Skript zur Vorlesung

Knowledge Discovery in Databases

im Sommersemester 2015

Kapitel 4: Outlier Detection

Vorlesung: PD Dr. Arthur Zimek
Übungen: Dr. Tobias Emrich

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)

Was ist ein Outlier?

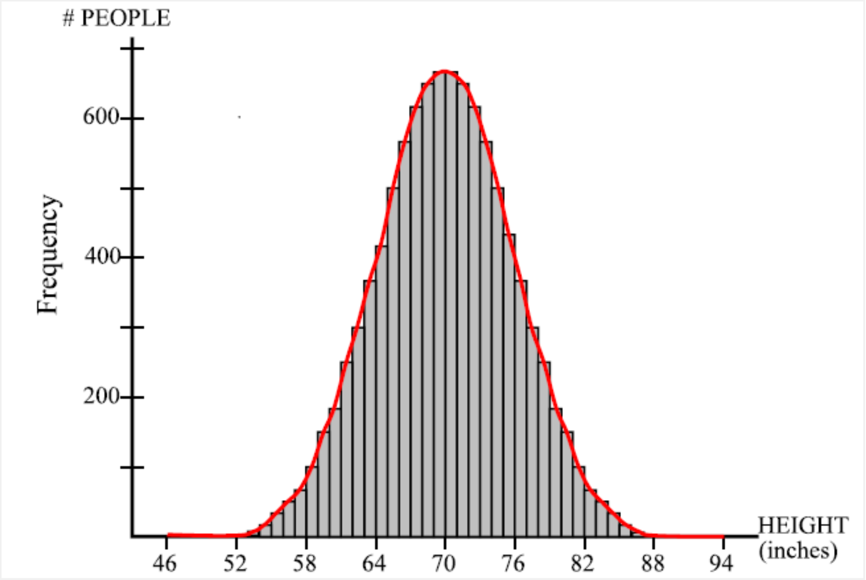
Definition nach Hawkins [Hawkins 1980]:

"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"?

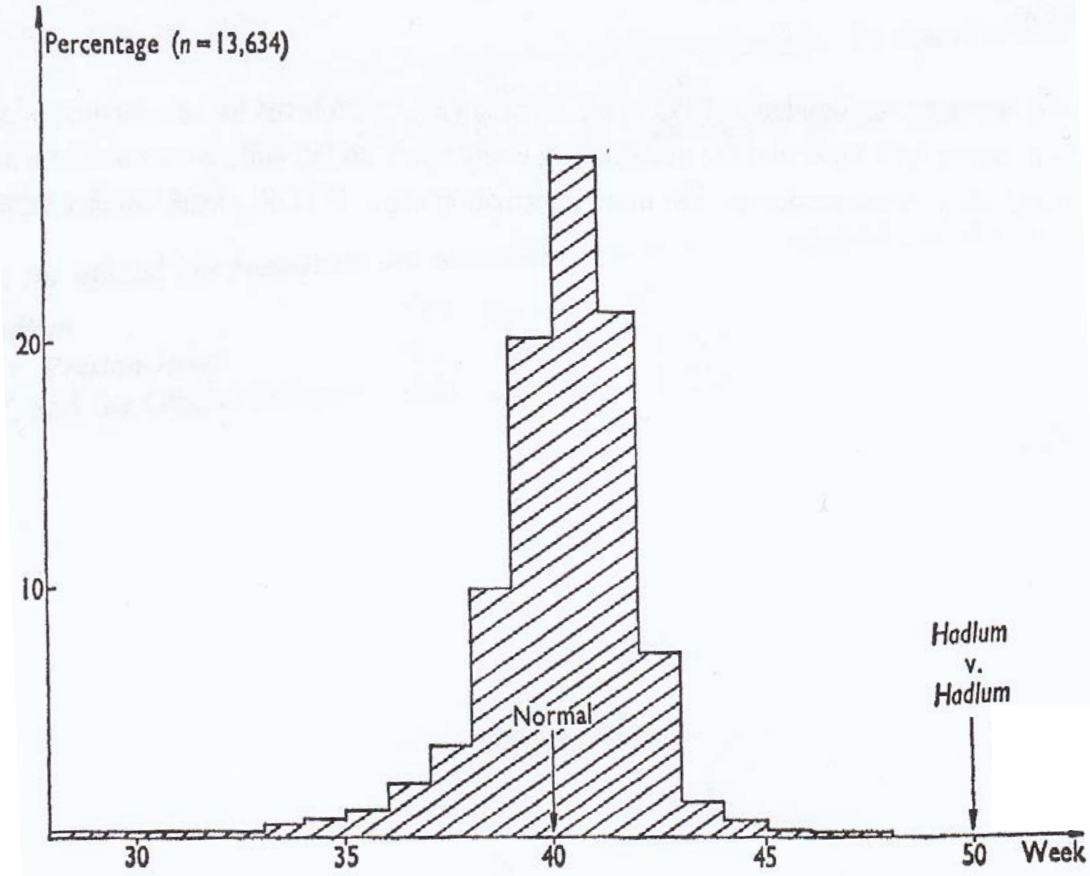
Intuition aus der Statistik: "erzeugender Mechanismus" ist ein (statistischer) Prozess.

Abnormale Daten (outlier) zeigen eine verdächtig geringe Wahrscheinlichkeit, aus diesem Prozess zu stammen.



