DATABASE SYSTEMS GROUP

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

LMU

Skript zur Vorlesung

Knowledge Discovery in Databases

im Wintersemester 2010/2011

Kapitel 6: Outlier Detection

Vorlesung+Übungen:
PD Dr. Peer Kröger, Dr. Arthur Zimek

basiert auf Tutorial von Hans-Peter Kriegel, Peer Kröger, Arthur Zimek:
Outlier Detection Techniques (PAKDD-09, Bangkok, Thailand)

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

Übersicht

Was ist ein Outlier?

Definition nach Hawkins [Hawkins 1980]:

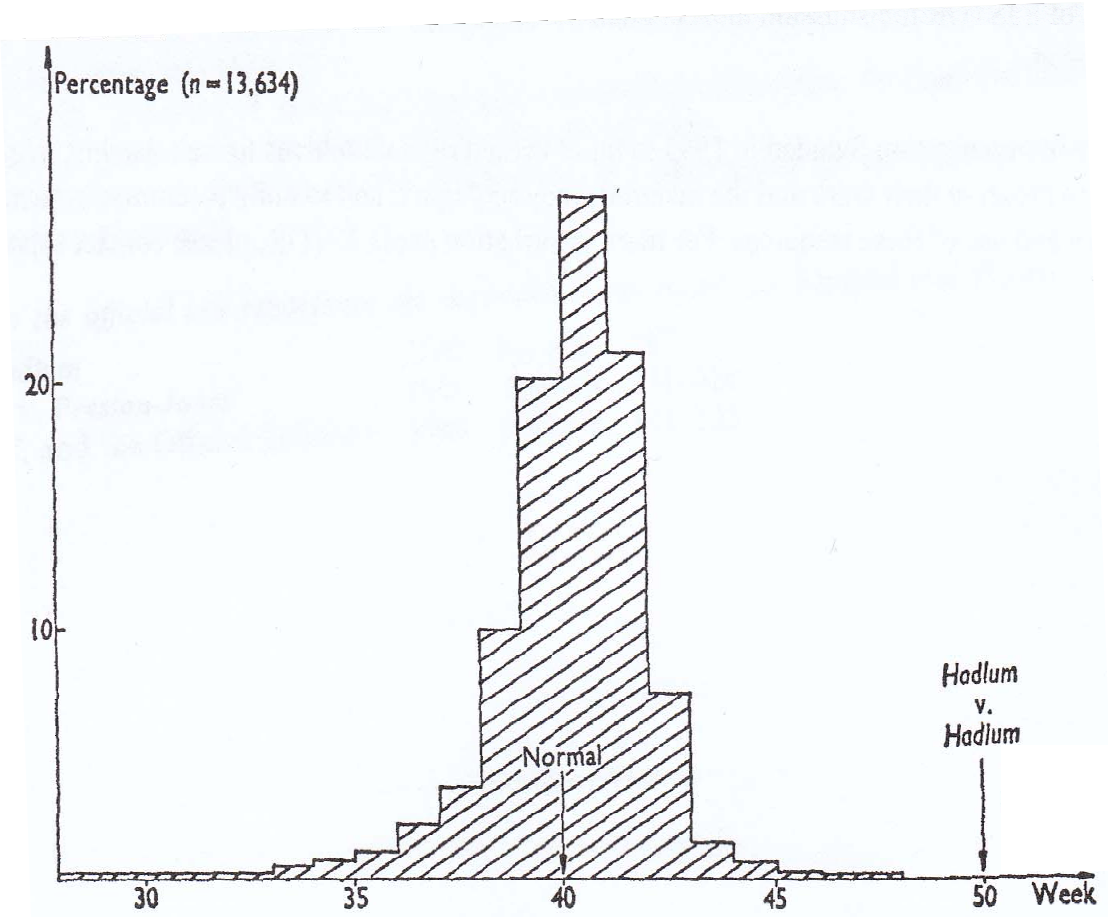
"Ein Outlier ist eine *Beobachtung*, die sich von den anderen *Beobachtungen* so deutlich unterscheidet, daß man denken könnte, sie sei von einem anderen Mechanismus generiert worden."

Was meint "Mechanismus"?

- Intuition aus der Statistik: "erzeugender Mechanismus" ist ein (statistischer) Prozess.
- Abnormale Daten (outlier) zeigen eine verdächtig geringe Wahrscheinlichkeit, aus diesem Prozess zu stammen.

