Data Mining Anomalieentdeckung

Jackelyn Wirri Miranda

Inhalt

- Einleitung
- Definition Anomalieentdeckung
- Ursache eines Ausreißers
- Merkmale von Ausreißer
- Anwendungen
- Variante von Anomalieentdeckung
- Anforderungen und Probleme
- Verfahren :

  → Grafik (2D-, Konvexe Hülle- Ausreißeranalyse)

  → Statistikbasierte

  → Distanzbasierte (K-Nächste Nachbarn, Distanzbasierte, Dichtebasierte, Clusterbasierte)
- Inhaltverzeichnis

Sinn und Nutzen von Ausreißer Entdeckung

„Das Rauschen für den einen ist für den anderen ein Signal.“

Im Rahmen der vielfältigen Betrachtung von Wissensentdeckung in Datenbanken, allgemein auch als KDD – Knowledge Discovery in Databases – bezeichnet, wurden Outlier eine lange Zeit im Bereich des maschinellen Lernens und des Data Mining von existierenden Anwendungen und ihren Algorithmen nur insoweit betrachtet, als dass sie gegenüber diesen Erscheinungen tolerant, bzw. robust waren. Es gibt jedoch eine breite Palette von Anwendungen, für die gerade das Wissen um außergewöhnliche Ereignisse und deren systematische Entdeckung von immenser Bedeutung ist.

Definition

- Entdeckung von Ausreißen (Outlier detection).
- Was ist ein „Ausreißer?
- keine allgemeine akzeptierte Definition
- [Hawkins1980]: „Ein Ausreißer ist eine Beobachtung, die sich von den anderen Beobachtung so deutlich unterscheidet, dass man denken könnte, si sei von einem anderen Mechanismus generiert worden “.
- [Lewis] Informell als Beobachtungen definiert, welche zum Rest einer Datenmenge inkonsistent erscheinen.
- Anders als die andere Objekte der Datenmenge.
- Starke Unterscheidung von anderen Datenobjekten.

Hawkins-Definition

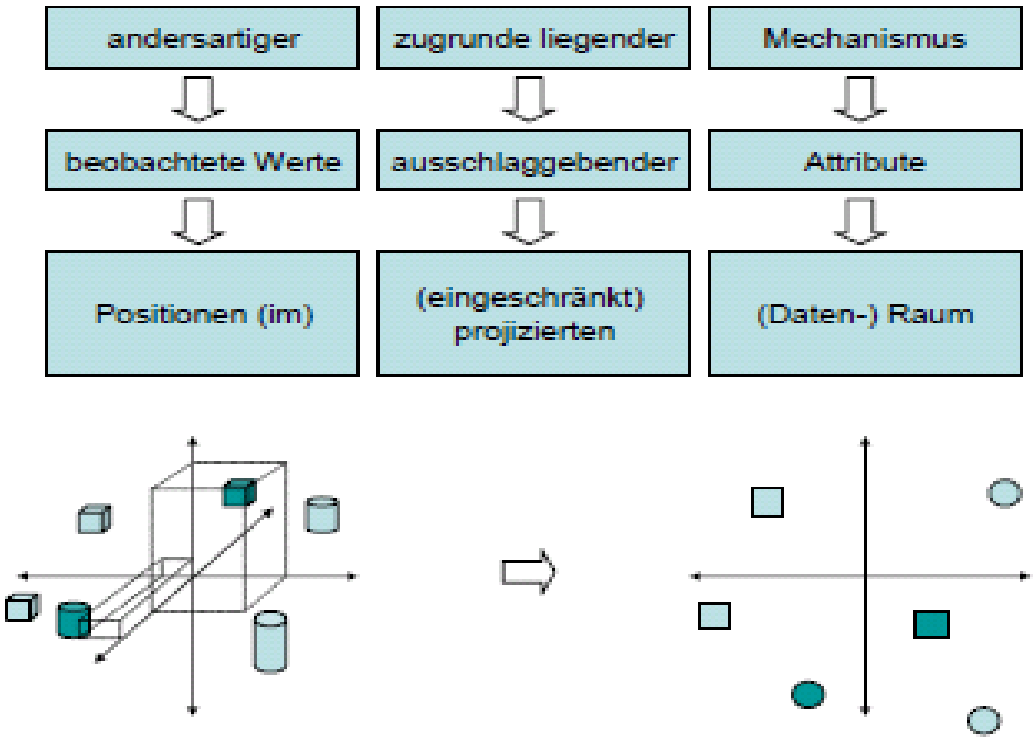


Abbildung 1 - Hawkins Definition von Outliern in Datenmengen

Abbildungsbeschreibung: Die Definition von Outliern nach Hawkins [45] wird grafisch gezeigt und drückt aus, dass sich Mechanismen, welche dem Verhalten von Objekten zugrunde liegen, in der Position dieser Objekte im Datenraum niederschlagen, der durch die Attribute des Objektes aufgespannt wird.

Ursachen eines Ausreißers

Drei Möglichen Ursachen für das Auftreten des Ausreißers:

1. Durch einen Verfahrenstechnischen Fehler verursacht.

   Bsp. Dateieingabe, Codieren der Daten, technischen Ausfall bei der Datenerfassung bzw. Speicherung.

2. Kennzeichnet einen Außergewöhnlichen Wert.

   Bsp. Eine einzelne aus dem Rahmen fallende Beobachtung (der einzige befragte Millionär).

   Anmerkung: mitunter können solche Ausreißer auch ein Hinweis sein, dass die Befragung falsch angelegt wurde und daher nicht repräsentativ ist.

3. Kennzeichnet einen korrekt erfassten außergewöhnlichen wert, für den es keinerlei Erklärung gibt.

Merkmale von Ausreißer

Ausreißer haben mehrere Dimensionen, die sowohl in Kombination, als auch allein auftreten können:

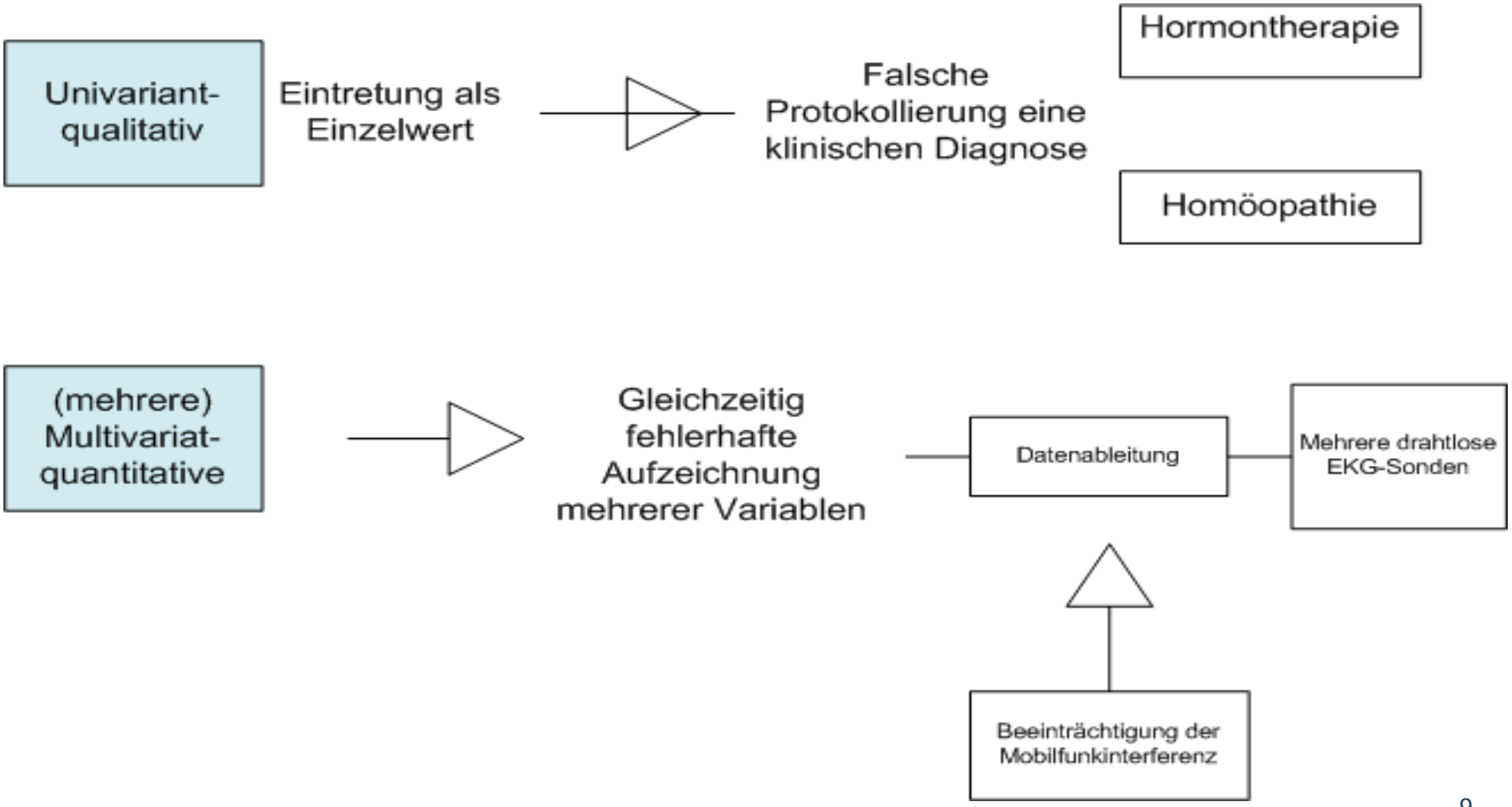
1. Normaler Ausreißer (Univariat)= außergewöhnlich großer oder kleiner Wert (persönliches Einkommen im Millionenbereich).
2. Multivariater Ausreißer (Syn. Hochdimensional) = für sich betrachtet im normalen Bereich liegende Einzelwerte, die in ihrer Kombination quer durch die Variablen einen einzigartigen Fall ergeben (86 jährige Frau mit Internetanschluss).

Merkmale von Ausreißer

3. Semantisch (qualitativ) oder formell (quantitativ) auffallen.
4. Nur bei einem Fall, aber auch in bestimmten Gruppierung auftreten.
5. Vereinzelt aber auch massiv auftreten.
6. Relativ zur Datenmenge (Stichprobengröße) sein.
7. Unterschiedliche Ursachen haben.

Beispiel-Merkmale von Ausreißer

- Ausreißer können mehrere Gesichter haben:



Anwendungen

- Telefonkunden-Betrug
- Medizinische Analyse
- Betrugsversuchen mit Kreditkarten

  - Erkennung von Datenobjekten mit vielen Zahlung großer Beträge innerhalb eines Kurzen Zeitraumes.

