Data Mining:

5. Spezialisierungen und Erweiterungen

A) Entdeckung von Anomalien
B) Kombination von DM-Verfahren

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a test point **x**, compute the anomaly score of **x** with respect to the data base
  - Find all the data points **x** with anomaly scores > some threshold t
  - Find all the data points **x** having the top-n largest anomaly scores
- Applications:
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection, unusual symptoms in medical diagnosis, ecosystem disturbances

Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!

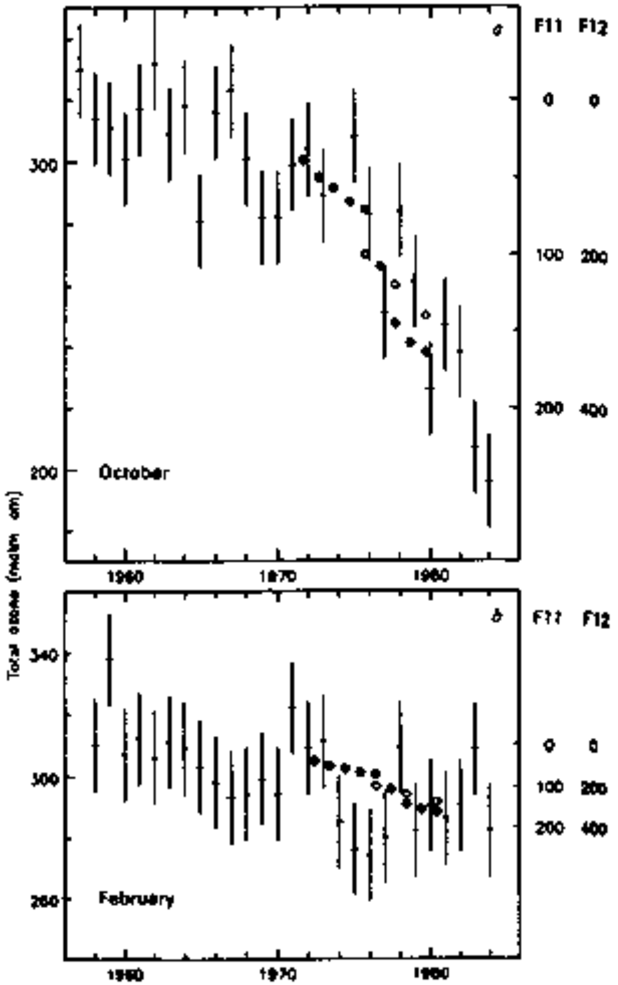
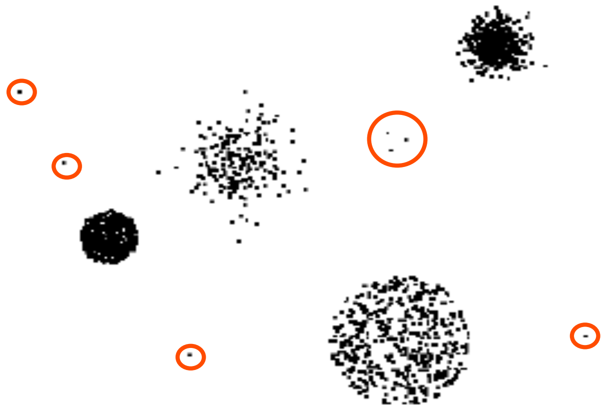


Fig. 2 Monthly means of total O_3 at Halley Bay, and Southern Hemisphere measurements of F-11 (●, p.p.t.v. (parts per thousand by volume) $CFCl_3$) and F-12 (○, p.p.t.v. CF_2Cl_2). *a*, October, 1957-84. *b*, February, 1958-84. Note that F-11 and F-12 amounts increase down the figure.

