse von Daten zur Generierung von Wissen
- Kerngedanke sehr ähnlich zu Knowledge Discovery in Databases



Unterschiede:

1. Data Science ist umfassender in den Aufgabengebieten definiert. (Ergebnispräsentation, Handlungsempfehlungen,..)
2. Versucht alle Forschungsrichtungen, die sich mit Datenanalyse und Wissensrepräsentation beschäftigen zu integrieren.
3. Betrachung neuer Berechnungsmodelle und Hardwarearchitekturen.+

Fazit: Viele Forschungsbereiche beschäftigen sich seit Jahrzenten mit ähnlichen Problemen und Techniken.

Data Science ist also als Überbegriff zu sehen.

Dimensionen von Datenbeständen

The four V's

- Volume: Anzahl an Datenobjekten, Anzahl der Merkmale

  => großes Datenvolumen, Skalierungsproblematiken

- Variety: Heterogenität der Daten, Struktur, Dimensionalität

  => Integration unterschiedlicher Quellen, Strukturierte Daten, Netzwerke

- Velocity: Veränderung der Daten über die Zeit

  => Alterung des Wissens, zeitliche Veränderung, Periodische Muster

- Veracity: Unsicherheit der beobachteten Daten

  => Umgang mit Messfehlern, falschen Daten, Unvollständigkeit

Behandelte Themen

- Variety: strukturierte Objekte (Kaptiel 4-6)
- Volumen:
  - Featureselektion und Featurereduktion, HD-Clustering (Kapitel 2-3)
  - Sampling, Micro-Clustering (Kapitel 7)
  - Paralleles und Verteiltes Data Mining (Kapitel 8)
- Velocity:
  - Stream Data Mining (Kaptiel 9)

- Veracity: Wird nicht explicit in der Vorlesung behandelt

  Aber: Messfehler und Umgang mit unsicheren Beobachtungen sind bereits inhärenter Bestandteil vieler Data Mining Ansätze.

  (Viele Verfahren modellieren Fehlerverteilungen)

Datenkompression

- Idee:
  - DM-Algorithmus auf der gesamten Datenmenge zu teuer
  - Komprimiere Daten, so dass sie in den Hauptspeicher passen
  - Wende DM-Algorithmus auf komprimierte Daten an

- Techniken:
  - Sampling

    Datenbank wird auf eine Stichprobe reduziert

  - Micro-Clustering

    bilde Micro-Cluster, die Teilmengen der Daten möglichst genau repräsentieren; Datenbank wird auf Micro-Cluster reduziert

Sampling

Indexbasiertes Sampling [Ester, Kriegel & Xu 1995]

- Zufälliges Sampling liefert u.U. schlechte Qualität
- Verwendung von räumlichen Indexstrukturen oder verwandten Techniken zur Auswahl des Samplings
  - Indexstrukturen liefern ein grobes Vor-Clustering
    räumlich benachbarte Objekte werden möglichst
    auf der gleichen Seite abgespeichert
  - Indexstrukturen sind effizient da nur einfache Heuristiken zum Clustering
  - schnelle Zugriffsmethoden für verschiedene Ähnlichkeitsanfragen
    z.B. Bereichsanfragen und *k*-Nächste-Nachbarn-Anfragen

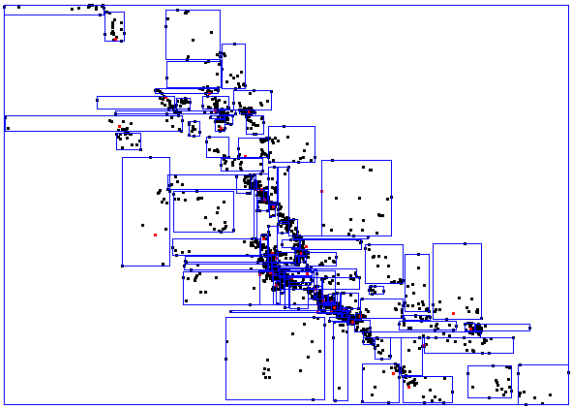
DATABASE SYSTEMS GROUP

Sampling

LMU

Methode

- Aufbau eines R-Baums
- Auswahl von Repräsentanten von den Datenseiten des R-Baums
- Anwendung des Clustering-Verfahrens auf die Repräsentantenmenge
- Übertragung des Clustering auf die gesamte Datenbank



Datenseitenstruktur
eines R*-Baums

Sampling

Auswahl von Repräsentanten

Wieviele Objekte sollen von jeder Datenseite ausgewählt werden?