- Network-Intrusion-Detection-Systeme (NIDS)

  Pakete + Eigenschaften = Objekte

  - Mögliche Angriffe auf das Netzwerk öffentlichen

- Unterstützung der Fehlerdiagnose in technischen Systemen.

  Ausreißer=Fehler im System

  - Beschreibung das normale Verhalten in einem System

Anwendungen

- Untersuchung der Transaktionen beim Einsatz von Kreditkarten oder ähnlichen Zahlungsmitteln (z.B. SmartCards)

Ziel:

- Missbrauch zu identifizieren und erfolgreich zu unterbinden.

Vorgang

- Die Unterscheidung zwischen normalen und außergewöhnlichen Transaktionsmustern → Kreditkartenfirmen

Vorteil

- schnell und zielgerichtet einzugreifen und die Kosten von missbräuchlicher Verwendung einzudämmen, Täter ggf. zu identifizieren und trotzdem dem Anwender einen normalen, in Bezug auf diese Aspekte transparenten Zahlungsverkehr zu gewährleisten.

Anwendungen

•Die Nutzung von Telefonverbindungen oder Mobilfunkanschlüssen die Identifizierung der Infiltration von Netzwerken (Intrusion Detection and Prevention)

•Die Analyse von Verkehrsmustern im Internet zur Vermeidung von Denial of Service Attacken (DoS)

•eCommerce Kriminalität im allgemeinen Sinn

•Wahl- und Steuerbetrug (z.B. über die IDEA Software der Prüfer des Finanzamtes)

•Strategischen Vorteil durch Wissensgewinn führen:
Die Identifizierung von Ausnahmesportlern in diversen Sportarten und ihren Ligen →Wissen um die extraordinären Fähigkeiten von Menschen schnell und effizient zu erwerben.

Variante von Anomalieentdeckung

Drei Varianten:

Auf einem gegebenen Datenbestand D finden

- Alle Objekte d E D ← über einem gegebenen Schwellenwert liegende Werte besitzen.

- N Objekte d E D ←am stärksten einem gegebenen Schwellenwert liegende Werte besitzen.

- D (größten Teil normalen Objekten + Testobjekt d) →Berechnung von Anomalienwert.

Anforderungen des Verfahren

- Skalierbar
- Lieferung dem Ergebnis auf einer großen Datenmenge in angemessener Zeit
- Lieferung nur den Objekten, die wirklich Ausreißer sind.

Probleme

- Identifizierte Daten Objekte als Ausreißer:

  Wahrheit

  - Korrektheit
  - Beschreibung der Realweltobjekte

„Die entscheidende Frage der Ausreißeranalyse lautet: Werden die Ausreißer beibehalten oder verworfen“

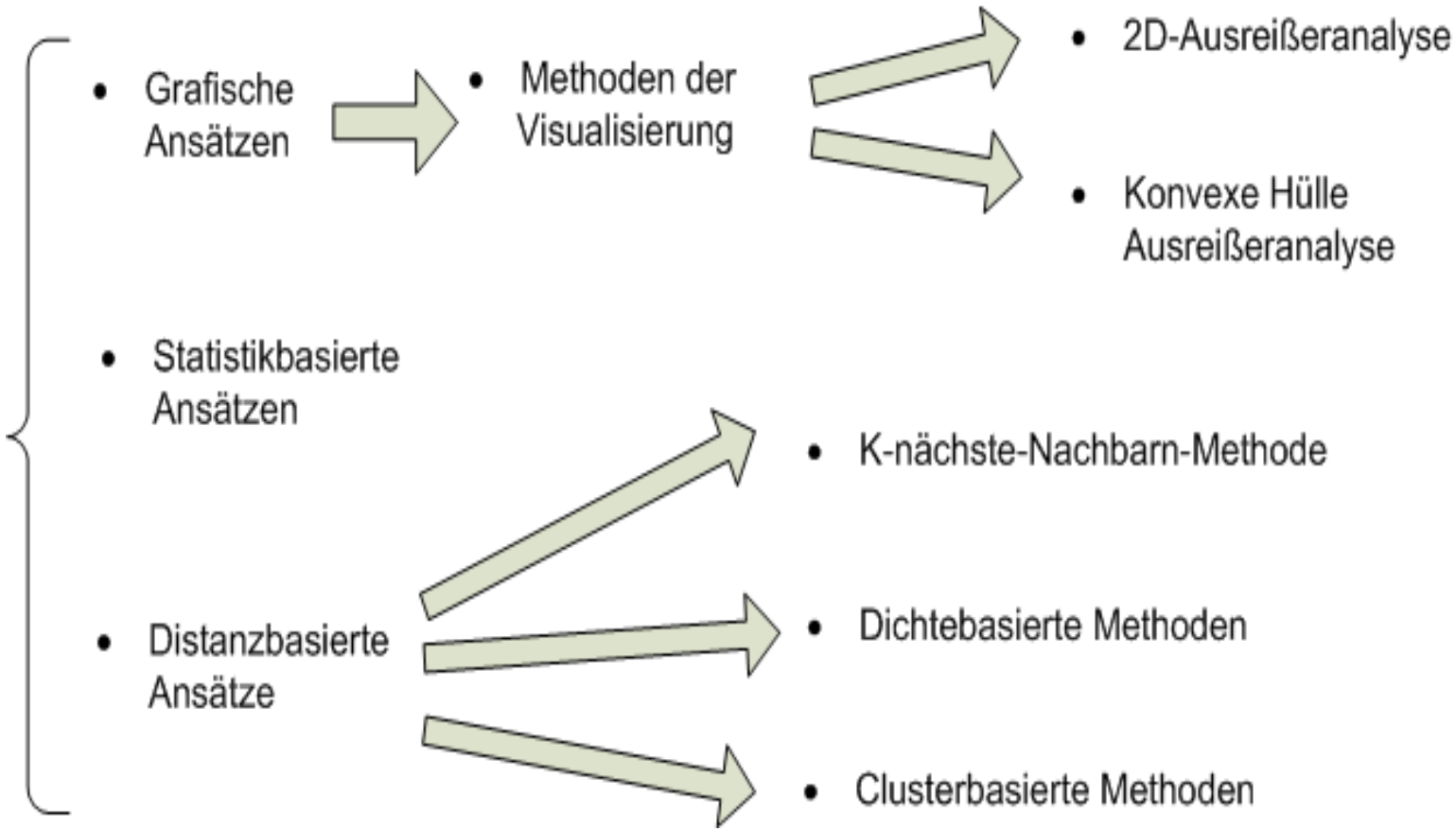
Verfahren

Zwei prinzipiellen Schritten:

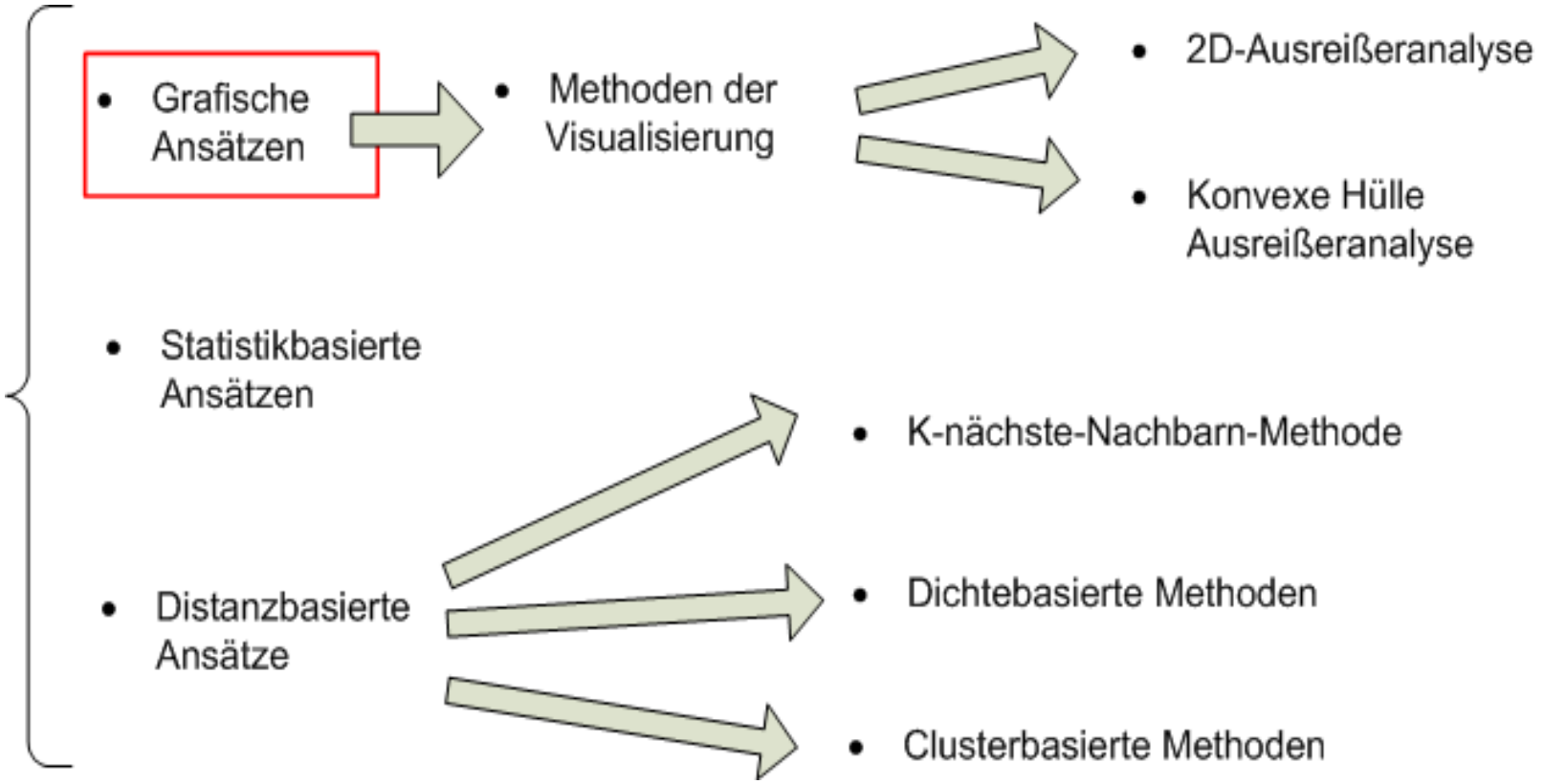
1. Erstellung ein Profil (normale) Objekte E D ← Form eines Musters/statischen Wertes.
2. Entdeckung des Ausreißer durch Verwendung des Profils →normale Objekte vom Profil Abweichen

-eine bestimmte Richtung verlassen

Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Grafische Verfahren

- Bei grafischen Ansätze:

  Erkennung des Ausreißers

- Visualisierung von Daten Objekten
- Signifikante Unterscheidung von anderen Objekten aufgrund des optischen Erscheinungsbildes .