Anomaly Detection Schemes

- General Steps
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

- Types of anomaly detection schemes
  - Visual/Graphical: by inspecting plots (subjective)
  - Statistical
  - Distance-/Density-based
  - Clustering-based
  - or Model-based
    - = supervised, using classification techniques (but training anomalies are rare)

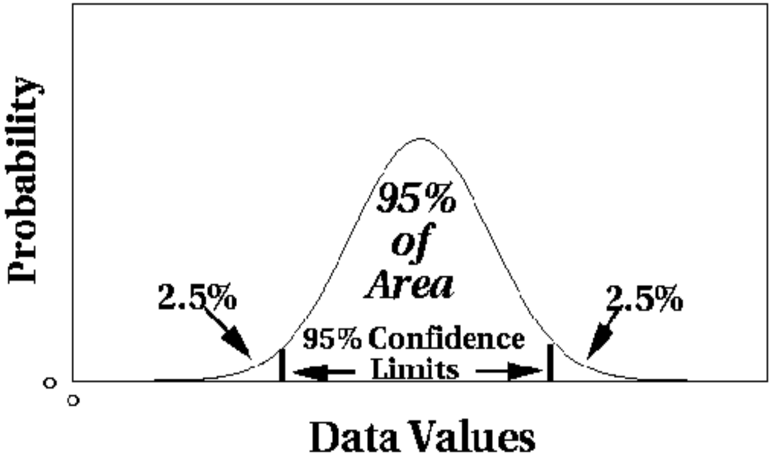


Statistical Approaches

- Assume a parametric model describing the distribution of the data, e.g., normal distribution

  (if known, often for a single attribute only)

- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution, e.g., mean/variance
  - Number of expected outliers, or: Confidence limit



Distance-Based Approach

- Data object is represented as a vector of features
- *Idea:* An object is an anomaly
  if it is distant from “most” points.
- *Approach:*
  - Compute the distance between every pair of data points; *needs quadratic time* O(N^2)
  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *d*
    - The top n data points whose distance to the kth nearest neighbor (= “outlier score”) is greatest; --- see figures -- *very parameter-sensitive*
    - *more robust:* The top n data points whose average distance to the k nearest neighbors is greatest