- hängt vom verwendeten Clusteringverfahren ab
- hängt von der Verteilung der Daten ab
- z.B. für CLARANS: ein Objekt pro Datenseite

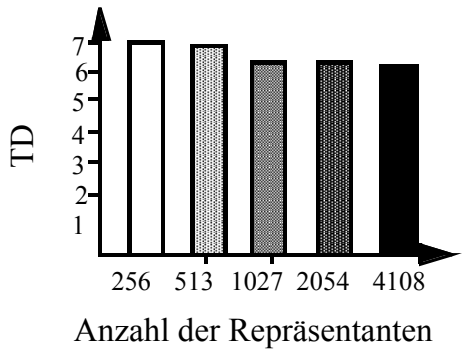
guter Kompromiss zwischen der Qualität des Clusterings und der Laufzeit

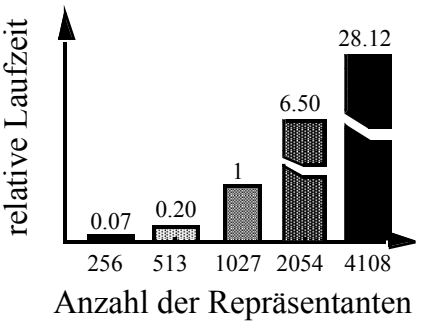
Welche Objekte sollen ausgewählt werden?

- hängt ebenfalls vom Clusteringverfahren und von der Verteilung der Daten ab
- einfache Heuristik: wähle das „zentralste“ Objekt auf der Datenseite

Sampling

Experimentelle Untersuchung für CLARANS





- Laufzeit von CLARANS ist etwa $O(n^2)$
- Qualität des Clusterings steigt bei mehr als 1024 Repräsentanten kaum noch

=> 1024 Repräsentanten guter Kompromiss zwischen Qualität und Effizienz

Micro-Clustering

BIRCH [Zhang, Ramakrishnan & Linvy 1996]

Methode

- Bildung kompakter Beschreibungen von Teil-Clustern (Clustering Features)
- hierarchische Organisation der Clustering Features in einem höhenbalancierten Baum (CF-Baum)
- Anwendung eines Clusteringverfahren wie z.B. CLARANS auf die Clustering Features in den Blättern des Baums

CF-Baum

- komprimierte, hierarchische Repräsentation der Daten
- berücksichtigt die Clusterstruktur

Micro-Clustering

Grundbegriffe

- *Clustering Feature* einer Menge *C* von Punkten X_i: $CF = (N, \overrightarrow{LS}, SS)$
- $N = |C|$ „Anzahl der Punkte in *C*“
- $\overrightarrow{LS} = \sum_{i=1}^{N} X_i$ „lineare Summe der *N* Datenpunkte“
- $SS = \sum_{i=1}^{N} X_i^2$ „Quadratsumme der *N* Datenpunkte“

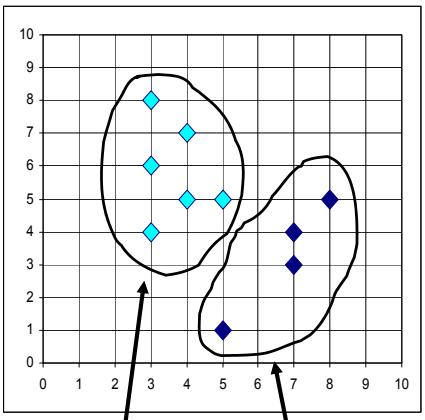
aus den CF‘s können berechnet werden:

- Centroid (Repräsentant)
- Kompaktheitsmaße und Distanzmaße für Cluster

Micro-Clustering

Beispiel:

(3,4)
(4,5)
(5,5)
(3,6)
(4,7)
(3,8)