Es gibt zwei Methoden:

1. 2D Auftragen de Datenobjekten
2. Konvexer Hülle

Grafische Verfahren

Methoden:

1. 2D Auftragen de Datenobjekten
2. Konvexer Hülle

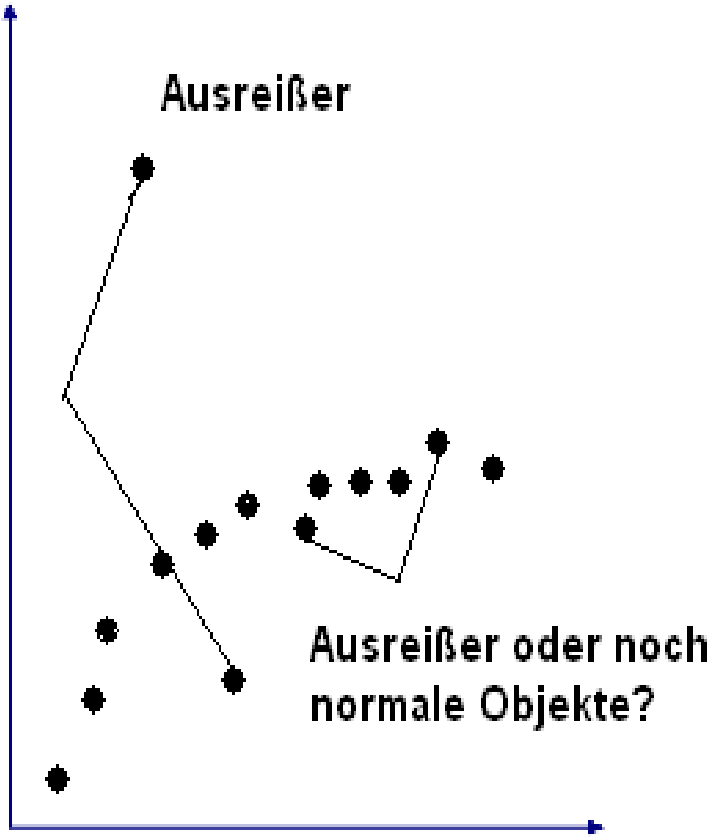
Probleme

-Subjektivität d.h. Unterschiede Betrachtungen durch Benutzern

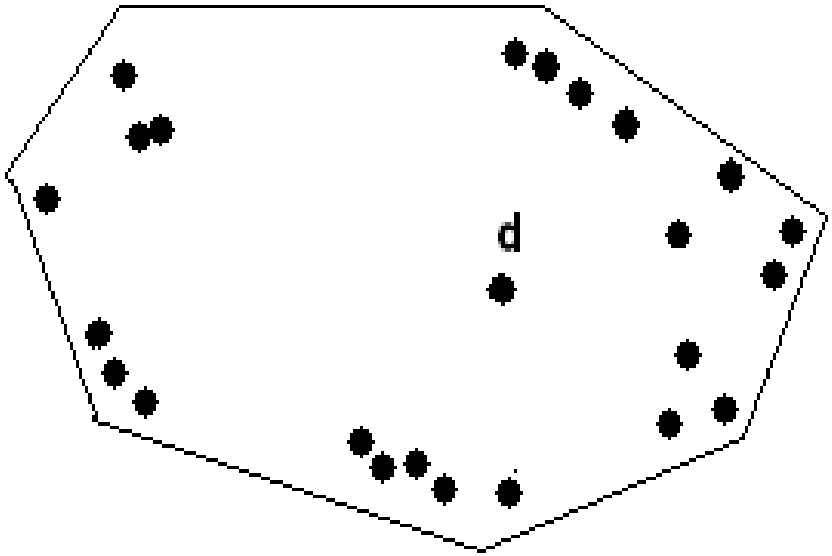
-Ausreißer im Zentrum der Daten

Methoden der Visualisierung

2D-Ausreißeranalyse



konvexer Hülle

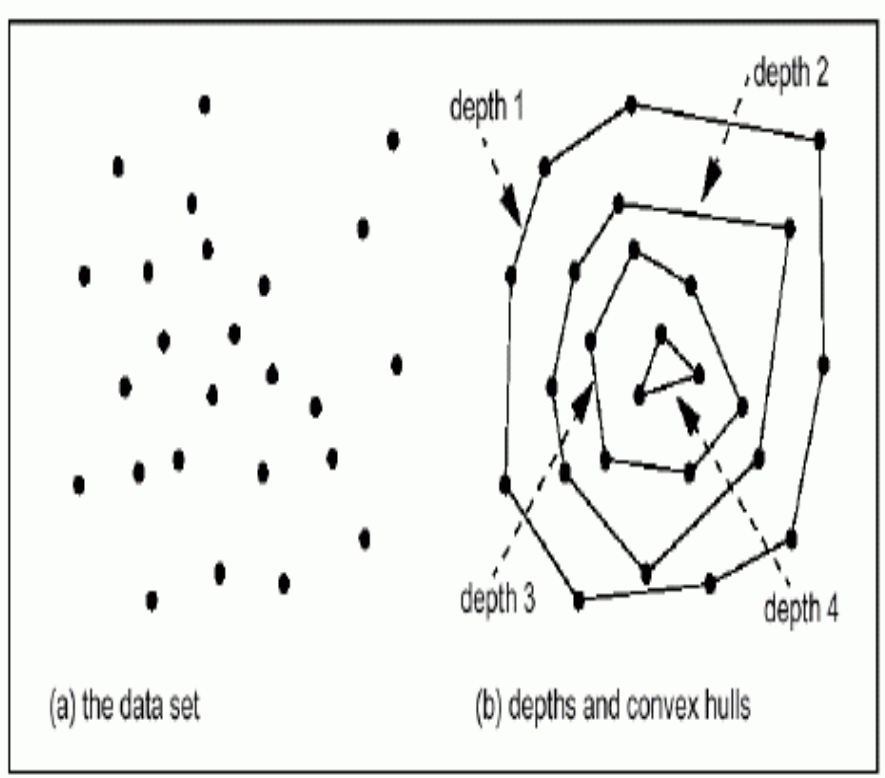


Konvexe Hülle

Identifikation von Ausreißer:

Depth-Based:

- Jedes Objekt wird als ein Punkt in einem d-Dimensionalesn raum betrachtet
- Die Tiefe der Objekte wird berechnet.
- Ausreißer haben eine kleinere Tiefe als die andere Objekte.



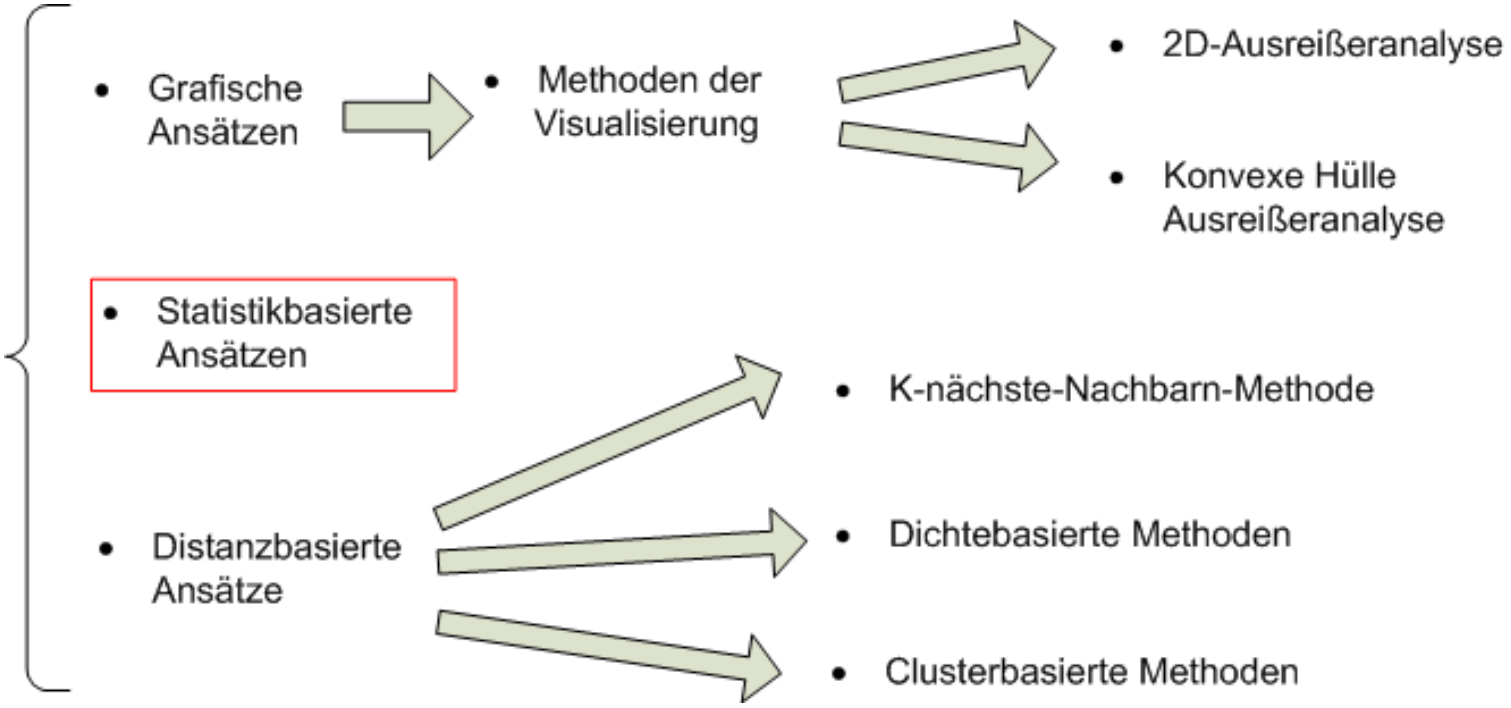
Theoretisch:

Gut auch für hoch.dim, Daten

Praktisch:

Ineffizient für d>= 4, da die Konvexe Hülle berechnet werden muss.

Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Statistikbasierte Verfahren

- Unterstellung einem Model der Verteilung der Datenobjekte. (z.B. Normalverteilung)
- Abhängig:
- Dieser Verteilung
- Dem Parameter Verteilung (Median/Varianz)
- Anzahl der erwarteten Ausreißer

→Durchführung ein statischer Test

- **Problemen:**
  - Die Verteilung der Objekten häufig nicht bekannt sind
  - meisten Test Beziehung nur auf ein Attribut

Statistikbasierte Verfahren

Idee

- Modelliere Daten als multivariate Normalverteilung
- Punkte deren Abstand (quadratische Formdistanz) zum Mittelwert μ größer als Grenzwert Θ (z.B. Θ = 3.σ) ist, sind Ausreißer Multivariate Normalverteilung

$$N(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2}[(x-\mu)^T \Sigma^{-1}(x-\mu)]}$$

Quadratische Formdistanz (Mahalanobis Distanz) des Punktes x vom Mittelwert μ der Normalverteilung

Statistikbasierte Verfahren

- Quadratische Formdistanz
  - Die quadratischen Formdistanzen der Punkte zum Mittelwert der Normalverteilung folgen einer χ2 (Chi-Square)-Verteilung mit *d* Freiheitsgraden (*d = Dimensionalität des Datenraums)*
- Algorithmus zur Erkennung multivariater Ausreißer
  - Input: *d-dimensionale Punktmenge DB*

$$\mu_{DB} = \frac{1}{|DB|} \sum_{x \in DB} x$$

•Berechne den Mittelwert μ*DB aller Punkte*

•Berechne die (*d x d) Kovarianzmatrix ΣDB aller Punkte*

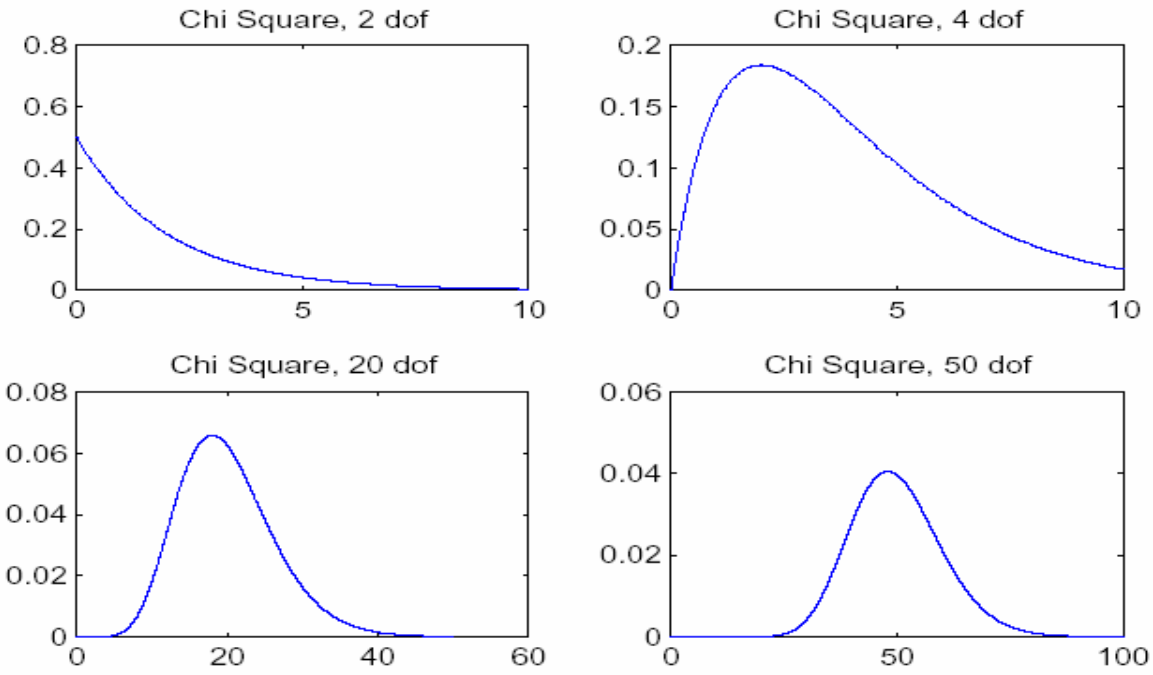
•Berechne für jeden Punkt *x∈DB die quadratische* Formdistanz von *x zum Mittelwert μDB*

$$D(x, \mu_{DB}) = (x - \mu_{DB})^{T} \Sigma^{-1} (x - \mu_{DB})$$

-Output: alle Punkte *x, deren Abstand zum Mittelwert größer als χ2(0,975)* ist $\text{OutlierSet} = \{x \in DB \mid D(x, \mu_{DB}) > \chi^2(0{,}975)\}$

Statistikbasierte Verfahren

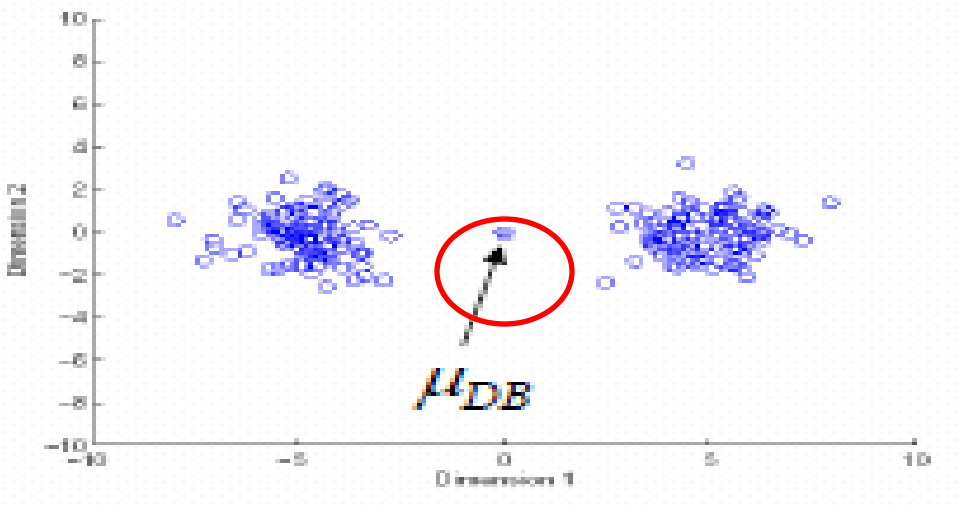
- Probleme
  - “Curse of Dimensionality”
- Distanzen werden in hochdimensionalen Räumen unaussagekräftig
- Je höher die Dimensionalität des Datenraums (Freiheitsgrade der Verteilung), desto ähnlicher werden die quadratischen Formdistanzen



Statistikbasierte Verfahren

- Probleme (cont.)
  – Robustheit
    • Mittelwert und Varianz/Kovarianz extrem sensitiv gegenüber Ausreißern
    • Verwendung der quadratische Formdistanz zur Outlier-Entdeckung obwohl diese Distanz selbst durch Ausreißer beeinflußt ist (da abhängig von der Kovarianzmatrix) => Minimum Covariance Determinant [Rousseeuw, Driessen 99] minimiert den Einfluss von Ausreißern auf die quadratische Formdistanz

  – Flexibilität
  • Datenverteilung muss vorher bekannt sein
  • Keine “Mixture of Gaussians”
  • Beispiel: Mittelwert der Daten ist ein Ausreißer!!!

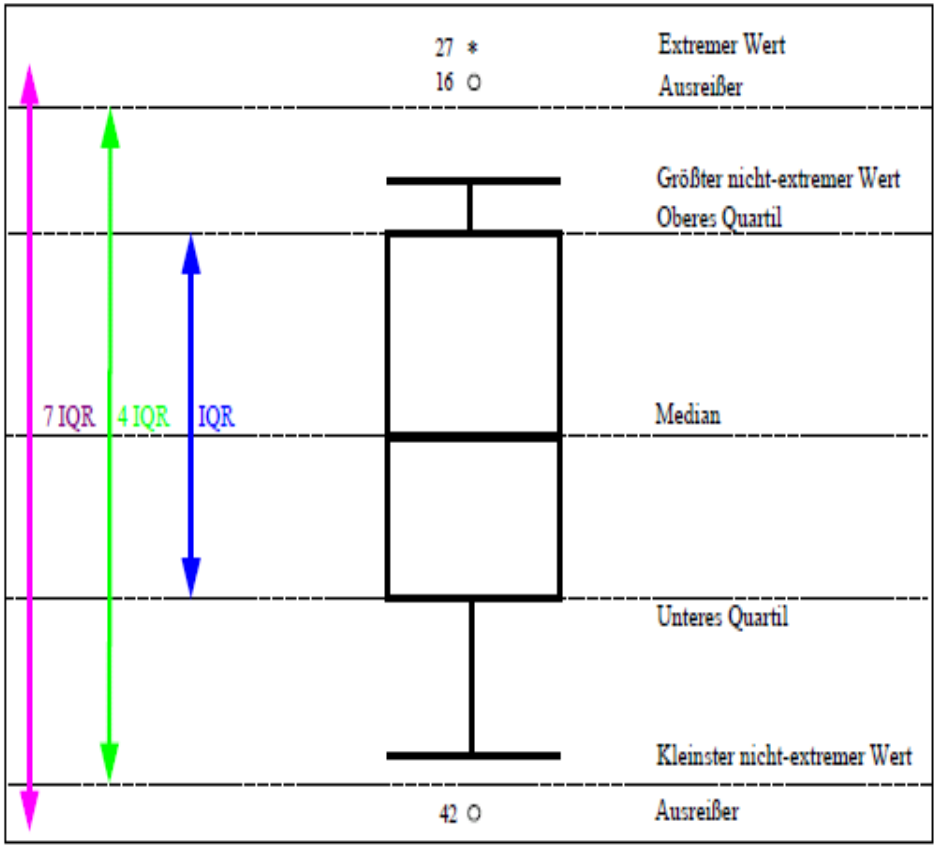


Grafik: Box-Plots

Zum Beispiel:

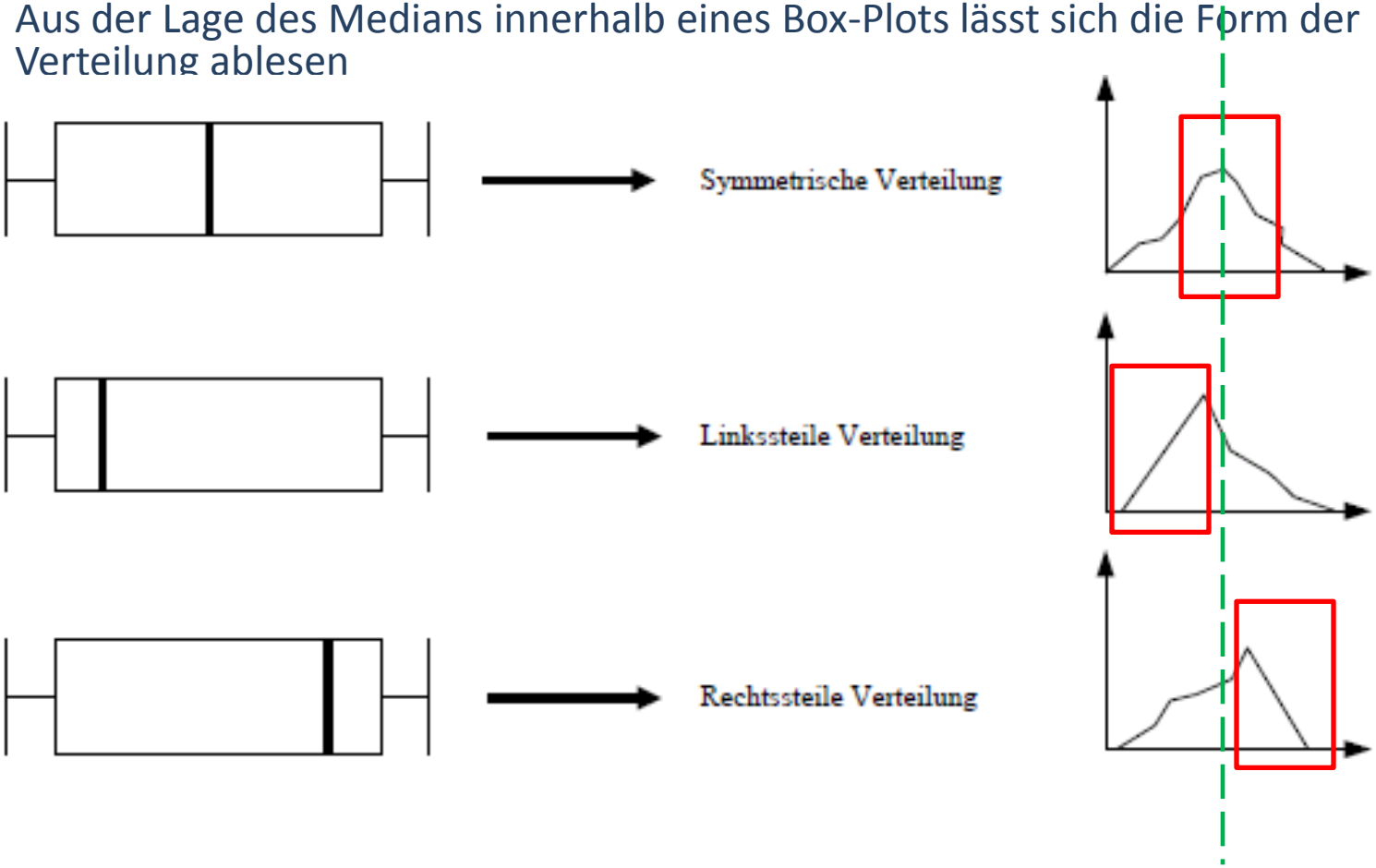
beim Box-Plot werden alle Werte außerhalb des dreifachen IQR-Bereichs um den Median als Ausreißer klassifiziert

- Box-Plots bieten einen direkten Verteilungsüberblick und eignen sich insbesondere zum Verteilungsvergleich
- Sie stellen sowohl Lage als auch Streuung der Verteilung dar und dienen zudem der Identifikation von Ausreißern

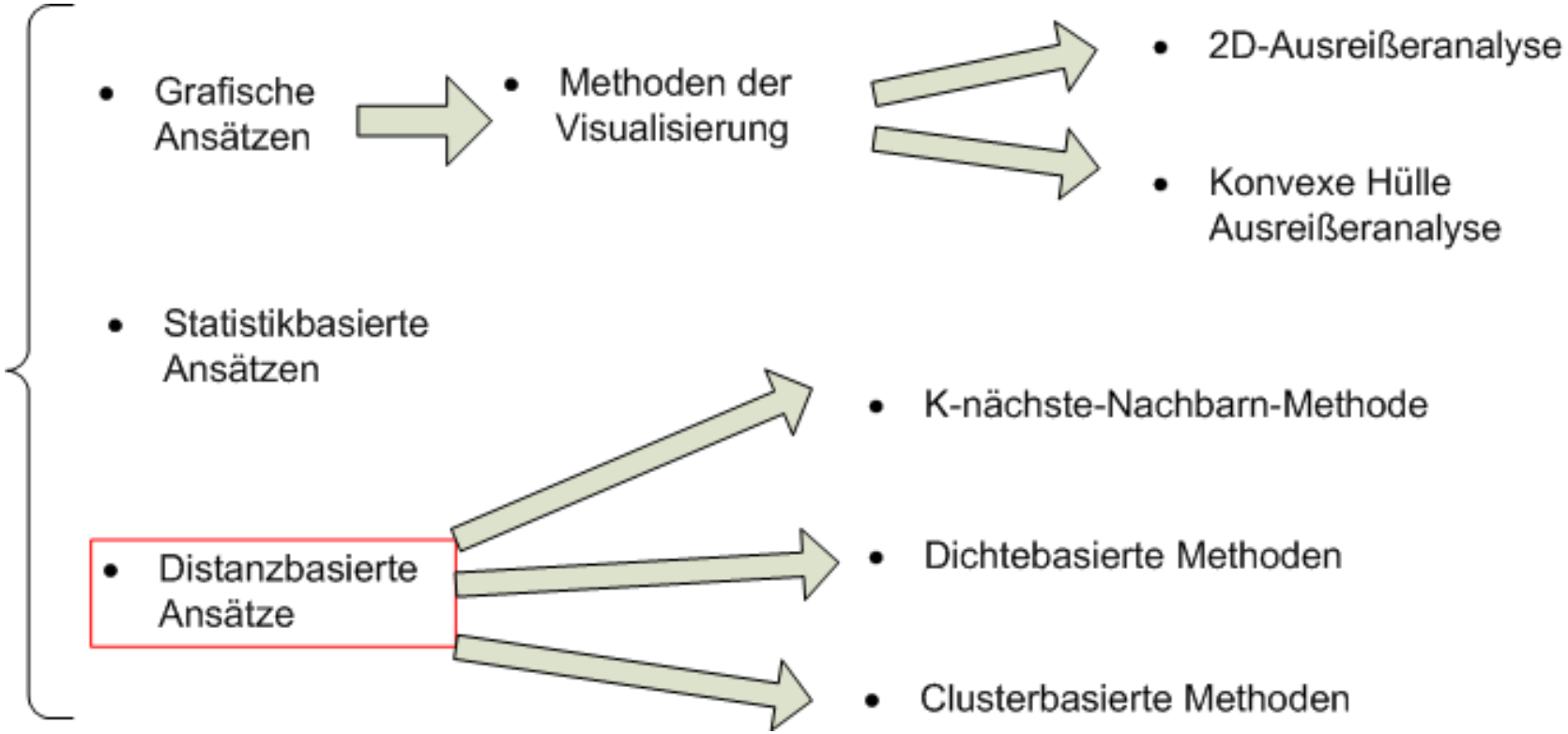


Grafik: Box-Plots

- Aus der Lage des Medians innerhalb eines Box-Plots lässt sich die Form der Verteilung ablesen



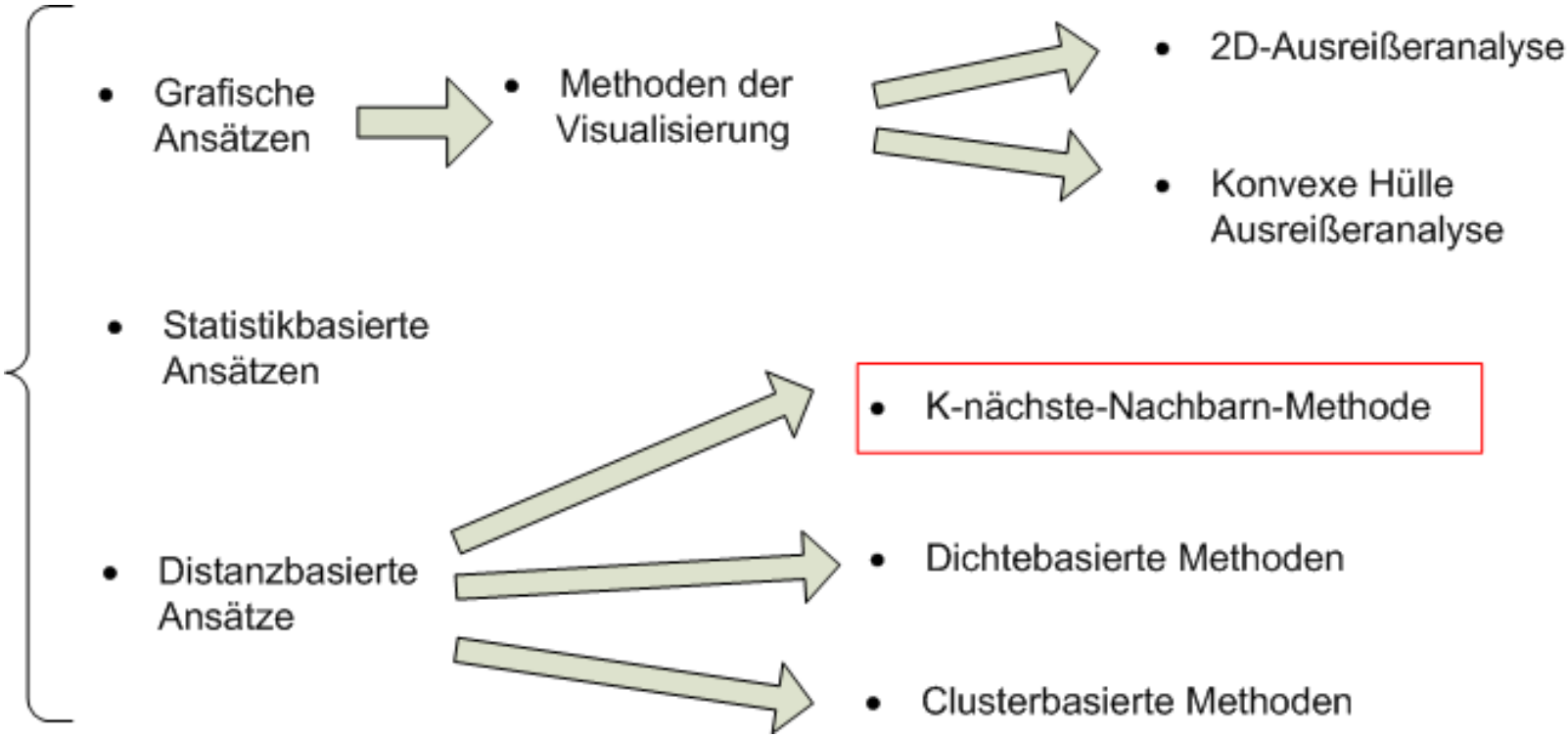
Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Distanzbasierte Ansätze

- Auffassung den Daten als Vektor von Features
- Anwendung der Folgende Methoden zur Entdeckung der Ausreißer :
  - Knächste-Nachbarn.Methode
  - Dichtebasierte Methoden
  - Clusterbasierte Methoden

Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

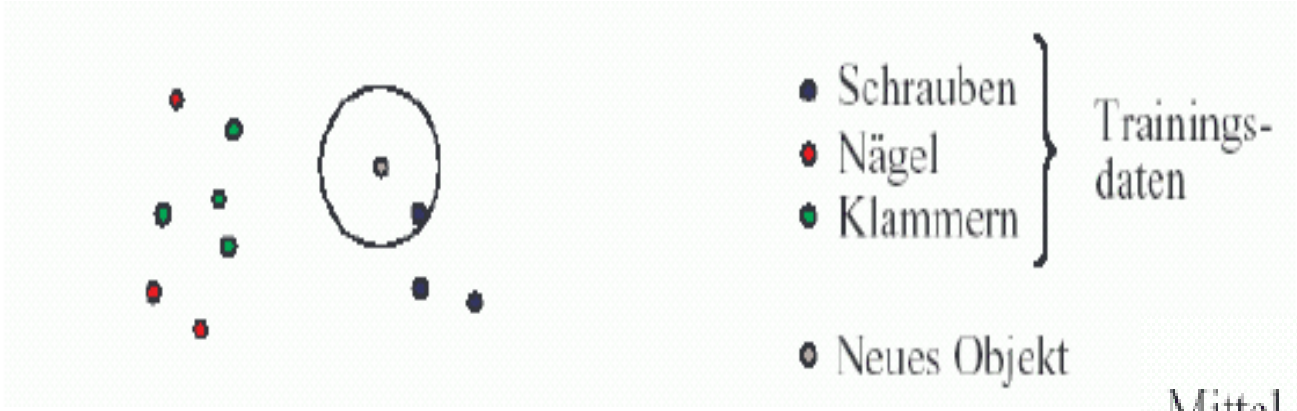
K-Nächste-Nachbarn (KNN)

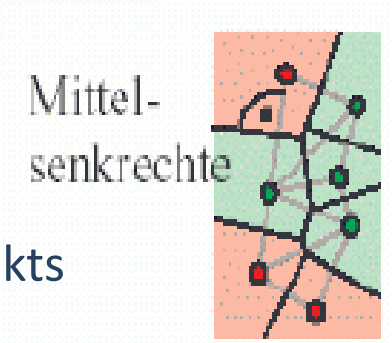
- Die Distanz zwischen allen Paaren von Objekten ermittelt.
- Definition von Ausreißer aus verschiedenen Art und Weise:

  - Objekte, für die es weniger als p Nachbarn bez. Einer vorgegebenen Distanz gibt.
  - Die top n Objekte, deren Distanz zum k-ten nächsten Nachbarn am Größten ist.
  - Die Top n Objekte, deren Durchschnittliche Distanz zu den k nächsten Nachbarn am größten ist.

K-Nächste-Nachbarn

- hier wieder Datensatz mit Klassen vorgegeben Beispiel:





- Instanz basiertes Lernen
- Einfachsten Nächste-Nachbar-Klassifikator
- Zuordnung zu der Klasse des nächsten Nachbarnpunkts

- Im Beispiel: Nächster Nachbar ist eine Schraube
- Regionen der Klasse Ordnung können als Voronoi-Diagramme dargestellt werden.

K-Nächste-Nachbarn

Fehlzuordnung durch Ausreißer möglich ⇒

-Besser: Betrachte mehr als nur einen Nachbarn

Entscheidungsmenge

- die Menge der zur Klassifikation betrachteten k nächsten Nachbarn

Entscheidungsregel

- Wie bestimmt man aus den Klassen der Entscheidungsmenge die Klasse des zu klassifizierenden Objekts?

→ Interpretiere Häufigkeit einer Klasse in der Entscheidungsmenge als Wahrscheinlichkeit der Klassezugehörigkeit.

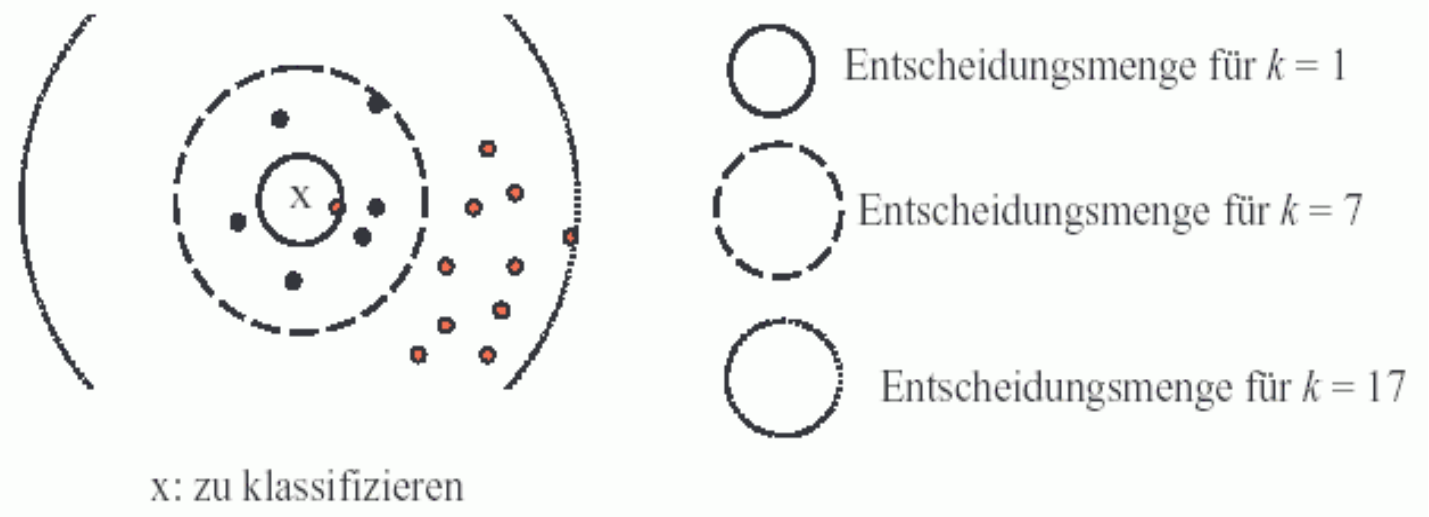
→Maximum –Likelihood-Prinzip: Mehrheitsentscheidung

Ggf. Gewichtung

K-Nächste-Nachbarn

Wahl des Parameters K

- „zu kleines“ k: hohe Sensitivität gegenüber Ausreißern.
- „zu großes“ k:viele Objekte aus anderen Clustern (Klassen) in der Entscheidungsmenge.
- Mittleres k: höchste Klassifikationsgüte, oft 1<k<10



K-Nächste-Nachbarn

Entscheidungsregel

Standardregel:

- wähle die Mehrheitsklasse der Entscheidungsmenge.

Gewichtete Entscheidungsmenge

- gewichte der Klasse de entscheidungsmenge

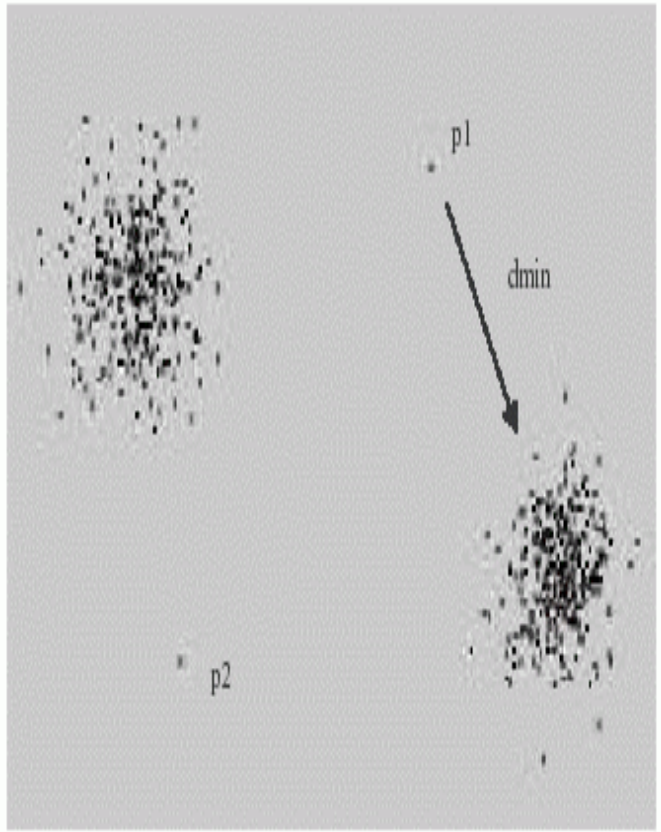
- nach Distanz, meist invers quadriert: $weight\ (dist) = 1/dist^2$
- nach Verteilung der Klassen (oft sehr ungleich!)
  Problem: Klasse mit zu wenig Instanzen ($< k/2$) in der Trainingsmenge bekommt keine Chance, ausgewählt zu werden, selbst bei optimaler Distanzfunktion
  - Klasse A: 95 %, Klasse B 5 %
  - Entscheidungsmenge = {A, A, A, A, B, B, B}
  - Standardregel $\Rightarrow$ A, gewichtete Regel $\Rightarrow$ B

Distanzbasierte Ansätze

Definition "*(pct,dmin)-Outlier" [Knorr, Ng 97]*

Ein Objekt p in einem Datensatz D ist ein DB(pct, dmin)-Ausreißer (DB=distance -Based), falls mindestens pct-Prozent von Objekten aus D eine größere Distanz als dmin zu p haben.

-Wahl von pct und dmin wird einem Experten überlassen.



Beispiel: p1 gehört D, pct=0.95, dmin=8

DB(0.95,8)=> 95% von Objekten aus D haben eine Distanz > 8 zu p1

Anmerkung: Jedoch diese Definition ist für viele Zwecke zu unflexible (Lösung Dichte basierte Ausreißer)

Distanzbasierte Ansätze

Alternative Definitionen:

– „(k,dmax)“-Outlier *[Kollios, Gunopulos, Kiudas, Berchtold 03]*

Ein Objekt p in einem Datensatz DB ist ein (k,dmax)-Outlier, falls höchstens k Objekte aus DB eine kleinere Distanz als dmax zu p haben.