Distance-Based Approach

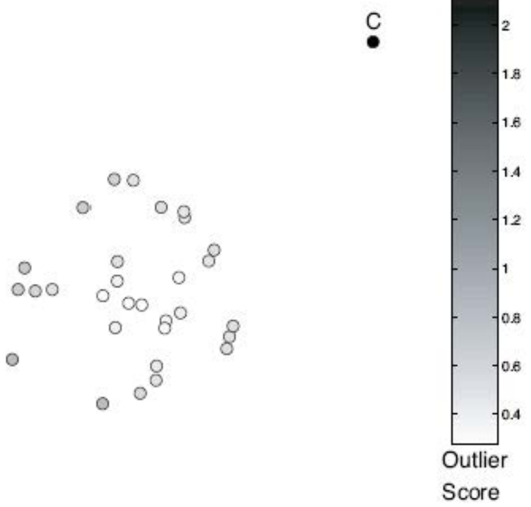


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

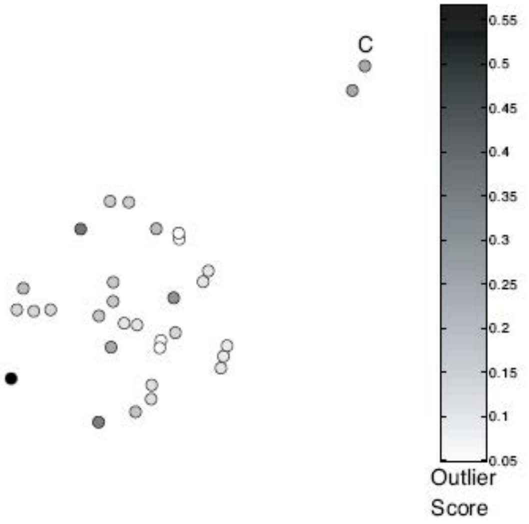


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

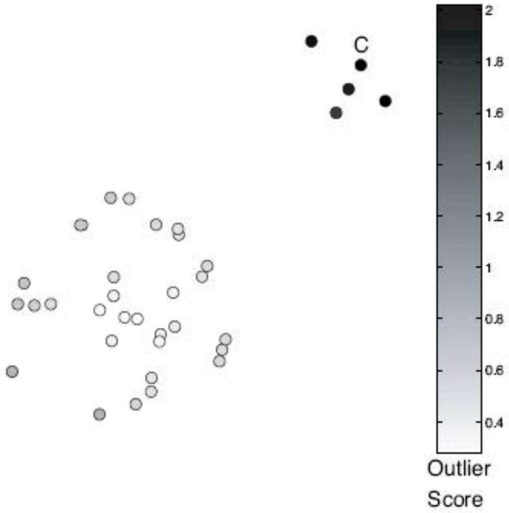


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

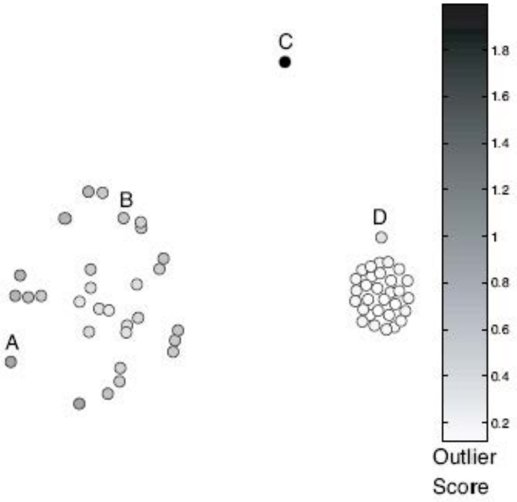


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

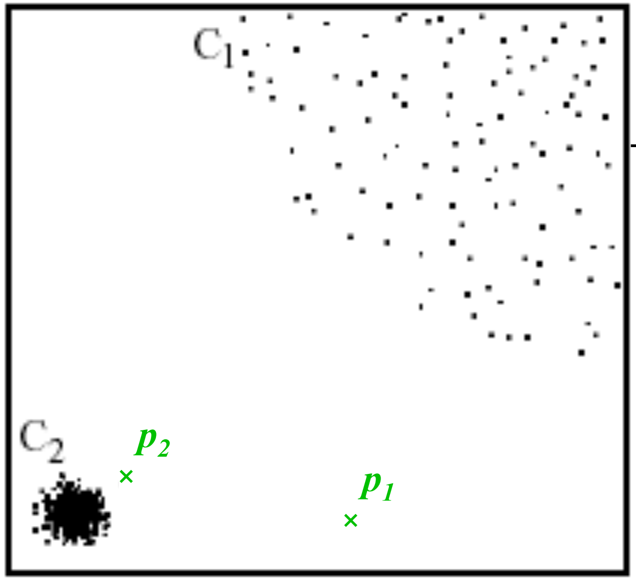
Outliers in Higher Dimensions

- In high-dimensional space, data is sparse and notion of distance becomes meaningless
  - Every point is an almost equally good outlier from the perspective of distance-based definitions

- Lower-dimensional projection methods
  - A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density

Density-based: LOF approach

- *Idea:* An object is an anomaly if it is in a region of low density.
- For each point, compute the density of its local neighborhood.
- *Def.* density= reciprocal of average distance to the k nearest neighbors.
  Or: density as in DBSCAN
- For a point x, compute its local outlier factor (LOF,or “relative density”) as the ratio of the density of x and the avg.density of its nearest neighbors
- Outliers are points with largest LOF value



In the nearest neighbors approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Complexity: O(N^2), can be reduced to O(N log N) with nearest neighbors data structures