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

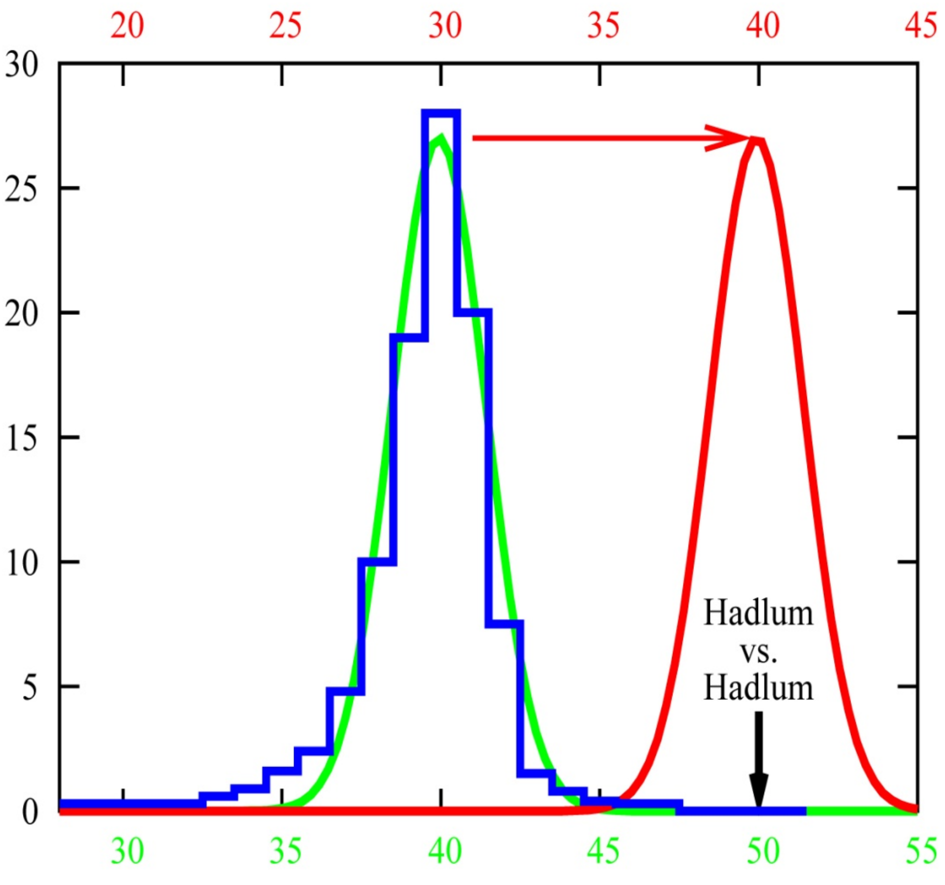
- Geburt eines Kindes von Mrs. Hadlum 349 Tage nachdem Mr. Hadlum zum Militärdienst abwesend war.
- Durchschnittliche Dauer einer menschlichen Schwangerschaft ist 280 Tage (40 Wochen)
- Ist eine Schwangerschaftsdauer von 349 Tagen ein Outlier?



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



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.

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."

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.

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 unbalanciert
- 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.

Überblick

- Einleitung
- Statistische Modellierung
- Depth-based Outliers
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Zusammenfassung

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

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

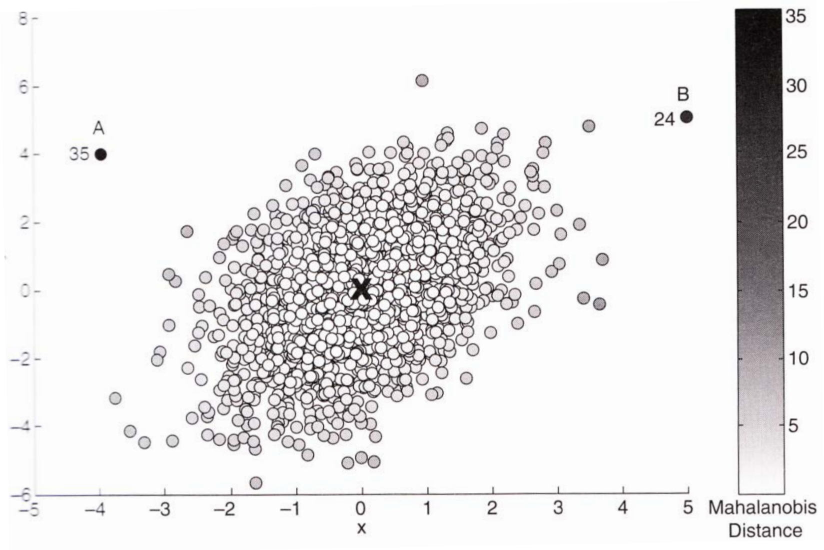
- Gaussian distribution
- Multivariate
- 1 model
- Parametric

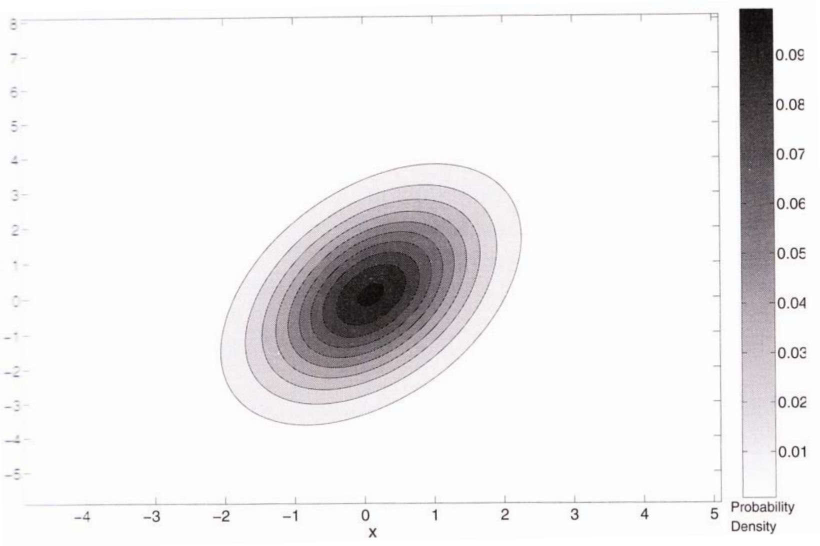
Probability density function of a multivariate normal distribution

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

- μ is the mean value of all points (usually data are normalized such that μ=0)
- Σ is the covariance matrix from the mean
- $MDist(x,\mu) = (x-\mu)^{\mathbf{T}} \mathbf{\Sigma}^{-1} (x-\mu)$ 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 $MDist(x,\mu) > \chi^2(0{,}975)$ [$\approx 3 \cdot \sigma$]