– kNN-Outlier *[Ramaswamy, Rastogi, Shim 03]*

Die n Objekte in DB mit den höchsten k-nächste-Nachbar-Distanzen sind Outlier

Beispiel Probleme Distanzbasierte Ansätze

Probleme (siehe Beispiel)

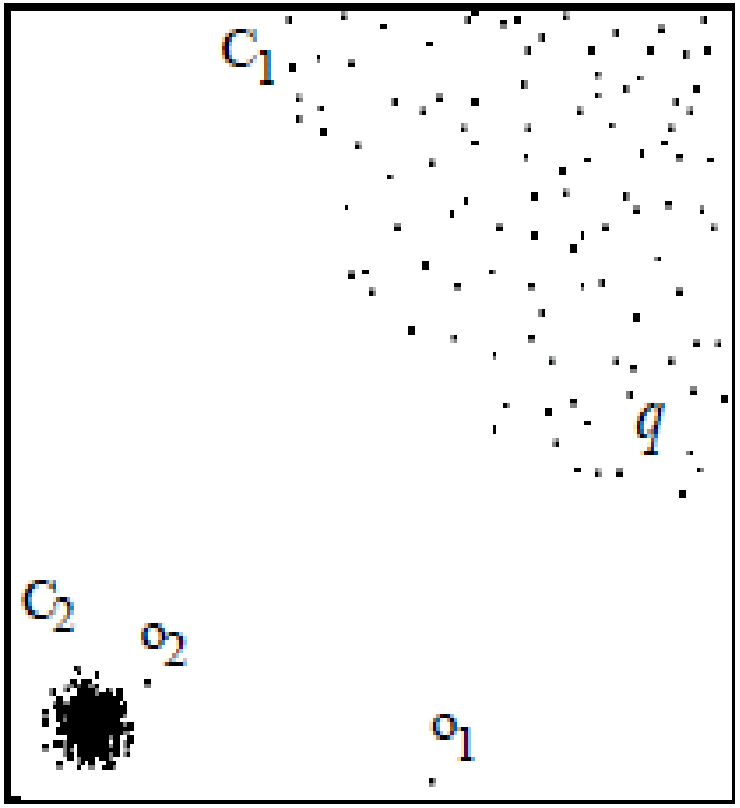
– (pct,dmin)-Outlier: welche Werte sollen pct und dmin annehmen, so daß o2 ein Outlier ist, nicht aber die Objekte

des Cluster C1 (z.B. q Î C1)?

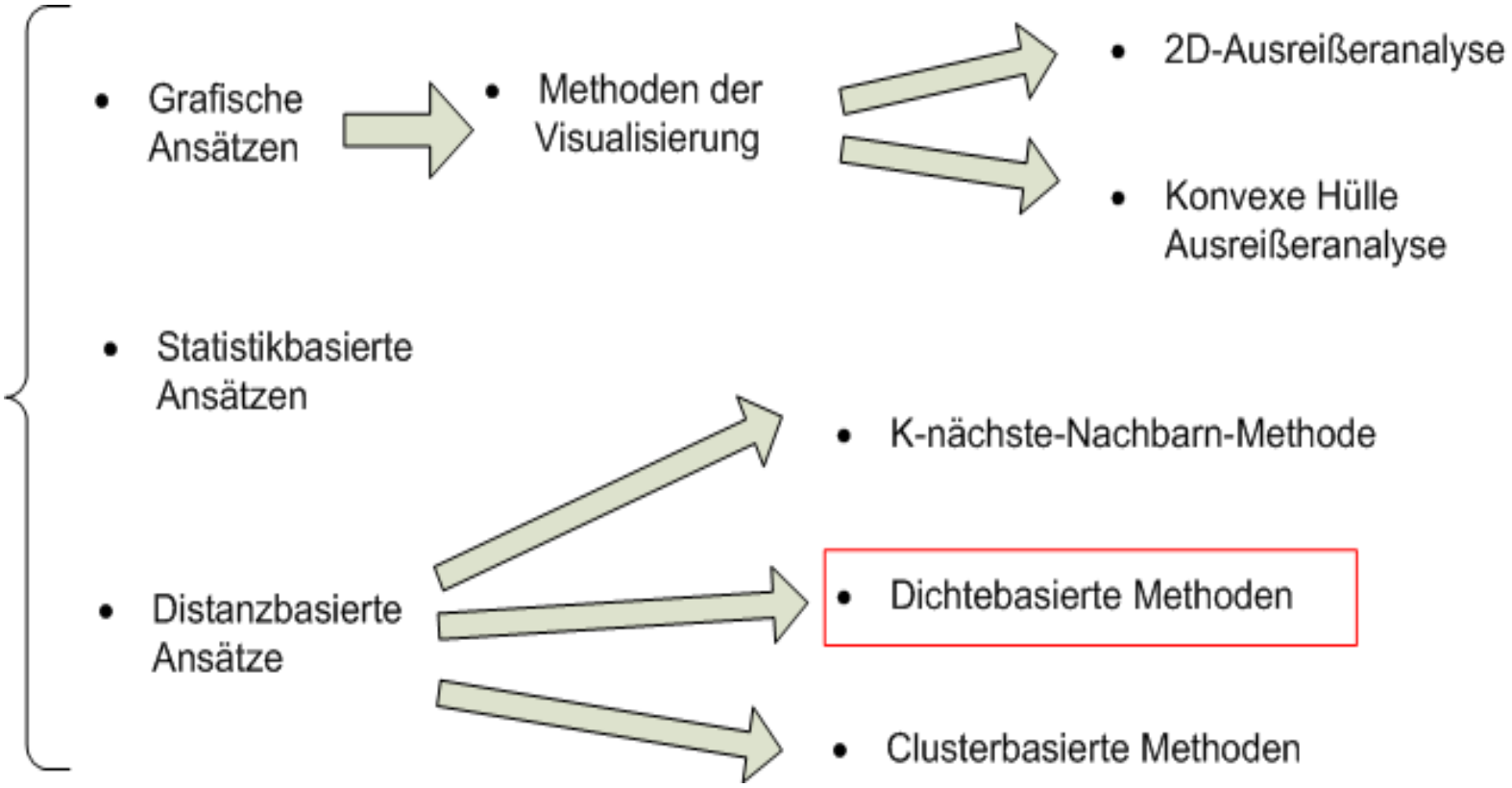
– (k,dmax)-Outlier: analog

– kNN-Outlier: kNN-Distanz der Objekte

in C1 größer als von o2



Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Dichtebasierte Methoden

- Berechnung für jedes Objekt die Dichte der lokalen Nachbarschaft.
- Beispielmenge p:

Bildung der Lokale Ausreißerfaktor als Durchschnitt des verhältnis:

-Der Dichte von p

- Dichte ihrer nächsten Nachbarn

- Objekte mit dem Größten Lokalen Ausreißerfaktor sind Ausreißer.

Dichtebasierte Ansätze

Lokale Identifikation von Ausreißer

- Nicht nur binäre Eigenschaften für Ausreißer
  d.h.(Ausreißer? Ja oder nein).
- Bei Clustern mit unterschiedlicher Dichte, können beim Distanzbasierte Ansätze Probleme Auftreten (DB(pct,dmin)).

- Dichte Basierte lokal Ausreißer:
  - weise jedem Objekt einen Grad zu, zu dem das Objekt ein Ausreißer ist.
  => Lokal Ausreißer Faktor LOF

  - Lokale Nachbarschaft von Objekten wird berücksichtig

Dichtebasierte Ausreißer

Local Outlier Factor (LOF) [Breunig, Kriegel, Ng, Sander 00]

• k-Distanz von p = dist(p,o), für jedes k, so dass gilt: (o Î DB)

(i) für mindestens k Objekte q Î DB gilt : dist(p,q) ≤ dist(p,o)

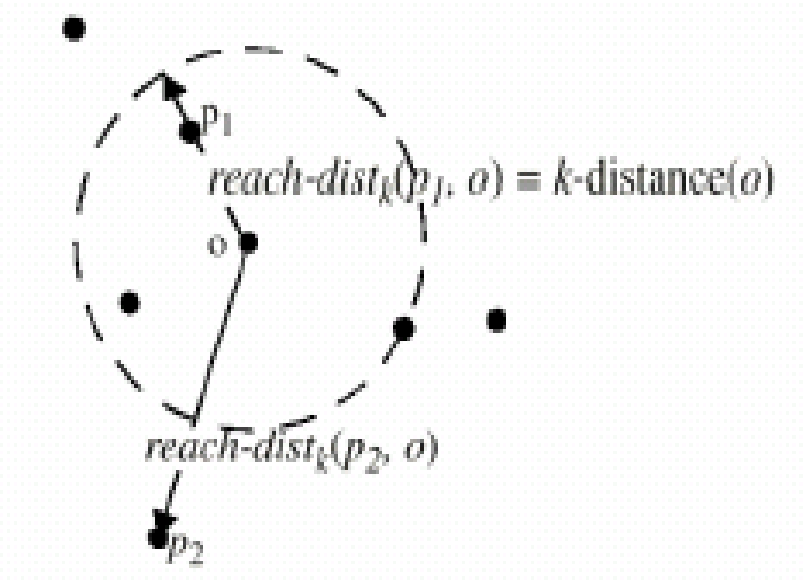
(ii) für höchstens k-1 Objekte q Î DB gilt : dist(p,q) < dist(p,o)

• k-Distanz - Nachbarschaft von p:

N k-distance(p) (p) = {q∈DB \ {p}|
dist(p,q) ≤ k-distance(p)}

• Erreichbarkeits-Distanz :

reach-distk(p,o) = max{k-distance(o),
dist(p,o)}



Dichtebasierte Ausreißer

Lokal Ausreißer Faktor(LOF)

- Als Parameter nur *MinPts*
- Lokale Erreichbarkeits-Distanz von *p:*

$$lrd_{MinPts}(p) = 1 / \left(\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p, o)}{|N_{MinPts}(p)|} \right)$$

- Lokal Ausreißer Faktor von *p (LOF):*

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

Dichtebasierte Ausreißer

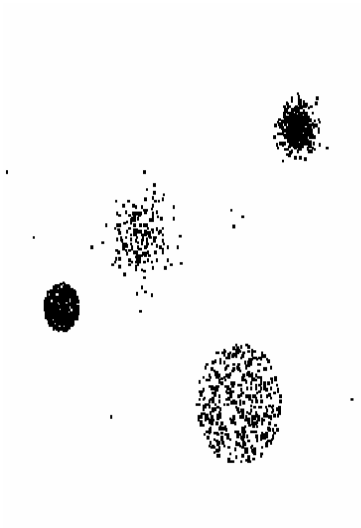
Lokal Ausreißer Faktor(LOF)

- LOF (*p) » 1:*

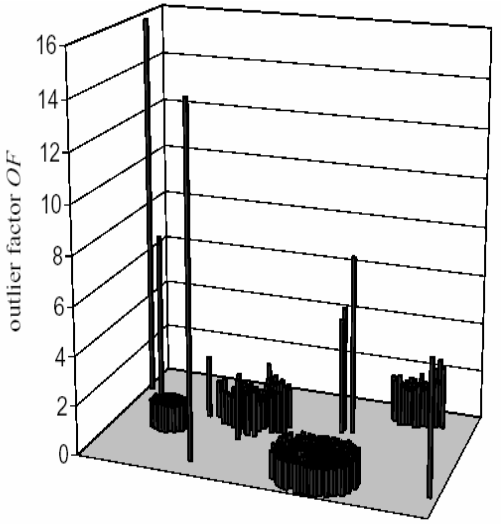
  Punkt liegt weit innen im Cluster

- LOF (*p) >> 1:*

  Punkt ist ein starker lokaler Ausreißer



Datensatz



LOFs (*MinPts = 40)*

Anwendungen

Szenario:

•Analisiere die Bewegung eines Objektes (z.B. des Menschen) von Punkt A zu Punkt B.