Density-based: LOF approach

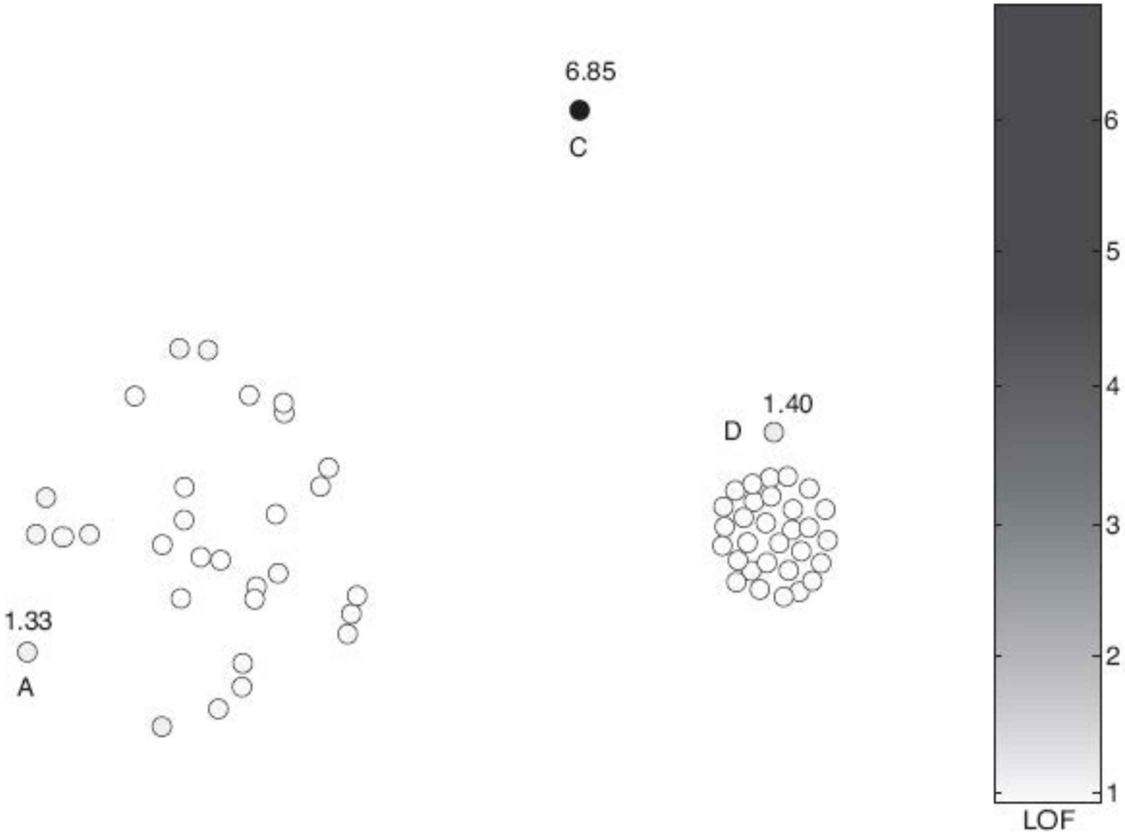
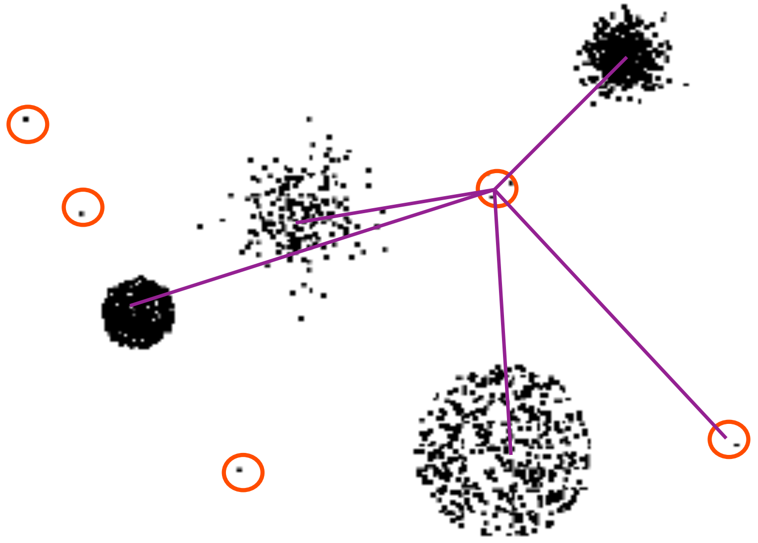


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

Clustering-Based (1)

- *First idea:* "Small" clusters that are "far" away from other clusters are outliers

- *Approach:*
  1. Cluster the data
  2. Choose points in small clusters as candidate outliers
  3. Compute the distance between candidate points and non-candidate clusters.
     - If candidate points are far from all other non-candidate points, they are outliers



– *But:* highly sensitive to number of clusters, no outlier score available

Clustering-Based (2)

- *Idea:* An object is a cluster-based outlier if it does not strongly belong to any cluster
- E.g., for prototype-based clustering like K-means, for an object:
  - Measure the (absolute) distance to its closest centroid
    - But: confusing values in presence of different densities
  - *Better:* Consider relative distance to closest centroid, i.e. the ratio of its absolute distance to average absolute distance of all points in its cluster

Clustering-Based (2)