(5,1)
(7,3)
(7,4)
(8,5)

$CF_1 = (6, (22,35), 299)$

$CF_2 = (4, (27,13), 238)$

Micro-Clustering

Grundbegriffe

Additivitätstheorem

für CF-Vektoren für zwei disjunkte Cluster C_1 und C_2 gilt:

$$CF(C_1 \cup C_2) = CF(C_1) + CF(C_2) = (N_1 + N_2, LS_1 + LS_2, QS_1 + QS_2)$$

d.h. CF's können inkrementell berechnet werden

Definition

Ein *CF-Baum* ist ein höhenbalancierter Baum zur Abspeicherung von CF's.

CF-Baum

Eigenschaften eines CF-Baums

- Jeder innere Knoten enthält höchstens *B* Einträge der Form $[CF_i, child_i]$ und CF_i ist der CF-Vektor des Subclusters des *i*-ten Sohnknotens.
- Ein Blattknoten enthält höchstens *L* Einträge der Form $[CF_i]$.
- Jeder Blattknoten besitzt zwei Zeiger *prev* und *next.*
- Für jeden Eintrag eines Blattknotens ist der Durchmesser kleiner als *T.*

CF-Baum

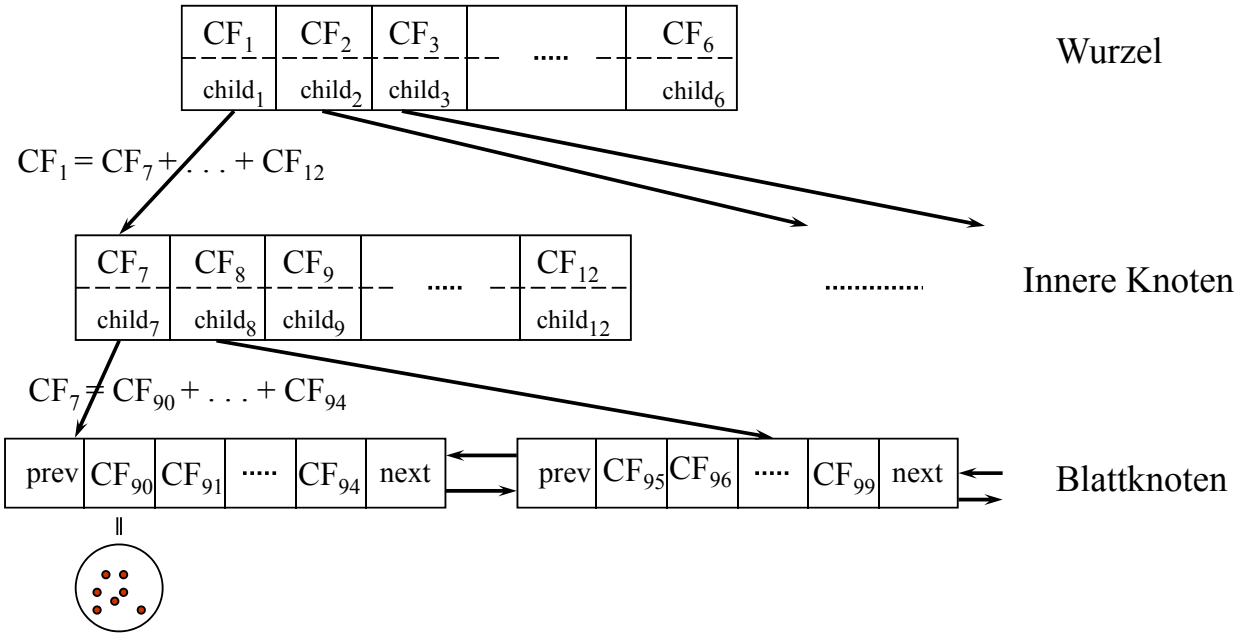
Aufbau eines CF-Baums (analog B^+-Baum)