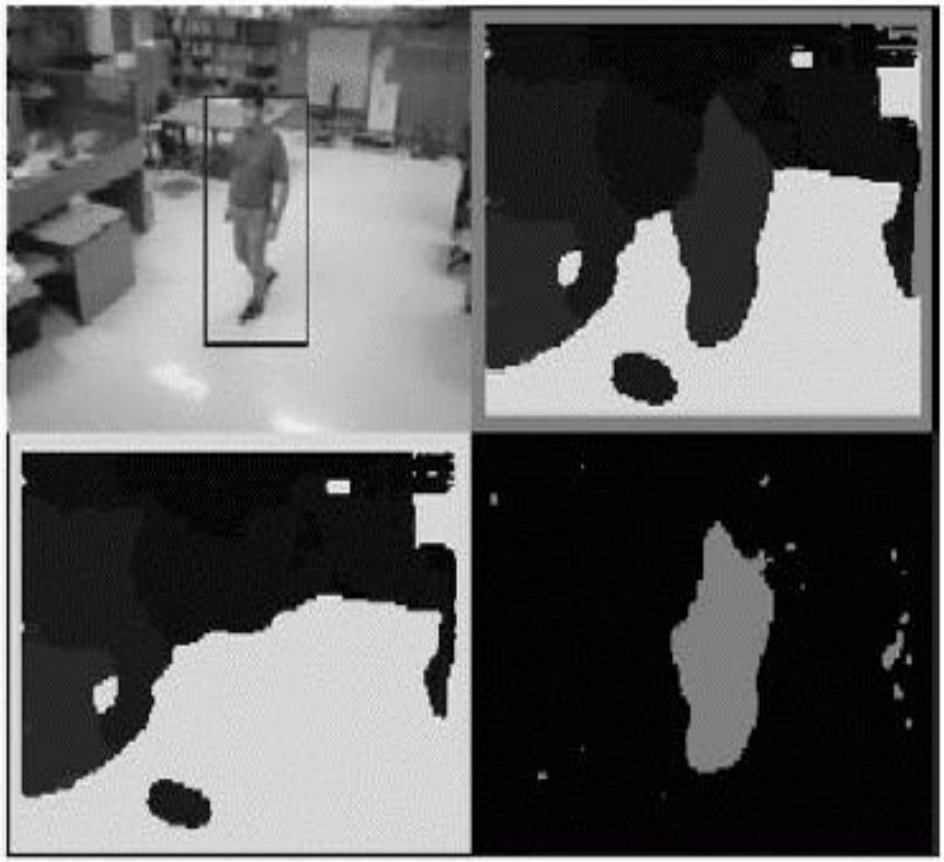
•Alarmiere Benutzer, falls sich das Objekt nicht an vorgeschriebene Routen hält.

•Andere Beispiele:

•Erdbeben früh erkennen (Quakefinder)

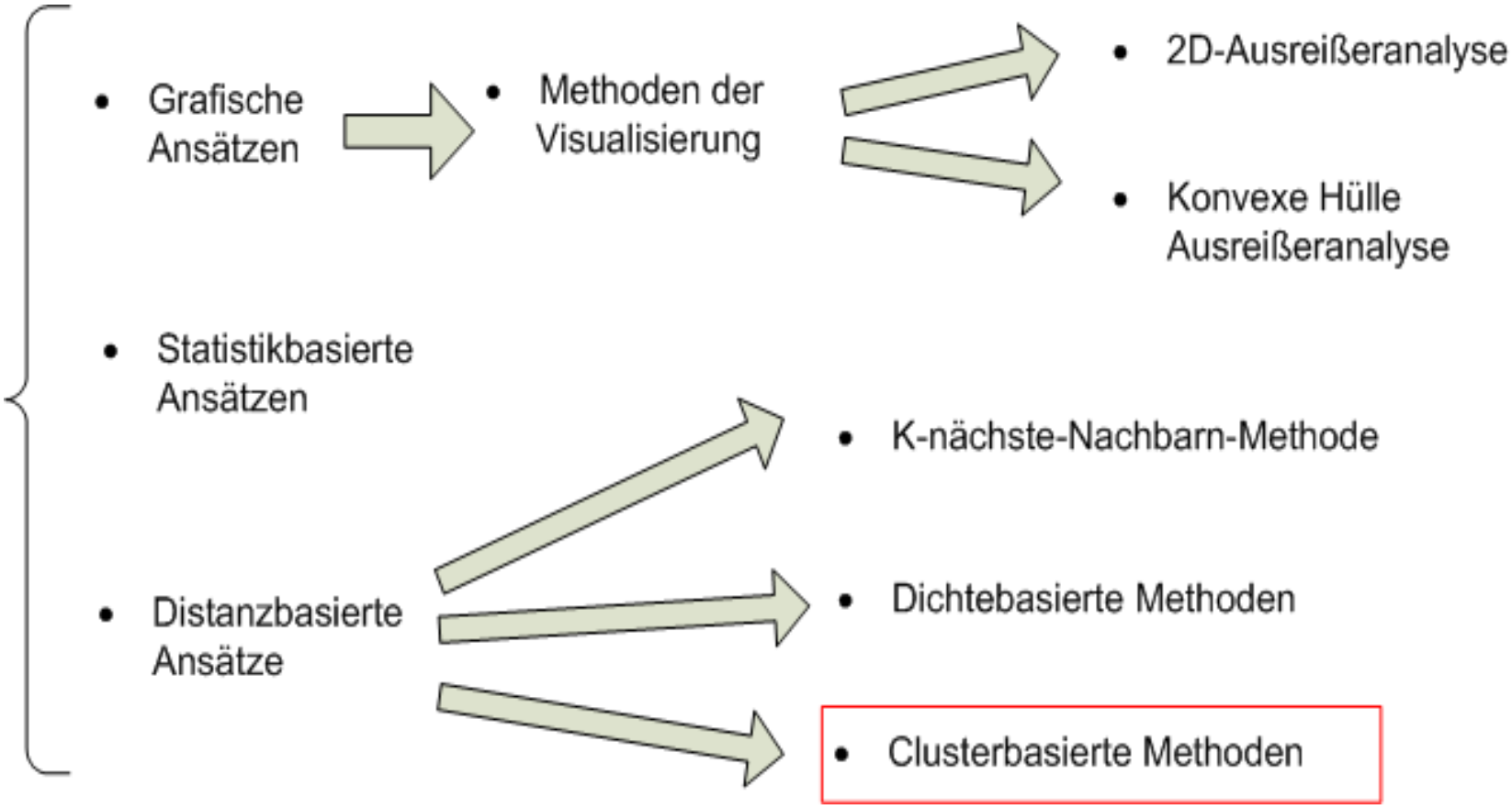
- durch Satellitenbeobachtungen.

. Vulkane aúf Venus erkennen (JARtool)



[Knorr, Ng, Tucakov 2000]: Triclops

Verfahren


Grafische Ansätzen
Methoden der Visualisierung
2D-Ausreißeranalyse
Konvexe Hülle Ausreißeranalyse
Statistikbasierte Ansätzen
Distanzbasierte Ansätze
K-nächste-Nachbarn-Methode
Dichtebasierte Methoden
Clusterbasierte Methoden

Clusterbasierte Methoden

- Zusammenfassung der Datenobjekten zunächst zu Clustern unterschiedlicher Dichte.
- Objekte in kleinen Clustern sind Kandidaten für Ausreißer.
- Ermittelung Distanz von:
  - Kandidaten Klassen
  - Nicht Kandidaten Klassen

Wenn:

Kandidaten > Nicht Kandidaten => Ausreißer

Zusammenfassung

- Distanzbasierte Ansätze: Basisansatz
- Dichtebasierte Ansatz: Für Datensätze mit unterschiedlich dichten Regionen.
- KNN erfordern keine Lernphase/Training (außer Evtl. zur Bestimmung von k). Bei sehr großen Datensuchen kann den suchen nach den K ähnlichsten Objekten sehr lange dauern. In diesem fall wählt man eine (repräsentative) Stichprobe aus dem Datensatz aus oder wendet zunächst eine Clusteranalise an.

Literaturverzeichnis

[ESSAN,2000] Martin Ester, Jörg Sander, Knowlwdge Discovery in Databases, Techniken und Anwendungen, Apringer Verlag Berlin Heidelberg, 2000.

[KUD, 2007] Thomas Kudrass, Taschen Buch Datenbanken, Carl Hanser Verlag München, 2007.

[GHO,98] Gholamreza Nakhaeizadeh, Data Mining, Theoretische Aspekte und AnwendungenPhysica Verlag, 1998.

[MENA, 2000] Jesus Mena, Data Mining und E-Commerce, Symposion Publisching GmbH, Düsseldorf, 2000.

[DEUT, 2006] Stephan Deutsch Diplomarbeit:Outlier Detection in USENET Newsgruppen, am Fachbereich Informatik der Universität Dortmund, Oktober 2006.

Vielen Dank für die Aufmerksamkeit

ENDE

K-Nächste-Nachbarn

Zusammenfassung

- KNN berechnet zur Klassifikation eines Objektes die K ähnlichsten Objekte aus dem Datensatz und Ordnet das gegebene Objekt in die Klasse, die am häufigsten unten den K ähnlichsten Objekten vorkommt.
- Die Ähnlichkeit bzw. die Unähnlichkeit zweier Objekte wird meistens über den euklidischen Abstand definiert. Dafür sollten die Attribute vorher normalisiert oder standarisiert werden.
- K wird üblicherweise durch Probieren ermittelt.

K-nächste-Nachbarn-Methode

Diskussion

- KNN eignen sich vor allem für Datensätze mit vorwiegend nummerischen Attributen.
- KNN erfordern keine Lernphase/Training (außer Evtl. zur Bestimmung von k). Bei sehr großen Datensuchen kann den suchen nach den K ähnlichsten Objekten sehr lange dauern. In diesem fall wählt man eine (repräsentative) Stichprobe aus dem Datensatz aus oder wendet zunächst eine Clusteranalise an.
- Bei Datensätzen mit sehr vielen Attributen ist die euklidische Distanz meistens wenig aussagekräftig, so dass NNK in diesem Fall meistens versagen.