Beispiel: Hadlum vs. Hadlum (1949) [Barnett 1978]

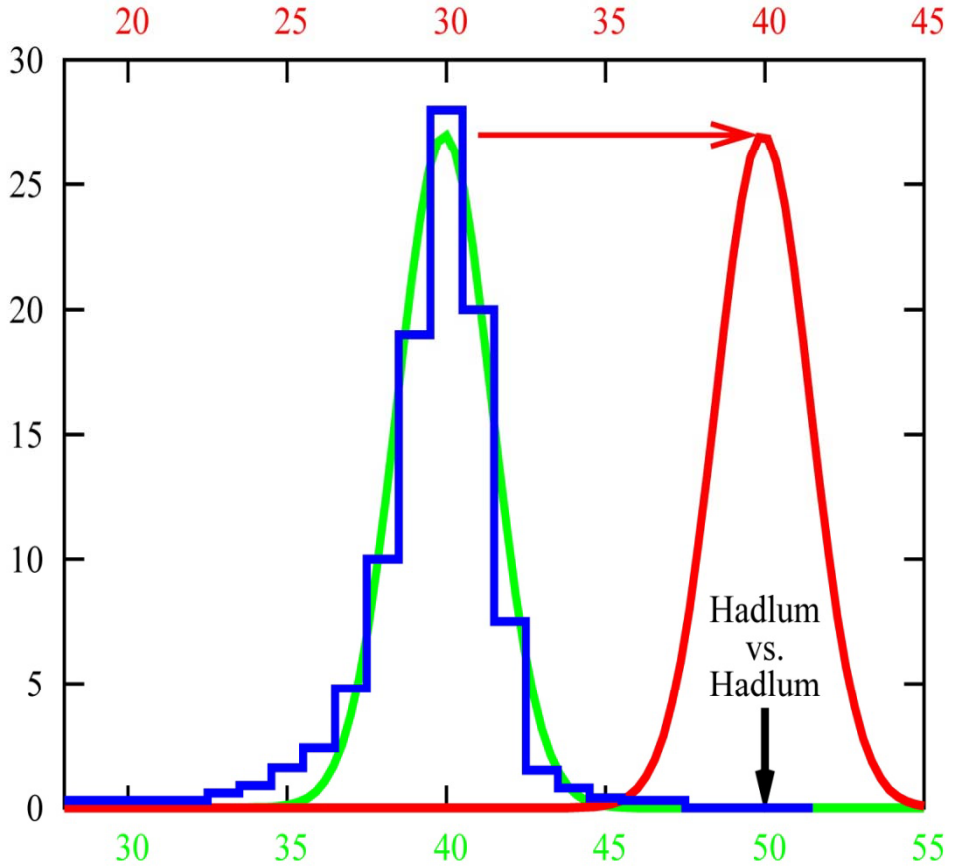
- Geburt eines Kindes von Mrs. Hadlum 349 Tage nachdem Mr. Hadlum zum Militärdienst abwesend war.
- Durchschnittliche Dauer einer menschlichen Schwangerschaft ist 280 Tage (40 Wochen)
- Ist eine Schwangerschaftsdauer von 349 Tagen ein Outlier?



Distribution of human gestation periods.

Beispiel: Hadlum vs. Hadlum (1949) [Barnett 1978]

- Blau: statistische Beobachtungsbasis (13634 erhobene Schwangerschaften)
- Grün: angenommener zugrundeliegender Gauss-Prozess
  - sehr geringe Wahrscheinlichkeit, dass die Geburt aus diesem Prozess stammt
- Rot: Annahme von Mr. Hadlum (ein anderer Gauss-Prozess, in dem die Schwangerschaft später beginnt, ist für die Geburt verantwortlich)
  - unter dieser Annahme hat die Schwangerschaftsdauer einen Durchschnittswert und höchst-mögliche Wahrscheinlichkeit



6.1 Einleitung

Anwendungsgebiete:

- Betrugsentdeckung
  - Kaufverhalten mit einer Kreditkarte ändert sich, wenn die Karte gestohlen wurde
  - Ungewöhnliche Kauf-Muster können Kreditkarten-Mißbrauch anzeigen
- Medizin
  - Ungewöhnliche Symptome oder Test-Ergebnisse können mögliche gesundheitliche Probleme eines Patienten anzeigen
  - Ob ein bestimmtes Testergebnis ungewöhnlich ist, kann von anderen Eigenschaften des Patienten abhängen (z.B. Geschlecht, Alter, Gewicht, ...)
- Öffentliches Gesundheitswesen
  - Auftauchen einer bestimmten Krankheit (z.B. Tetanus) verstreut über verschiedene Krankenhäuser einer Stadt zeigt Probleme mit dem zugehörigen Impfprogramm an
  - Ob das Auftreten der Krankheit unnormal ist hängt von verschiedenen Aspekten ab, z.B. Häufigkeit, räumliche Korrelation etc.