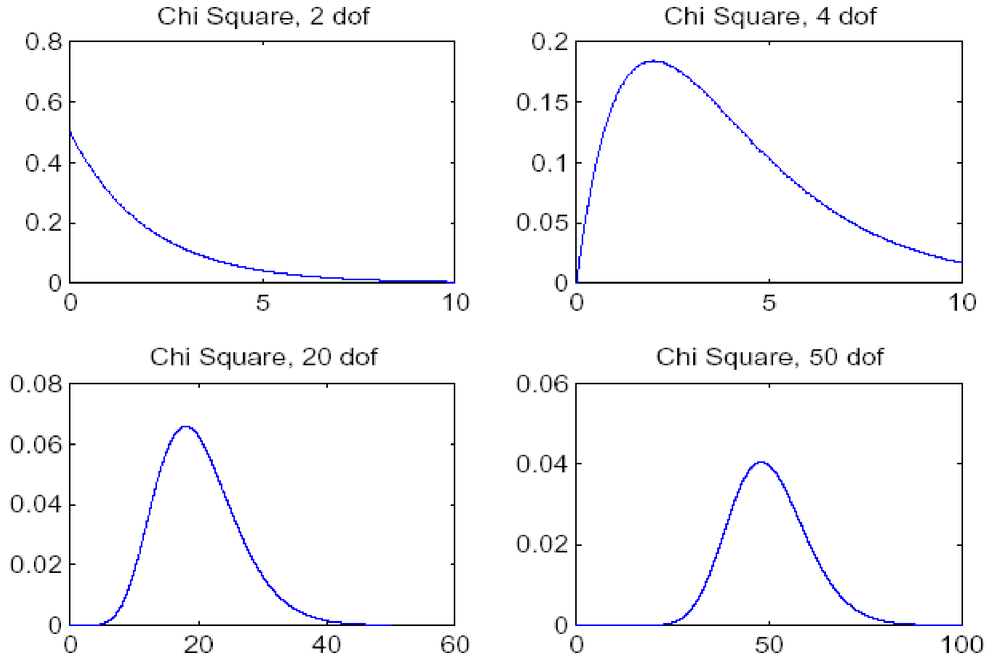
Visualization (2D) [Tan et al. 2006]





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

Problems (cont.)

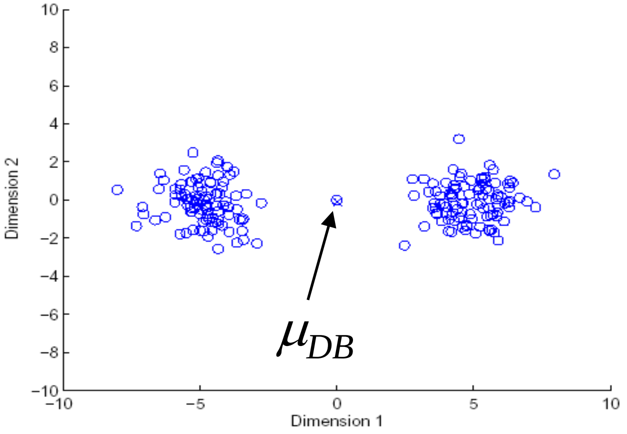
- Robustness
  - Mean and standard deviation are very sensitive to outliers
  - These values are computed for the complete data set (including potential outliers)
  - The *MDist* is used to determine outliers although the *MDist* values are influenced by these outliers

    => 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



DATABASE SYSTEMS GROUP

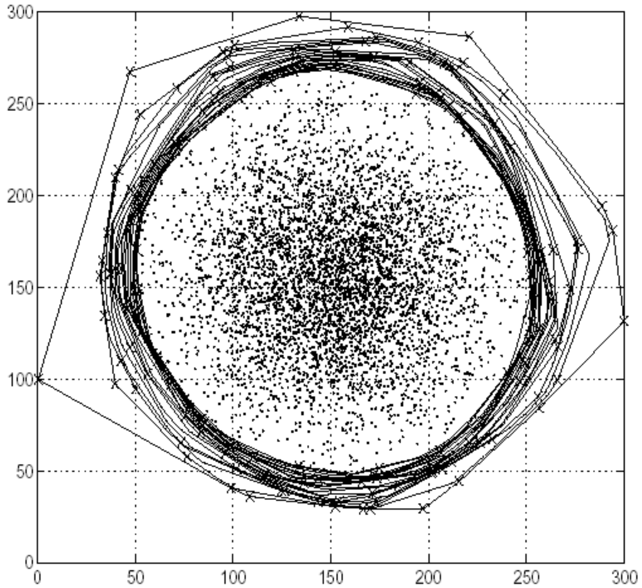
Überblick

- Einleitung
- Statistische Modellierung
- Depth-based Outliers
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Zusammenfassung

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]

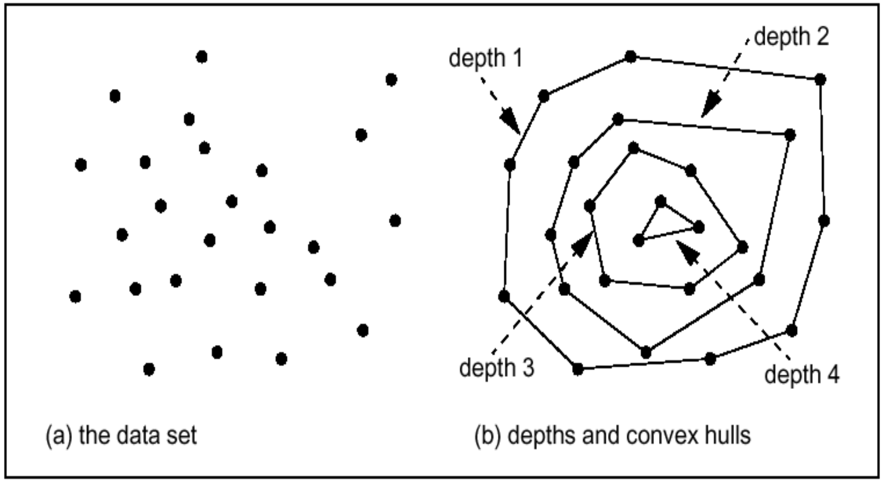
Basic assumption

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

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]

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

Überblick

- Einleitung
- Statistische Modellierung
- Depth-based Outliers
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Zusammenfassung