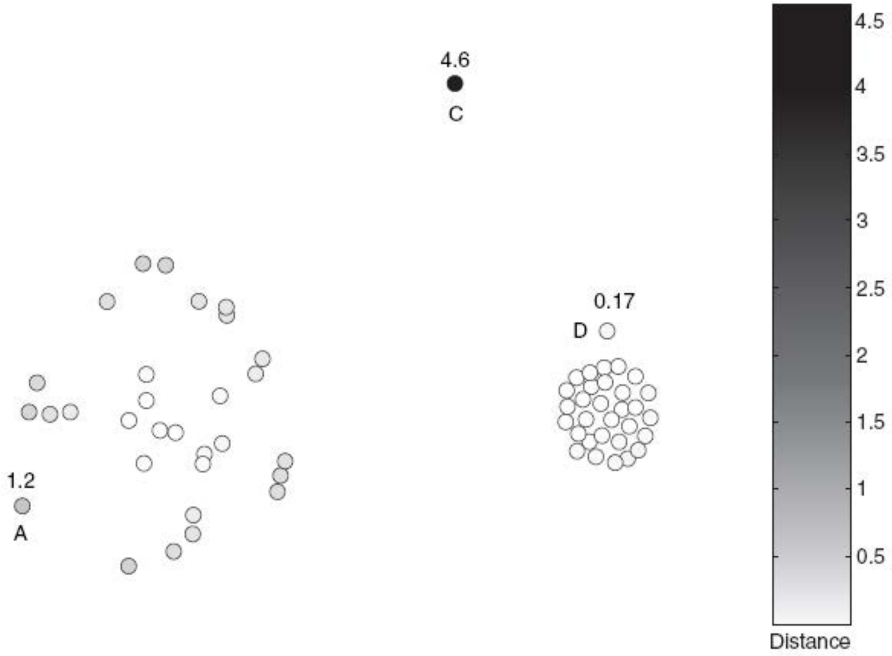


Figure 10.9. Distance of points from closest centroid.

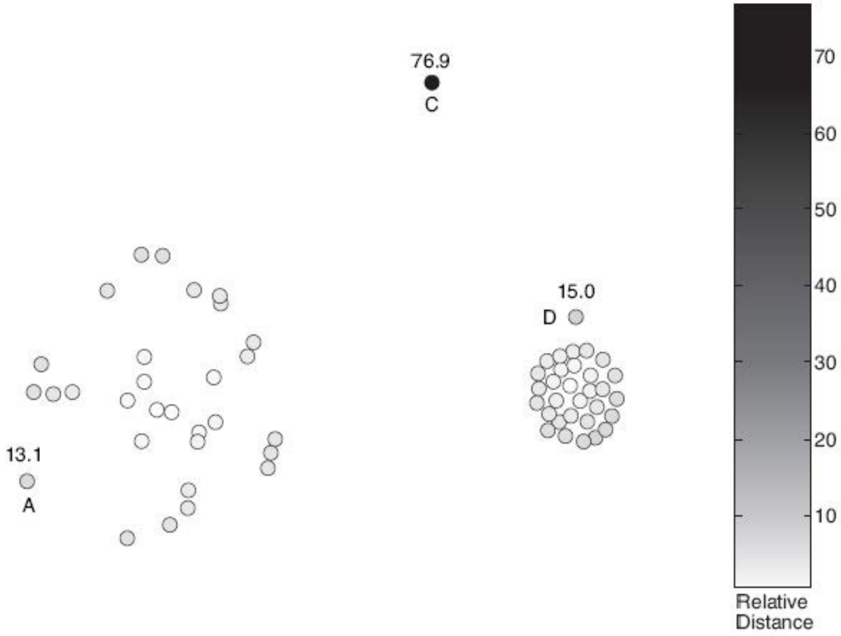
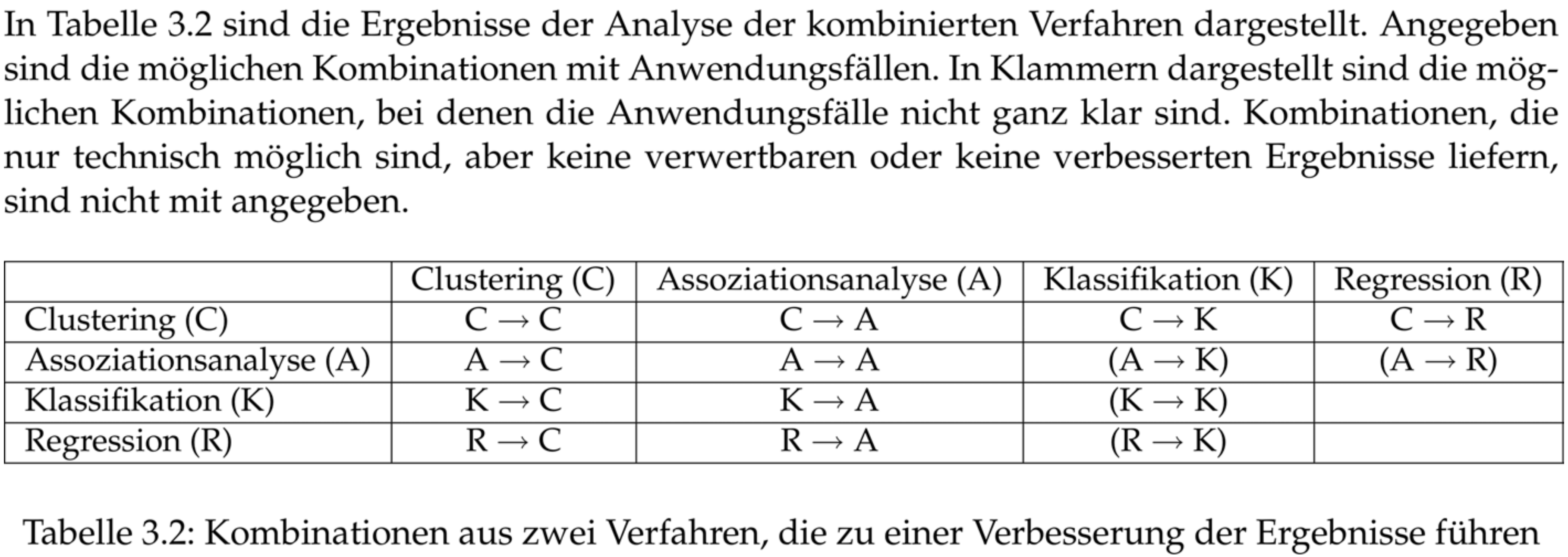