6.1 Einleitung

Anwendungsgebiete:

- Sport Statistiken
  - In vielen Sportarten werden diverse Parameter aufgezeichnet, um die Leistung eines Spielers zu bewerten
  - Außergewöhnliche (in positivem wie negativem Sinne) Spieler können durch ungewöhnliche Werte bestimmt werden
  - Manchmal ist nur eine Teilmenge der Parameter ungewöhnlich
- Entdecken von Messfehlern
  - Daten aus Sensoren (z.B. in einem wissenschaftlichen Experiment) können Meßfehler enthalten
  - Ungewöhnliche Werte können ein Hinweis auf Meßfehler sein
  - Solche Meßfehler aus den Daten zu entfernen, kann wichtig sein für erfolgreiche Datenanalyse und Data Mining

„One person's noise could be another person's signal."

6.1 Einleitung

Diskussion der Intuition von Hawkins

- Daten sind gewöhnlich multivariat (mehr-dimensional)

  => Basis-Modell ist univariat (ein-dimensional)
- Ein Datensatz stammt oft aus mehr als einem erzeugenden Prozess

  => Basis-Model nimmt nur einen einzelnen genuinen erzeugenden Mechanismus an
- Anomalien können eine andere Klasse von Objekten sein (aus einem anderen Prozess erzeugt), die nicht besonders selten sind

  => Basis-Model nimmt an, dass Outlier sehr selten sind

Eine große Zahl von Methoden wurde entwickelt, um über die Basis-Annahmen hinauszugelangen. Dabei liegen jedoch stets andere, oft nicht explizite Annahmen zugrunde.

6.1 Einleitung

Generelle Szenarien der Anwendung:

- supervised
  - in manchen Anwendungsgebieten gibt es Trainingsdaten mit normalen und ungewöhnlichen Fällen
  - es kann mehrere normale und ungewöhnliche Klassen geben
  - meist ist das Klassifikationsproblem unbalanziert
- semi-supervised
  - in manchen Szenarien gibt es Trainingsdaten nur für die normale oder nur für die ungewöhnliche Klasse
- unsupervised
  - in den meisten Szenarien gibt es keine Trainingsdaten

In dieser Vorlesung konzentrieren wir uns auf das unsupervised Szenario.

Erkennung von Outliern

- Nebenprodukt von Clustering?
- Manche Cluster-Algorithmen ordnen nicht jeden Punkt einem Cluster zu, sondern lassen “Noise” übrig.
- Idee: Wende Cluster-Verfahren an, betrachte Noise als Outlier.

- Problem:
  - Clustering Algorithmen sind daraufhin entwickelt und optimiert, Cluster zu finden.
  - Qualität der Outlier Detection hängt von Qualität der Cluster-Struktur und der Eignung des Clustering Algorithmus für diese Struktur ab.
  - Mehrere Outlier, die einander ähnlich sind, bilden eventuell auch selbst ein (kleines) Cluster, können also nicht entdeckt werden.

Klassifikation von Outlier Detection Algorithmen

- Globaler vs. lokaler Ansatz:

  Wird die "Outlierness" bestimmt bezüglich des gesamten Datensatzes (global) oder nur bezüglich einer Auswahl?

- Labeling vs. Scoring

  Bestimmt der Algorithmus den Outlier-Grad eines Punktes (Scoring) oder wird für jeden Punkt eine Entscheidung getroffen (Label: Outlier/kein Outlier)

- Eigenschaften des Outlier Modells

  Auf welchen Eigenschaften beruht die Modellierung von "Outlierness"

- Global vs. Lokal
  - bezieht sich auf die Auflösung der Referenzmenge bezüglich derer die "Outlierness" bestimmt wird
  - Globale Ansätze:
    - Referenzmenge enthält gesamten Datensatz
    - Basis-Annahme: nur ein einziger (normaler) erzeugender Mechanismus
    - Grundlegendes Problem: Outlier sind auch in Referenzmenge und verfälschen die Ergebnisse
  - Lokale Ansätze:
    - Referenzmenge enthält nur eine (kleine) Teilmenge des Datensatzes
    - Meist keine Annahme über Anzahl der Mechanismen
    - Grundlegendes Problem: wie ist eine geeignete Referenzmenge zu bestimmen?
  - Beachte: Manche Ansätze liegen dazwischen
    - Auflösung der Referenzmenge wird im Verfahren variiert

6.1 Einleitung

- Labeling vs. Scoring
  - bezieht sich auf das Ergebnis, das der Algorithmus liefert
  - Labeling Ansätze:
    - binäre Entscheidung
    - Daten-Objekt wird als Outlier markiert oder als normal
  - Scoring Ansätze:
    - kontinuierlicher Output: für jedes Objekt wird ein Score geliefert (z.B. die Wahrscheinlichkeit, ein Outlier zu sein)
    - Objekte können nach ihrem Score geordnet werden
  - Beachte:
    - Viele Scoring-Ansätze bestimmen nur die top-n Outlier (Parameter n wird durch Benutzer angegeben)
    - Scoring-Ansätze können grundsätzlich in Labeling-Ansätze transformiert werden, wenn ein geeigneter Grenzwert angegeben werden kann, dessen Überschreitung zum Label “Outlier” führt