Distance-based Approaches

General Idea

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

Basic Assumption

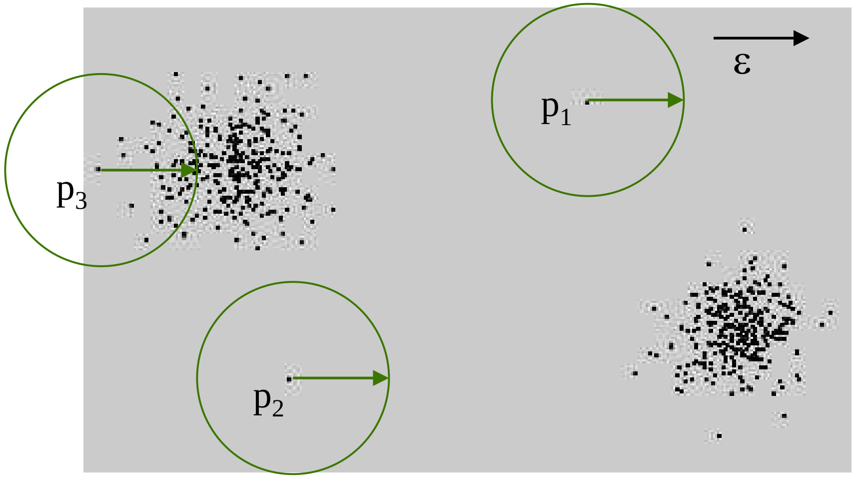
- Normal data objects have a dense neighborhood
- Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

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

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

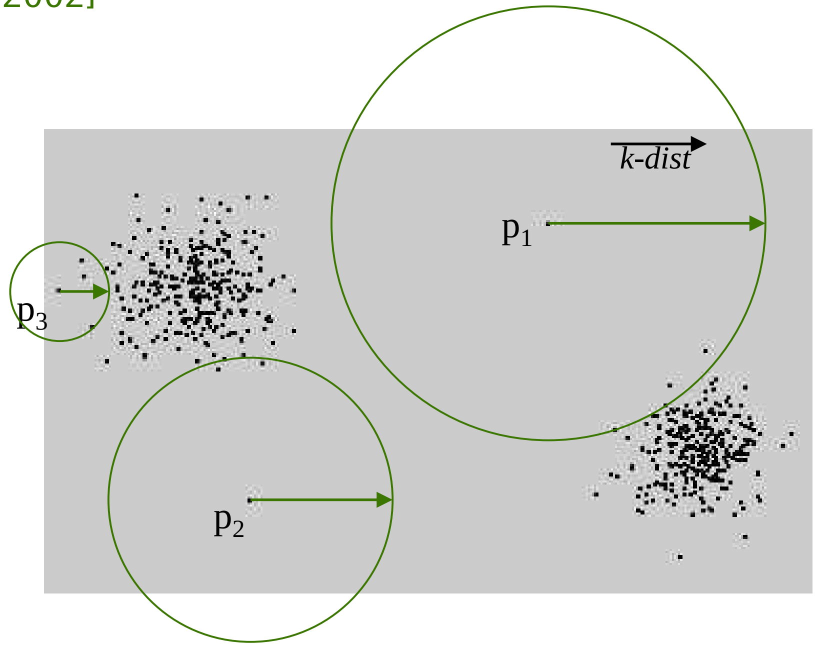
range-query with radius ε



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]

- DB-Outlier:
  binary-decision
- *k*NN-Outlier:
  ranking



- *k*NN-Outlier can be roughly considered Schönfinkeled or Curried version of DB-Outlier

k^{th} nearest neighbor based I

The outlier score of an object is given by the distance to its *k*-nearest neighbor.

- theoretically lowest outlier score: 0

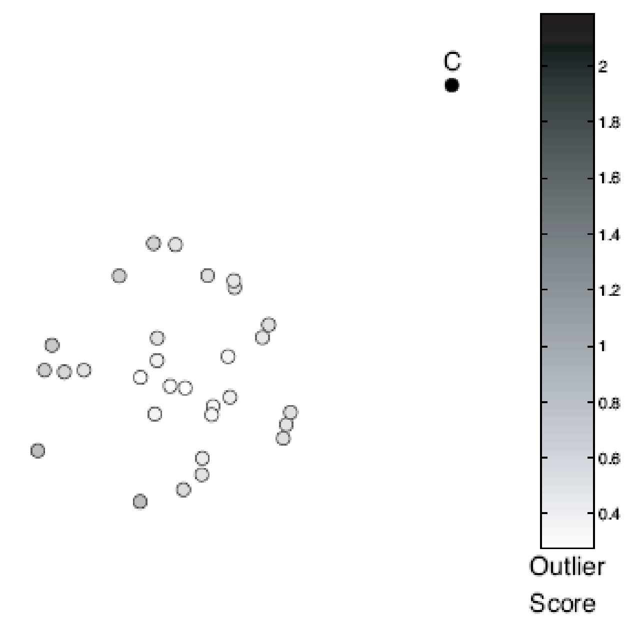


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

[Tan, Steinbach, Kumar 2006]

k=5

k^{th} nearest neighbor based II

- The outlier score is highly sensitive to the value of k

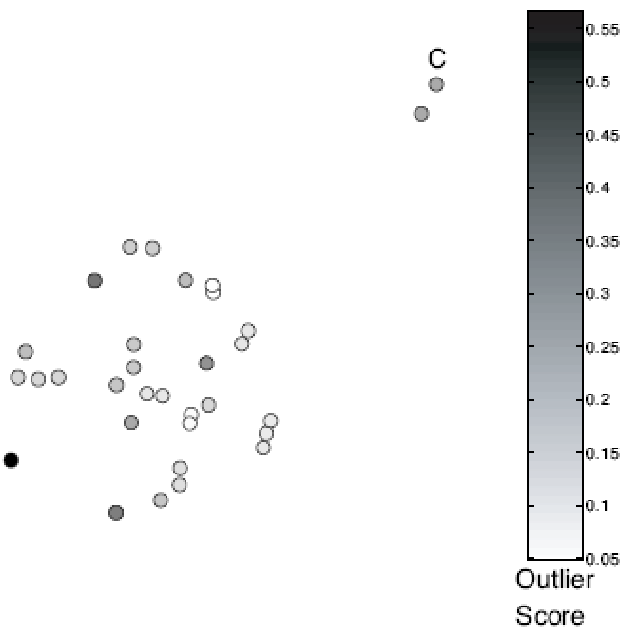


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

If k is to small, then a small number of close neighbors can cause low outlier scores.