Figure 10.10. Relative distance of points from closest centroid.

Kombination von Data Mining Verfahren

- z.B. Clustering ➔ Klassifikation
  - Klassifikation bzgl. Clusterattribut soll ein Modell für die Einordnung von Objekten in Cluster liefern
  - ==>> Klassifikationsregeln = Clusterbeschreibungen
- z.B. Assoziationsanalyse ➔ Clustering
  - Clustering der Assoziationsregeln zwecks besserer Übersicht / Zusammenfassung in Clustern (erfordert geeignete Attribute für Regeln, wie Prämisse, Konklusion, Confidence)
  - oder Clustering der Frequent Itemsets (erfordert Distanzen zwischen Itemsets)
- z.B. Klassifikation/Clustering ➔ Assoziationsanalyse
  - erlaubt Einbeziehung von Klassifikationsattribut/ Cluster-Id in Assoziationsregeln
- z.B. Clustering ➔ Assoziationsanalyse pro Cluster
  - zur Verbesserung der Assoziationsregeln, *s.Extrafolien*

Kombination von Data Mining Verfahren



In Tabelle 3.2 sind die Ergebnisse der Analyse der kombinierten Verfahren dargestellt. Angegeben sind die möglichen Kombinationen mit Anwendungsfällen. In Klammern dargestellt sind die möglichen Kombinationen, bei denen die Anwendungsfälle nicht ganz klar sind. Kombinationen, die nur technisch möglich sind, aber keine verwertbaren oder keine verbesserten Ergebnisse liefern, sind nicht mit angegeben.

	Clustering (C)	Assoziationsanalyse (A)	Klassifikation (K)	Regression (R)
Clustering (C)	$C \rightarrow C$	$C \rightarrow A$	$C \rightarrow K$	$C \rightarrow R$
Assoziationsanalyse (A)	$A \rightarrow C$	$A \rightarrow A$	$(A \rightarrow K)$	$(A \rightarrow R)$
Klassifikation (K)	$K \rightarrow C$	$K \rightarrow A$	$(K \rightarrow K)$	
Regression (R)	$R \rightarrow C$	$R \rightarrow A$	$(R \rightarrow K)$	

Tabelle 3.2: Kombinationen aus zwei Verfahren, die zu einer Verbesserung der Ergebnisse führen

basiert auf Masterarbeit von Ulf Mewe, 2008

Vergleich von Data Mining Verfahren

- *Zum Nachdenken:*

  (Wann) Kann man

  - Assoziationsanalyse
  - Klassifikation
  - Clustering

  gegenseitig austauschen ?

  Welche Verfahren sind datenbank-
  (big data) -tauglich?