6.1 Einleitung

- Klassen von zugrundeliegenden Modellen
  - Statistisches Modell
    - Überlegung:
      - Wende ein Modell an, das die normalen Daten statistisch beschreibt (z.B. Gauss-Verteilung)
      - Outlier sind Punkte, die nicht gut zu diesem Modell passen (eine geringe Erzeugungswahrscheinlichkeit haben)
    - Beispiele:
      - Wahrscheinlichkeitstests basierend auf statistischen Modellen
      - Tiefen-basierte Ansätze
      - Deviation-based Ansätze
      - Manche Subspace Outlier Detection Ansätze

6.1 Einleitung

- Modellierung durch räumliche Nähe
  - Überlegung:
    - Untersuche die räumliche Nachbarschaft jedes Punktes im Datenraum
    - Wenn die Nachbarschaft deutlich andere Struktur (z.B. geringere Dichte) aufweist als die Nachbarschaften von anderen Punkten, kann der betreffende Punkt als Outlier angesehen werden.
  - Beispiele:
    - Distanz-basierte Ansätze
    - Dichte-basierte Ansätze
    - Manche Subspace Outlier Detection Ansätze

6.1 Einleitung

- Modellierung durch Winkel-Spektrum
  - Überlegung:
    - Bestimme das Spektrum paarweiser Winkel zwischen einem gegebenen Punkt und anderen (alle? Auswahl?) Punkten
    - Outlier sind Punkte, die eine geringe Varianz haben

Im Folgenden:

Orientierung an den verschiedenen Modellierungen

Übersicht

6.1 Einleitung √

6.2 Statistical Tests

6.3 Depth-based Approaches

6.4 Deviation-based Approaches

Statistisches Modell

6.5 Distance-based Approaches

6.6 Density-based Approaches

Modellierung durch räumliche Nähe

6.7 High-dimensional Approaches

Anpassung verschiedener Modelle an spezielles Problem

6.8 Summary

Literatur

6.2 Statistical Tests

General idea

- Given a certain kind of statistical distribution (e.g., Gaussian)
- Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)
- Outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean)

Basic assumption

- Normal data objects follow a (known) distribution and occur in a high probability region of this model
- Outliers deviate strongly from this distribution

6.2 Statistical Tests

A huge number of different tests are available differing in

- Type of data distribution (e.g. Gaussian)
- Number of variables, i.e., dimensions of the data objects (univariate/multivariate)
- Number of distributions (mixture models)
- Parametric versus non-parametric (e.g. histogram-based)