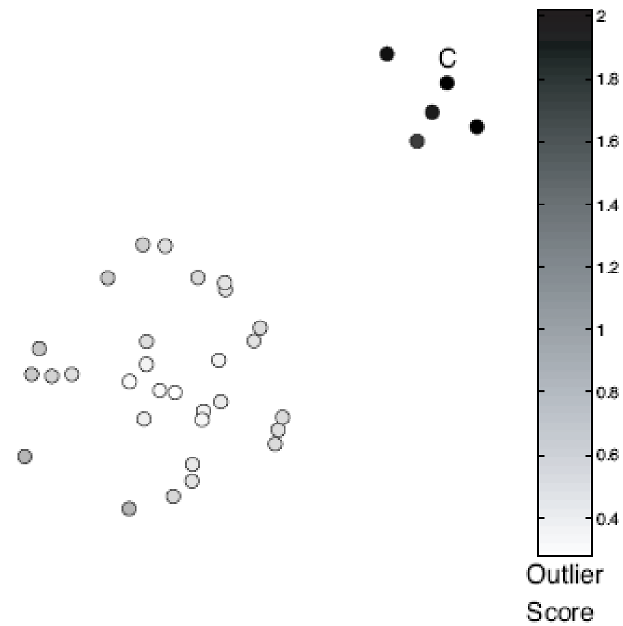


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

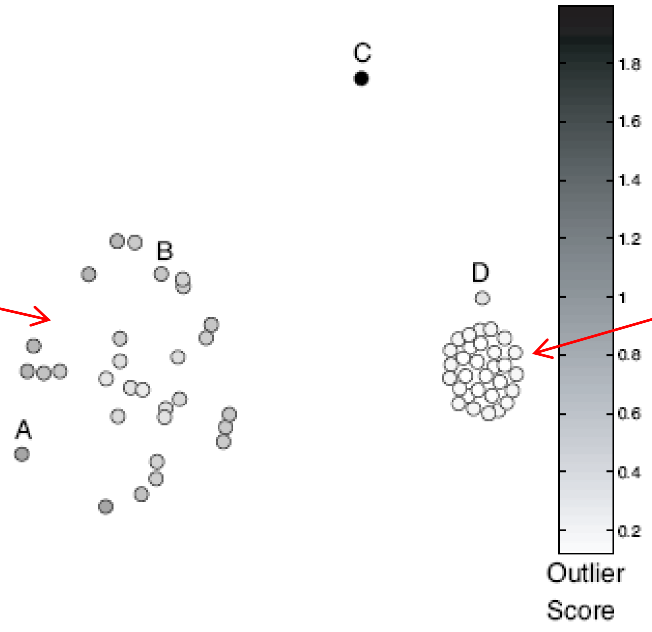
If k is to large, then all objects in a cluster with less than k objects might become outliers.

[Tan, Steinbach, Kumar 2006]

k^{th} nearest neighbor based III

- cannot handle datasets with regions of widely different densities due to the global threshold

sparse cluster



dense cluster

Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density. [Tan, Steinbach, Kumar 2006]

Überblick

- Einleitung
- Statistische Modellierung
- Depth-based Outliers
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Zusammenfassung

Density-based Approaches

General idea

- Compare the density around a point with the density around its local neighbors.
- The relative density of a point compared to its neighbors is computed as an outlier score.
- Approaches also differ in how to estimate density.

Basic assumption

- The density around a normal data object is similar to the density around its neighbors.
- The density around an outlier is considerably different to the density around its neighbors.

Density-based Approaches

- Different definitions of density:
  - e.g., # points within a specified distance *d* from the given object
- The choice of d is critical
  - If *d* is to small many normal points might be considered outliers
  - If *d* is to large, many outlier points will be considered as normal
- A global notion of density is problematic (as it is in clustering)
  - fails when data contain regions of different densities
- Solution: use a notion of density that is relative to the neighborhood of the object

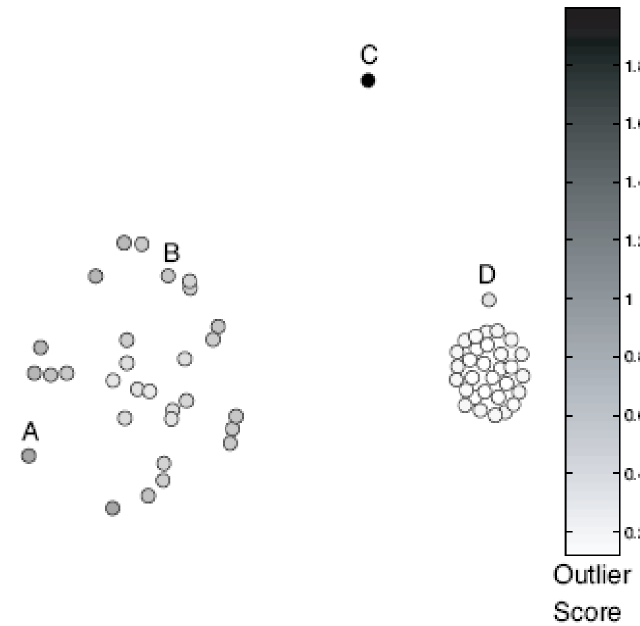


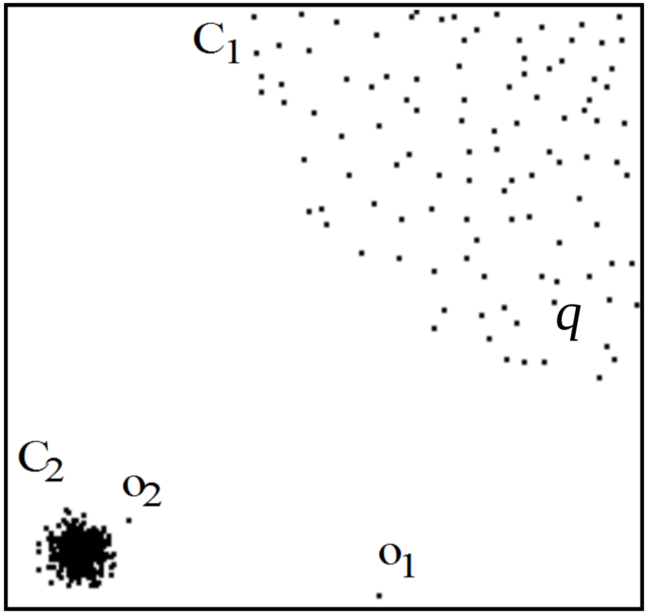
Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

[Tan, Steinbach, Kumar 2006]

D has a higher absolute density than A but compared to its neighborhood, D's density is lower.