- Transformation eines Datensatzes *p* in einen CF-Vektor $CF_p=(1, p, p^2)$
- Einfügen von CF_p in den Teilbaum des CF-Vektors mit kleinster Distanz
- bei Verletzung des Schwellwerts *T* wird ein neuer Eintrag CF_p eingefügt
- sonst absorbiert der nächste Eintrag im Blatt CF_p
- bei Verletzung des Schwellenwerts *B* oder *L* wird Knoten gesplittet:
  - die entferntesten CF‘s bilden die beiden neuen Knoten
  - die restlichen CF‘s werden dem neuen Knoten mit geringster Distanz zugeordnet

Micro-Clustering

Beispiel

B = 7, L = 5



Micro-Clustering

Gesamter Algorithmus

Phase 1

- ein Scan über die gesamte Datenbank
- Aufbau eines CF-Baums B_1 bzgl. T_1 durch sukzessives Einfügen der Datensätze

Phase 2

- falls der CF-Baum B_1 noch zu groß ist, wähle ein $T_2 > T_1$
- Aufbau eines CF-Baums B_2 bzgl. T_2 durch Einfügen der CF‘s der Blätter von B_1

Phase 3

- Anwendung eines Clusteringalgorithmus auf die Blatteinträge des CF-Baums
- Clusteringalgorithmus muss evtl. an Clustering Features angepasst werden

Micro-Clustering

Diskussion

+ Komprimierungsfaktor frei wählbar

+ Effizienz:

  Aufbau eines sekundärspeicherresidenten CF-Baums: O(n log n)

  Aufbau eines hauptspeicherresidenten CF-Baums: O(n)

zusätzlich: Aufwand des Clusteringalgorithmus
(wenn CF-Baum im Hauptspeicher, ist dieser Aufwand vernachlässigbar)

- nur für numerische Daten euklidischer Vektorraum

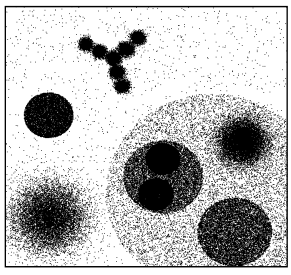
- abhängig von der Reihenfolge der Daten

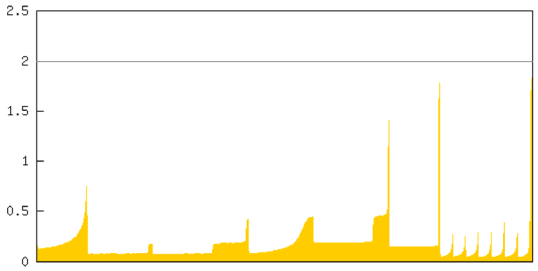
Micro-Clustering

Data Bubbles [Breunig, Kriegel, Kröger, Sander 2001]

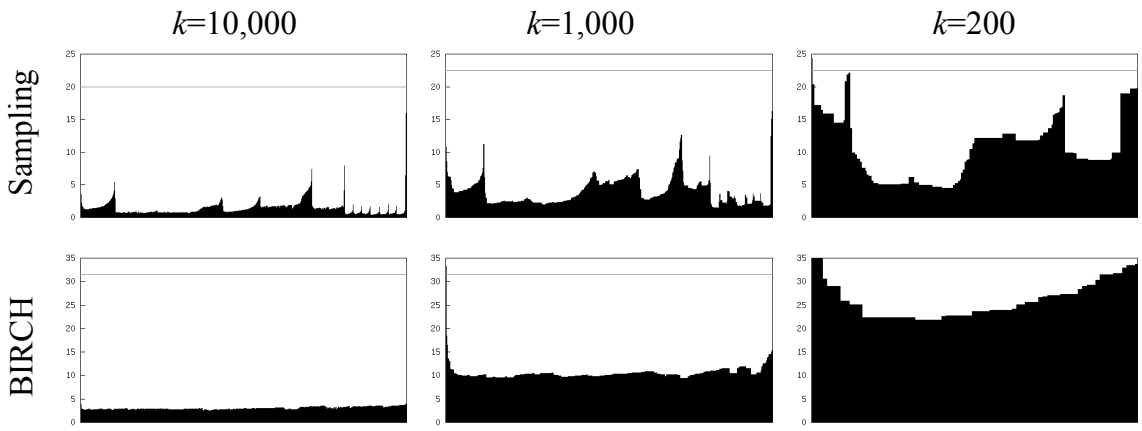
Original DB und OPTICS-Plot

1 Mio. Datenpunkte





OPTICS-Plots auf Komprimierten Daten



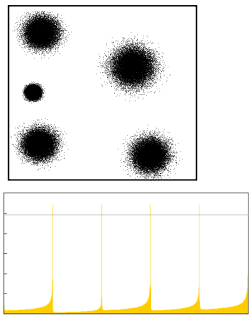
Micro-Clustering

Drei Hauptprobleme bei BIRCH und Sampling für hierarchisches Clustering:

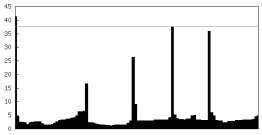
1. Verlorengegangene Objekte (Lost Objects)
   Viele Objekte der DB fehlen im Plot/Dendrogram
2. Größenverzerrung (Size Distortions)
   Cluster sind gequetscht und gestreckt
3. Strukturelle Verzerrung (Structural Distortions)
   Hierarchische Cluster Struktur ist zerstört

Lösungen

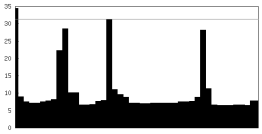
- Post-Processing für Problem 1 und 2 (Lost Objects, Size Distortions)
  nn-Klassifikation, ersetze repräsentative Objekte durch die Menge der repräsentierten Objekte
- Data Bubbles für Problem 3 (Structural Distortions)



OPTICS original



Sampling 100 Obj.



CF 100 Obj.

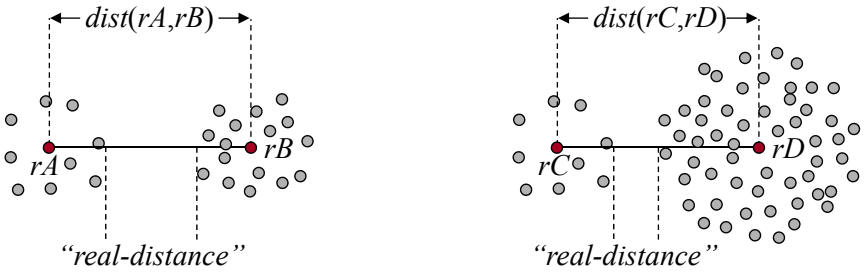
DATABASE SYSTEMS GROUP

Micro-Clustering

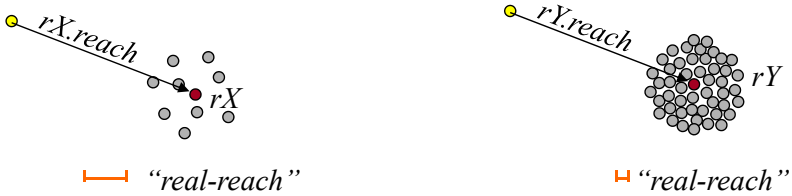
LMU

Gründe für strukturelle Verzerrungen:

- Distanz zwischen den Originalobjekten wird schlecht durch die Distanz zwischen Repräsentanten approximiert



- Erreichbarkeitsdistanz die den Repräsentanten zugeordnet werden, approximieren die Erreichbarkeitsdistanz der repräsentierten Objekte sehr schlecht



Micro-Clustering

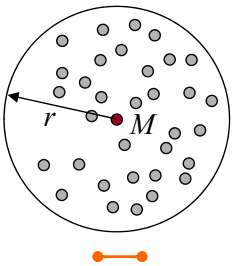
Data Bubble: reichere Abstraktion

Definition: **Data Bubble**

- Data Bubble $B=(n, M, r)$ für eine Menge von n Objekten $X=\{X_i\}$

$$M = \left(\sum_{i=1}^{n} X_i \right) / n$$ ist der *Mittelpunkt* von X und

$$r = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2}{n \cdot (n-1)}}$$ ist der *Radius* von X.