Example on the following slides

- Gaussian distribution
- Multivariate
- 1 model
- Parametric

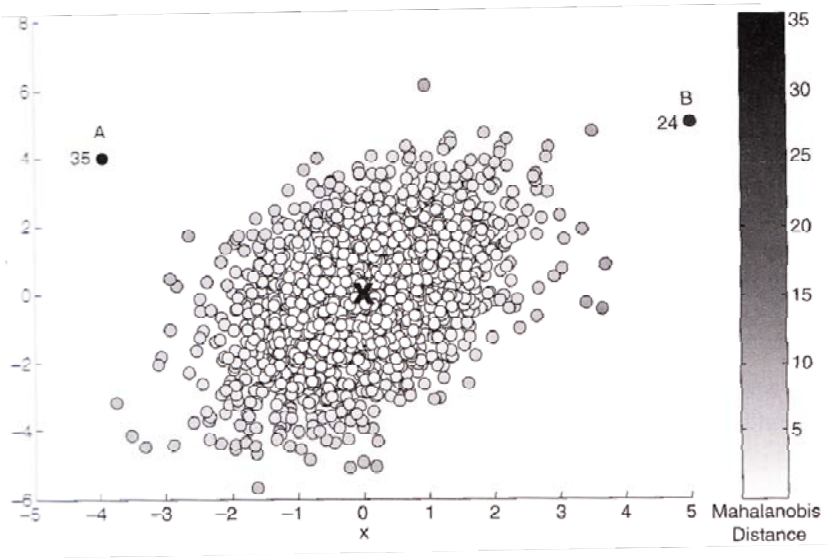
6.2 Statistical Tests

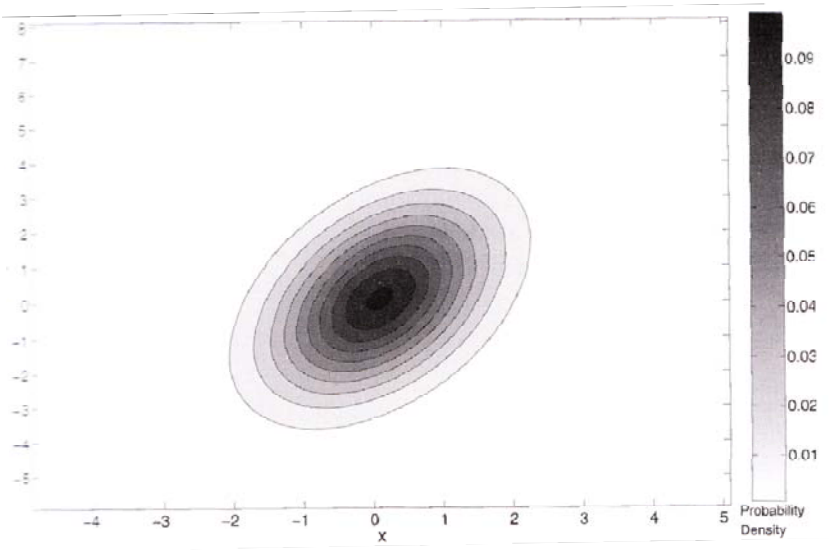
Probability density function of a multivariate normal distribution

$$N(x) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} e^{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}$$

- μ is the mean value of all points (usually data are normalized such that $\mu=0$)
- Σ is the covariance matrix from the mean
- $MDist(x,\mu) = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is the Mahalanobis distance of point x to μ
- MDist follows a χ^2-distribution with d degrees of freedom (d = data dimensionality)
- All points x, with $MDist(x,\mu) > \chi^2(0{,}975)$ $[\approx 3\cdot\sigma]$

Visualization (2D) [Tan et al. 2006]

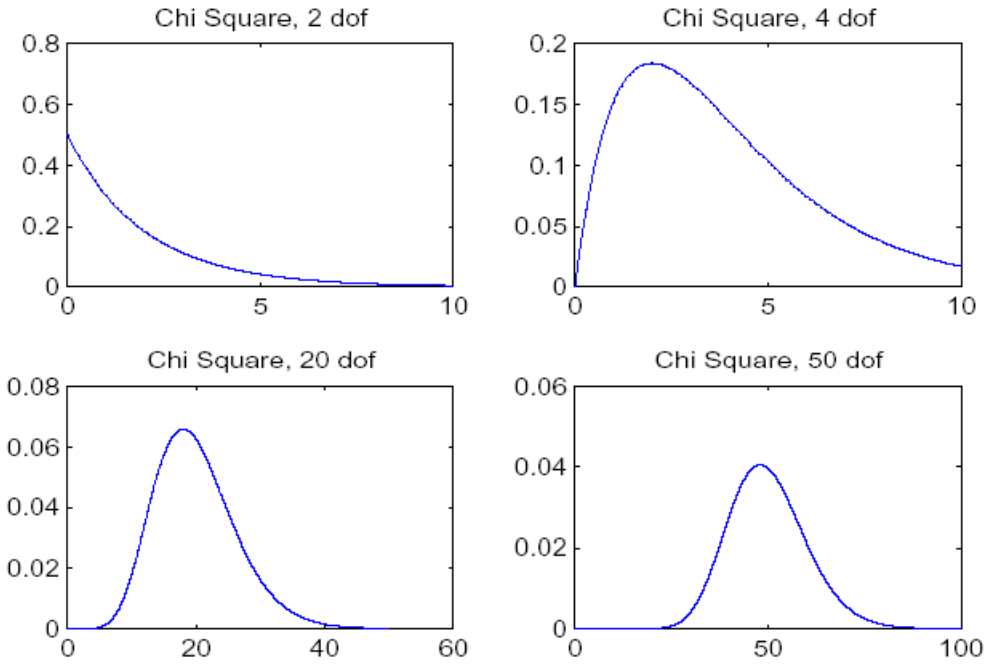




6.2 Statistical Tests

Problems

- Curse of dimensionality
  - The larger the degree of freedom, the more similar the MDist values for all points



x-axis: observed *MDist* values

y-axis: frequency of observation

6.2 Statistical Tests

Problems (cont.)