Local Outlier Factor (LOF) [Breunig et al. 1999, 2000]

- Motivation:
  - Distance-based outlier detection models have problems with different densities
  - How to compare the neighborhood of points from areas of different densities?
  - Example
    - DB(ε,π)-outlier model
      - » Parameters ε and π cannot be chosen so that o_2 is an outlier but none of the points in cluster C_1 (e.g. *q*) is an outlier
    - Outliers based on kNN-distance
      - » kNN-distances of objects in C_1 (e.g. *q*) are larger than the kNN-distance of o_2

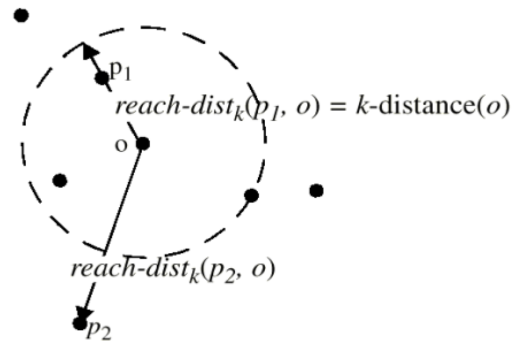


- Solution: consider relative density

Density-based Approaches

- Model
  - Reachability "distance"
    - Introduces a smoothing factor

$$reach\text{-}dist_k(p,o) = \max\{k\text{-distance}(o), dist(p,o)\}$$



  - Local reachability density (*lrd*) of point *p*
    - Inverse of the average reach-dists of the *k*NNs of *p*

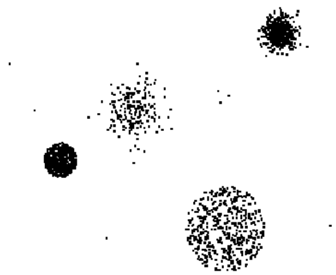
$$lrd_k(p) = \left(\frac{\sum_{o \in kNN(p)} reach\text{-}dist_k(p,o)}{Card(kNN(p))} \right)^{-1}$$

  - Local outlier factor (LOF) of point *p*
    - Average ratio of *lrd*s of neighbors of *p* and *lrd* of *p*

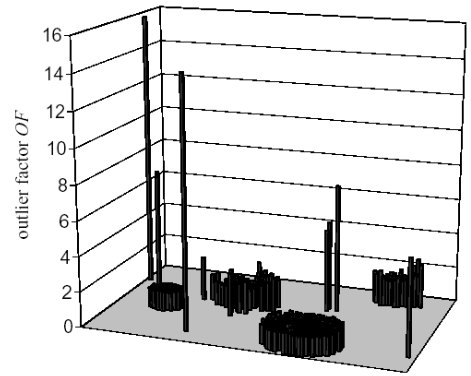
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card(kNN(p))}$$

Density-based Approaches

- Properties
  - LOF ≈ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
  - LOF >> 1: point is an outlier



Data set



LOFs (*MinPts* = 40)

- Discussion
  - Choice of *k* (*MinPts* in the original paper) specifies the reference set
  - Originally implements a local approach (resolution depends on the user's choice for *k*)
  - Outputs a scoring (assigns an LOF value to each point)

Density-based Approaches

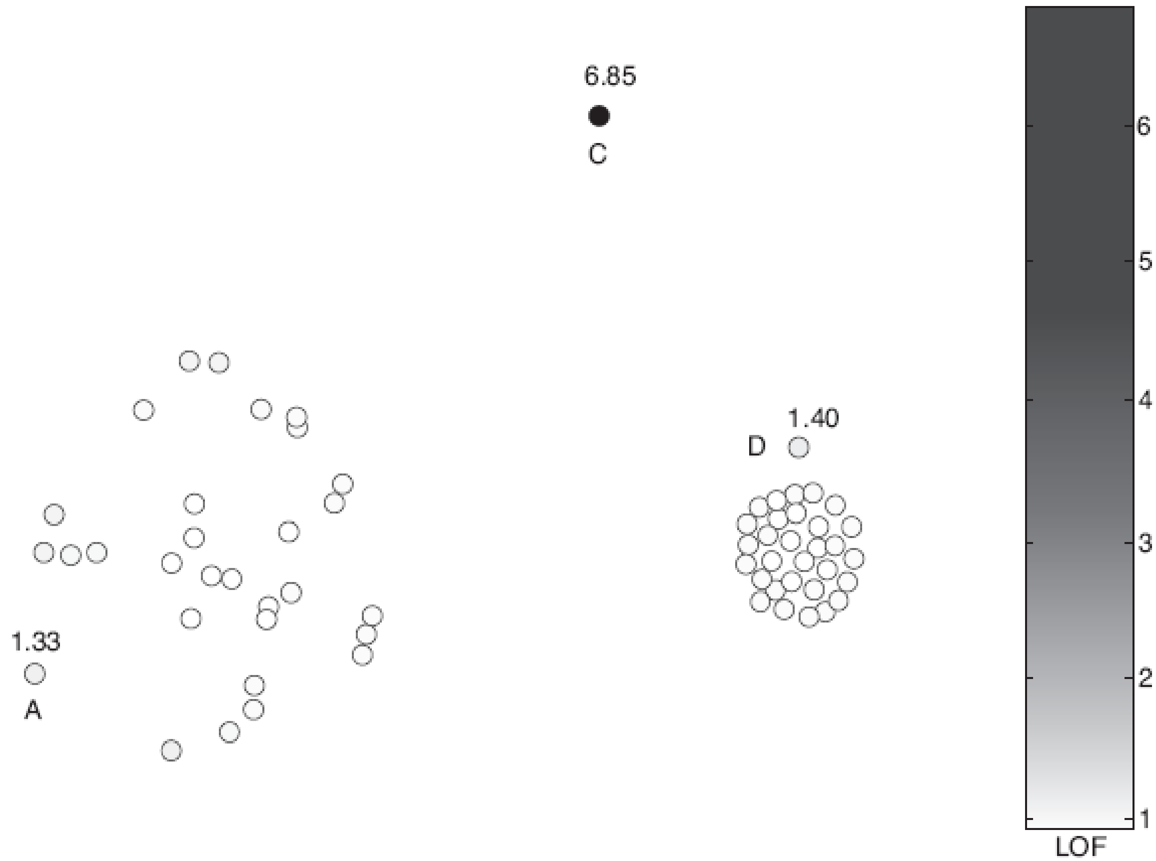


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

[Tan, Steinbach, Kumar 2006]