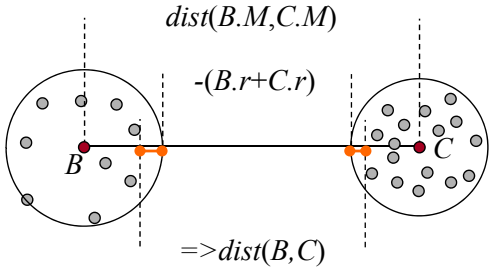


erwartete k-nn Distanz

- Erwartete k-nn Distanz der Objekte X_i im Data Bubble (bei Gleichverteilung) $$nnDist(k, B) = r \cdot \left(\frac{k}{n} \right)^d$$
- Data Bubbles können entweder aus einem Sample oder aus CFs berechnet werden

Micro-Clustering

- Definition: Distanz zwischen Data Bubbles

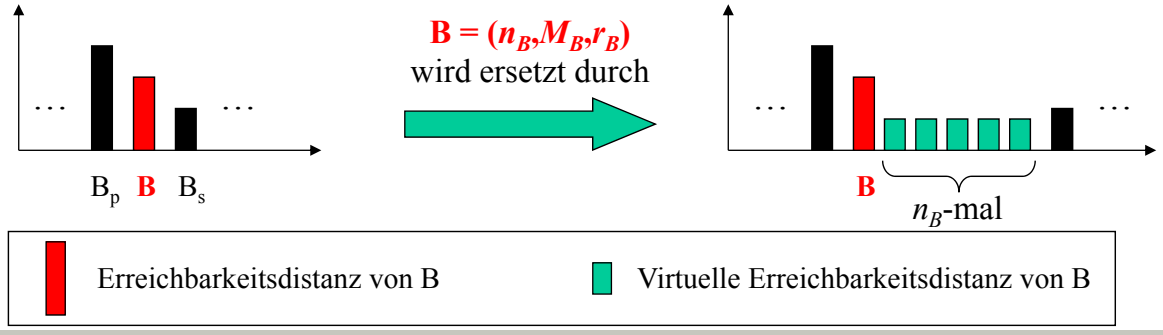


- Definition: Kern- und Erreichbarkeitsdistanz für Data Bubbles
  - analog zur Kern- und Erreichbarkeitsdistanz von Punkten
- Definition: virtuelle Erreichbarkeitsdistanz eines Data Bubble
  - Erwartete k-nn-Distanz innerhalb des Data Bubble
  - bessere Approximation der Erreichbarkeitsdistanz der repräsentierten Objekte

Micro-Clustering

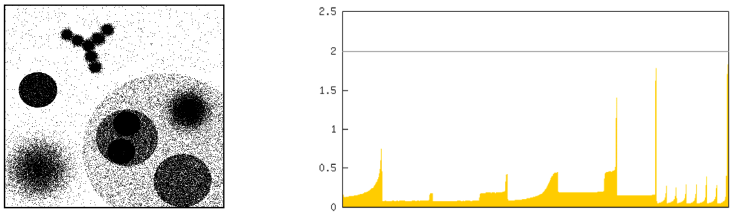
- Clustering mit Data Bubbles:
  - Generiere *m* Data Bubbles aus *m* Sampleobjekten oder CF-Features
  - Cluster die Menge der Data Bubbles mit OPTICS
  - Generiere das Erreichbarkeitsdiagramm:

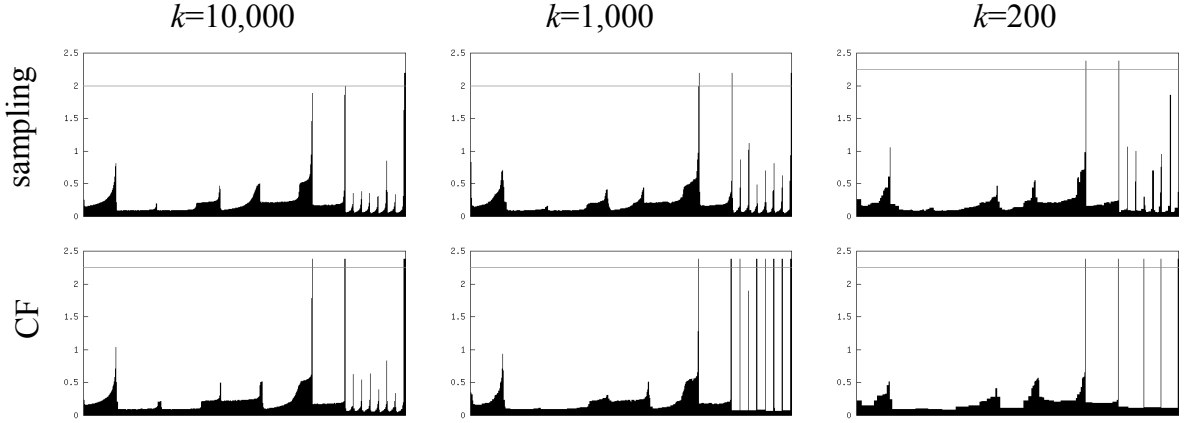
    Für jedes Data Bubble *B*:
    - „Plotte“ die Erreichbarkeitsdistanz *B.reach* (vom OPTICS-Lauf auf den Data Bubbles erzeugt)
    - „Plotte“ alle Punkte, die von *B* repräsentiert werden mit der virtuellen Erreichbarkeitsdistanz von *B*