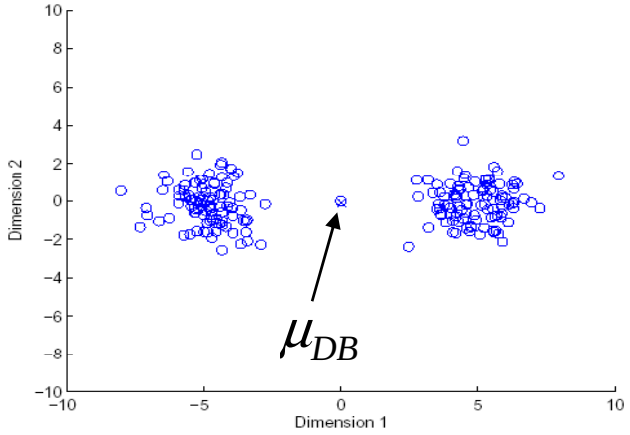
- Robustness
  - Mean and standard deviation are very sensitive to outliers
  - These values are computed for the complete data set (including potential outliers)
  - The *MDist* is used to determine outliers although the *MDist* values are influenced by these outliers

    => Minimum Covariance Determinant [Rousseeuw and Leroy 1987]

    minimizes the influence of outliers on the Mahalanobis distance

Discussion

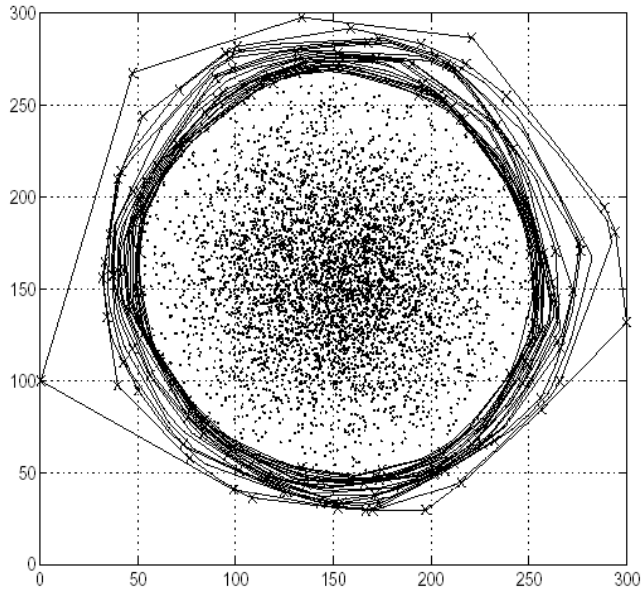
- Data distribution is fixed
- Low flexibility (no mixture model)
- Global method
- Outputs a label but can also output a score



6.3 Depth-based Approaches

General idea

- Search for outliers at the border of the data space but independent of statistical distributions
- Organize data objects in convex hull layers
- Outliers are objects on outer layers



Picture taken from [Johnson et al. 1998]

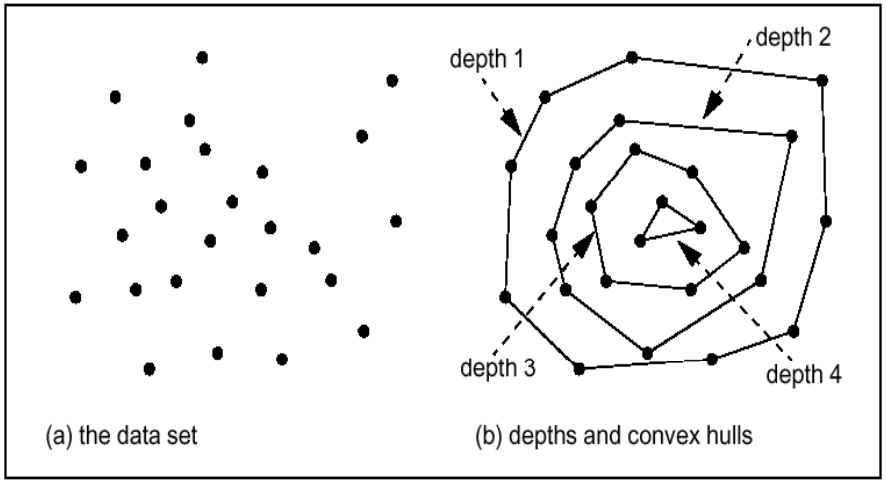
Basic assumption

- Outliers are located at the border of the data space
- Normal objects are in the center of the data space

6.3 Depth-based Approaches

Model [Tukey 1977]

- Points on the convex hull of the full data space have depth = 1
- Points on the convex hull of the data set after removing all points with depth = 1 have depth = 2
- …
- Points having a depth ≤ *k* are reported as outliers



Picture taken from [Preparata and Shamos 1988]

6.3 Depth-based Approaches

Sample algorithms

- ISODEPTH [Ruts and Rousseeuw 1996]
- FDC [Johnson et al. 1998]

Discussion

- Similar idea like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution
- Convex hull computation is usually only efficient in 2D / 3D spaces
- Originally outputs a label but can be extended for scoring easily (take depth as scoring value)
- Uses a global reference set for outlier detection