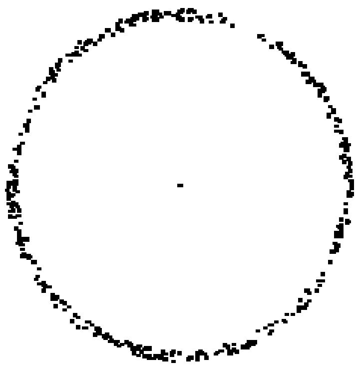
Variants of LOF

- Mining top-*n* local outliers [Jin et al. 2001]
  - Idea:
    - Usually, a user is only interested in the top-*n* outliers
    - Do not compute the LOF for all data objects => save runtime
  - Method
    - Compress data points into micro clusters using the CFs of BIRCH [Zhang et al. 1996]
    - Derive upper and lower bounds of the reachability distances, lrd-values, and LOF-values for points within a micro clusters
    - Compute upper and lower bounds of LOF values for micro clusters and sort results w.r.t. ascending lower bound
    - Prune micro clusters that cannot accommodate points among the top-*n* outliers (*n* highest LOF values)
    - Iteratively refine remaining micro clusters and prune points accordingly

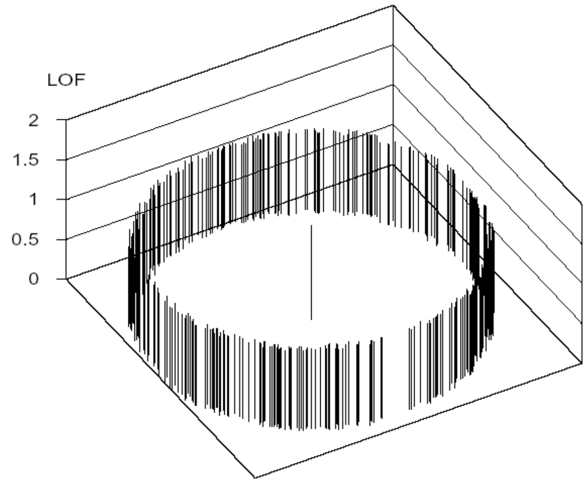
Variants of LOF (cont.)

– Connectivity-based outlier factor (COF) [Tang et al. 2002]

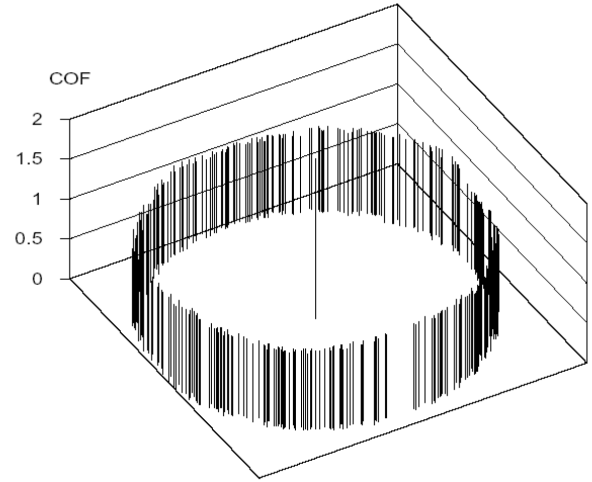
- Motivation
  - In regions of low density, it may be hard to detect outliers
  - Choose a low value for *k* is often not appropriate
- Solution
  - Treat "low density" and "isolation" differently
- Example



Data set



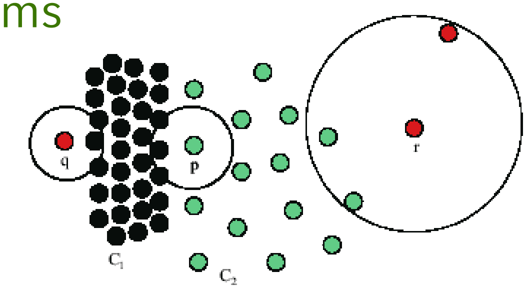
LOF



COF

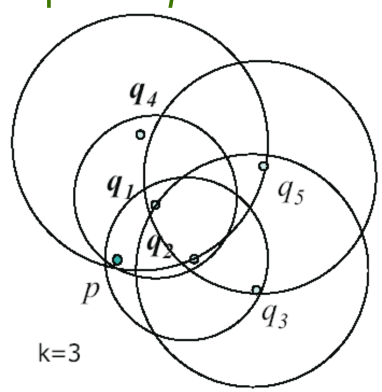
Influenced Outlierness (INFLO) [Jin et al. 2006]

– Motivation
  - If clusters of different densities are not clearly separated, LOF will have problems



Point *p* will have a higher LOF than points *q* or *r* which is counter intuitive

– Idea
  - Take symmetric neighborhood relationship into account
  - Influence space (*k*IS(*p*)) of a point *p* includes its kNNs (kNN(*p*)) and its reverse kNNs (RkNN(*p*))



$$kIS(p) = kNN(p) \cup RkNN(p)) = \{q_1, q_2, q_4\}$$

Density-based Approaches

- Model
  - Density is simply measured by the inverse of the *k*NN distance, i.e.,