Micro-Clustering

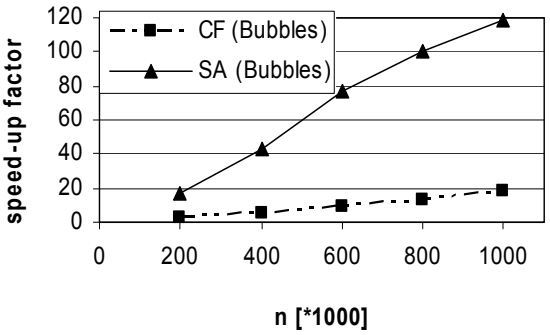
Ergebnisse (Kompression auf *k* Objekte)





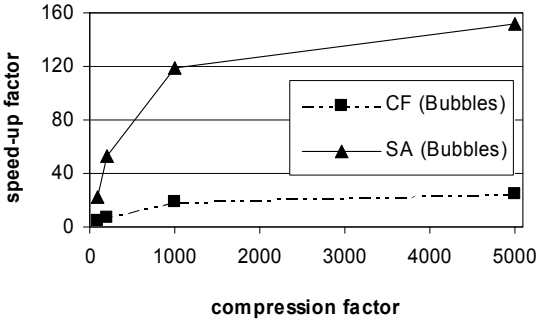
Micro-Clustering

Speed-Up-Faktoren



bzgl. Datenbank Größe

Speed-Up-Faktoren



bzgl. Kompressions Faktor

für Datenbank mit (1 Mio. Objekte)

Diskussion Micro-Clustering und Sampling

- Häufig lassen sich die gleichen Muster auf einer wesentlich geringeren Stichprobengröße extrahieren.
- Entscheidend ist nicht die Anzahl der Stichproben sondern die gut sie die Datenverteilung repräsentieren.
- Sampling und Micro-Clustering versuchen die Datenverteilung zu approximieren und dann gezielt eine Teilmenge zu repräsentieren
- In der Klassifikation gibt es ähnliche Ansätze

  Instanzselektion: Selektion von Trainingsobjekten, die nahe an der Klassengrenze liegen. (vgl. Idee mit Support Vektoren)

  (aus Zeitgründe kann nicht weiter darauf eingegangen werden)

Literatur

- M. Ester, H.-P. Kriegel, X. Xu:
  A Database Interface for Clustering in Large Spatial Databases
  In Proceedings of the 1st ACM International Conference on Knowledge Discovery and Data Mining (KDD), Montreal, QC, Canada: 94–99, 1995.

- Tian Zhang, Raghu Ramakrishnan, and Miron Livny:
  BIRCH: an efficient data clustering method for very large databases
  SIGMOD Rec. 25, 2 (June 1996), 103-114. DOI=10.1145/235968.233324
  http://doi.acm.org/10.1145/235968.233324

- Markus M. Breunig, Hans-Peter Kriegel, Peer Kröger, and Jörg Sander:
  Data bubbles: quality preserving performance boosting for hierarchical clustering
  SIGMOD Rec. 30, 2 (May 2001), 79-90. DOI=10.1145/376284.375672
  http://doi.acm.org/10.1145/376284.375672