6.4 Deviation-based Approaches

General idea

- Given a set of data points (local group or global set)
- Outliers are points that do not fit to the general characteristics of that set, i.e., the variance of the set is minimized when removing the outliers

Basic assumption

- Outliers are the outermost points of the data set

6.4 Deviation-based Approaches

Model [Arning et al. 1996]

- Given a smoothing factor SF(I) that computes for each $I \subseteq DB$ how much the variance of DB is decreased when I is removed from DB
- With equal decrease in variance, a smaller exception set is better
- The outliers are the elements of the **exception set** $E \subseteq DB$ for which the following holds:

$$SF(E) \geq SF(I) \quad \text{for all } I \subseteq DB$$

Discussion:

- Similar idea like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution
- Naïve solution is in $O(2^n)$ for n data objects
- Heuristics like random sampling or best first search are applied
- Applicable to any data type (depends on the definition of SF)
- Originally designed as a global method
- Outputs a labeling

Übersicht

6.1 Einleitung √
6.2 Statistical Tests √
6.3 Depth-based Approaches √
6.4 Deviation-based Approaches √

Statistisches Modell

6.5 Distance-based Approaches
6.6 Density-based Approaches

Modellierung durch räumliche Nähe

6.7 High-dimensional Approaches

Anpassung verschiedener Modelle an spezielles Problem

6.8 Summary
Literatur

6.5 Distance-based Approaches

General Idea

- Judge a point based on the distance(s) to its neighbors
- Several variants proposed

Basic Assumption

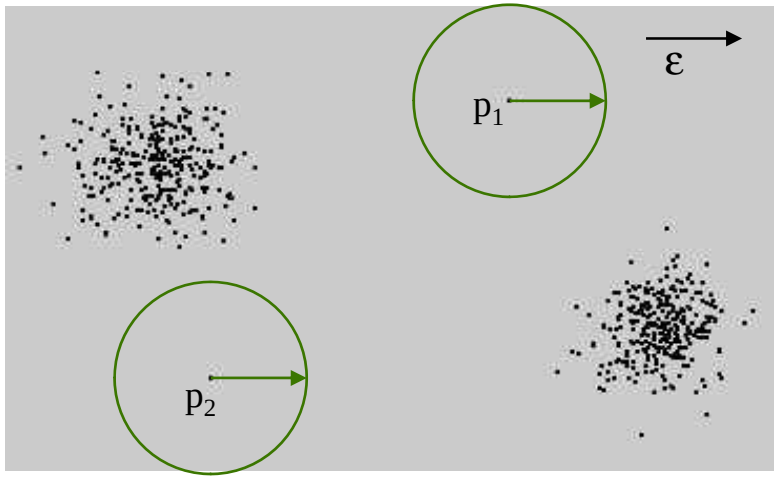
- Normal data objects have a dense neighborhood
- Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

DB(ε,π)-Outliers

- Basic model [Knorr and Ng 1997]
  - Given a radius ε and a percentage π
  - A point *p* is considered an outlier if at most π percent of all other points have a distance to *p* less than ε

$$OutlierSet(\varepsilon,\pi) = \{p \mid \frac{Card(\{q \in DB \mid dist(p,q) < \varepsilon\})}{Card(DB)} \leq \pi\}$$

range-query with radius ε



6.5 Distance-based Approaches

- Algorithms
  - Index-based [Knorr and Ng 1998]
    - Compute distance range join using spatial index structure
    - Exclude point from further consideration if its ε-neighborhood contains more than $Card(DB) \cdot \pi$ points
  - Nested-loop based [Knorr and Ng 1998]
    - Divide buffer in two parts
    - Use second part to scan/compare all points with the points from the first part
  - Grid-based [Knorr and Ng 1998]
    - Build grid such that any two points from the same grid cell have a distance of at most ε to each other
    - Points need only compared with points from neighboring cells

6.5 Distance-based Approaches

- Deriving intensional knowledge [Knorr and Ng 1999]
  - Relies on the DB(ε,π)-outlier model
  - Find the minimal subset(s) of attributes that explains the "outlierness" of a point, i.e., in which the point is still an outlier
  - Example
    - Identified outliers

Player Name	Power-play Goals	Short-handed Goals	Game-winning Goals	Game-tying Goals	Games Played
MARIO LEMIEUX	31	8	8	0	70
JAROMIR JAGR	20	1	12	1	82
JOHN LECLAIR	19	0	10	2	82
ROD BRIND'AMOUR	4	4	5	4	82

    - Derived intensional knowledge (sketch)

```
MARIO LEMIEUX:
  (i)  An outlier in the 1-D space of Power-play goals
  (ii) An outlier in the 2-D space of Short-handed goals and
       Game-winning goals
       (No player is exceptional on Short-handed goals alone;
        No player is exceptional on Game-winning goals alone.)
ROD BRIND'AMOUR:
  (i)  An outlier in the 1-D space of Game-tying goals
JAROMIR JAGR:
  (i)  An outlier in the 2-D space of Short-handed goals and
       Game-winning goals
       (No player is exceptional on Short-handed goals alone;
        No player is exceptional on Game-winning goals alone.)
  (ii) An outlier in the 2-D space of Power-play goals and
       Game-winning goals
```

6.5 Distance-based Approaches

Outlier scoring based on *k*NN distances

- General models
  - Take the *k*NN distance of a point as its outlier score [Ramaswamy et al 2000]
  - Aggregate the distances of a point to all its 1NN, 2NN, …, *k*NN as an outlier score [Angiulli and Pizzuti 2002]
- Algorithms
  - General approaches
    - Nested-Loop
      - » Naïve approach:
        For each object: compute *k*NNs with a sequential scan
      - » Enhancement: use index structures for *k*NN queries
    - Partition-based
      - » Partition data into micro clusters
      - » Aggregate information for each partition (e.g. minimum bounding rectangles)
      - » Allows to prune micro clusters that cannot qualify when searching for the *k*NNs of a particular point

6.5 Distance-based Approaches

- Sample Algorithms (computing top-*n* outliers)
  - Nested-Loop [Ramaswamy et al 2000]
    - Simple NL algorithm with index support for *k*NN queries
    - Partition-based algorithm (based on a clustering algorithm that has linear time complexity)
    - Algorithm for the simple *k*NN-distance model
  - Linearization [Angiulli and Pizzuti 2002]
    - Linearization of a multi-dimensional data set using space-fill curves
    - 1D representation is partitioned into micro clusters
    - Algorithm for the average *k*NN-distance model
  - ORCA [Bay and Schwabacher 2003]
    - NL algorithm with randomization and simple pruning
    - Pruning: if a point has a score greater than the top-*n* outlier so far (cut-off), remove this point from further consideration

      => non-outliers are pruned

      => works good on randomized data (can be done in linear time)

      => worst-case: naïve NL algorithm
    - Algorithm for both *k*NN-distance models and the DB(ε,π)-outlier model

6.5 Distance-based Approaches

- Sample Algorithms (cont.)
  - RBRP [Ghoting et al. 2006],
    - Idea: try to increase the cut-off as quick as possible => increase the pruning power
    - Compute approximate *k*NNs for each point to get a better cut-off
    - For approximate *k*NN search, the data points are partitioned into micro clusters and *k*NNs are only searched within each micro cluster
    - Algorithm for both *k*NN-distance models
  - Further approaches
    - Also apply partitioning-based algorithms using micro clusters [McCallum et al 2000], [Tao et al. 2006]
    - Approximate solution based on reference points [Pei et al. 2006]
- Discussion
  - Output can be a scoring (*k*NN-distance models) or a labeling (*k*NN-distance models and the DB(ε,π)-outlier model)
  - Approaches are local (resolution can be adjusted by the user via ε or *k*)

Variant

- Outlier Detection using In-degree Number [Hautamaki et al. 2004]
  - Idea
    - Construct the *k*NN graph for a data set
      - » Vertices: data points
      - » Edge: if $q \in k\text{NN}(p)$ then there is a directed edge from p to q
    - A vertex that has an indegree less or equal T (user defined threshold) is an outlier
  - Discussion
    - The indegree of a vertex in the *k*NN graph equals to the number of reverse kNNs (R*k*NN) of the corresponding point
    - The R*k*NNs of a point p are those data objects having p among their *k*NNs
    - Intuition of the model: outliers are
      - » points that are among the *k*NNs of less than T other points
      - » have less than T R*k*NNs
    - Outputs an outlier label
    - Is a local approach (depending on user defined parameter *k*)

Resolution-based outlier factor (ROF) [Fan et al. 2006]

- Model
  - Depending on the resolution of applied distance thresholds, points are outliers or within a cluster
  - With the maximal resolution *Rmax* (minimal distance threshold) all points are outliers
  - With the minimal resolution *Rmin* (maximal distance threshold) all points are within a cluster
  - Change resolution from *Rmax* to *Rmin* in certain steps: points change from being outlier to being a member of a cluster
  - Cluster is defined similar as in DBSCAN as a transitive closure of *r*-neighborhoods (where *r* is the current resolution)
  - ROF value

$$ROF(p) = \sum_{R\min \le r \le R\max} \frac{clusterSize_{r-1}(p) - 1}{clusterSize_r(p)}$$

- Discussion
  - Outputs a score (the ROF value)
  - Resolution is varied automatically from local to global