    $den(p) = 1/k\text{-distance}(p)$

  - Influenced outlierness of a point p

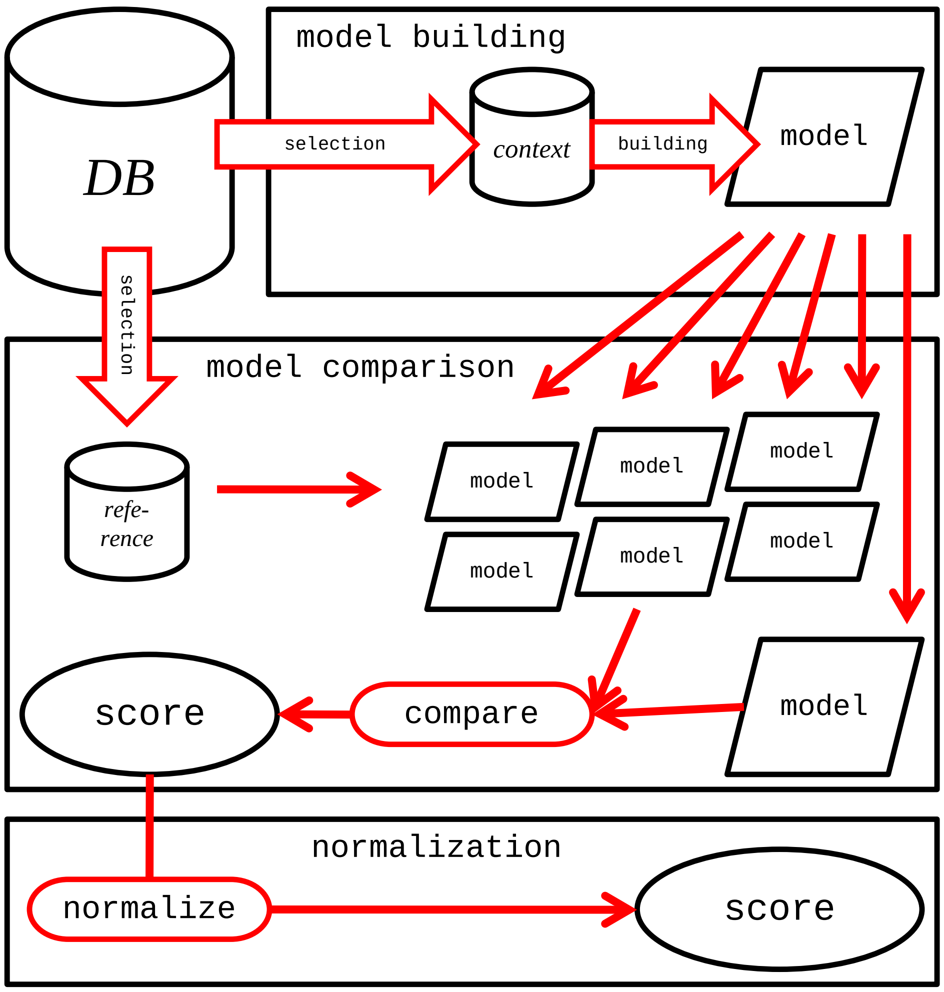
$$INFLO_k(p) = \frac{\sum_{o \in kIS(p)} den(o) \Big/ Card(kIS(p))}{den(p)}$$

  - INFLO takes the ratio of the average density of objects in the neighborhood of a point *p* (i.e., in *k*NN(*p*) ∪ R*k*NN(*p*)) to *p*'s density

Local vs. Global Approaches

general scheme [Schubert et al. 2012]:

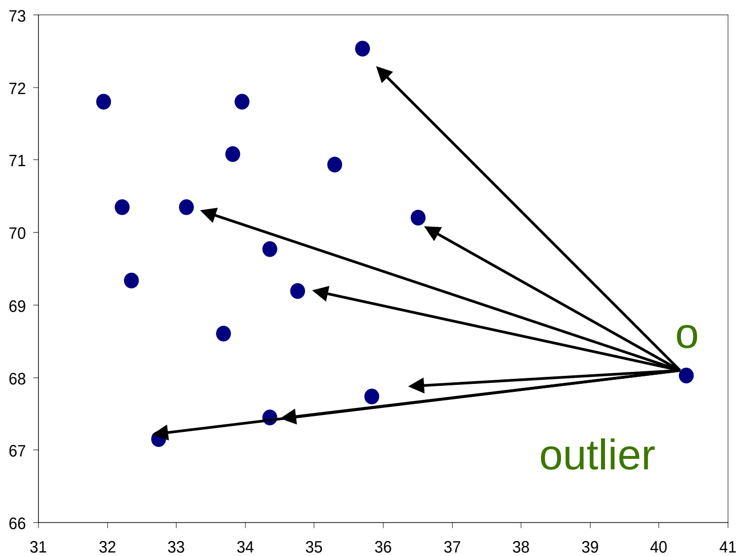
- context set for model building
- reference set for model comparison
- both, context set and reference set, could be:
  - identical
  - local
  - global
- outlier models come in degrees of "locality"

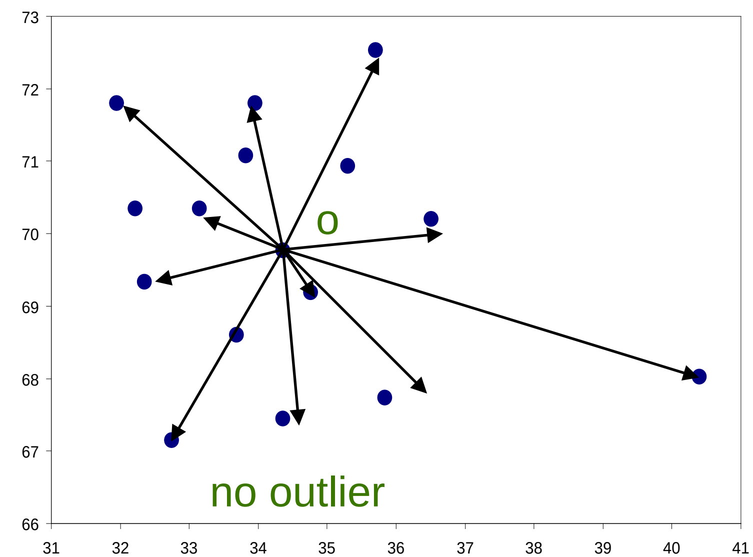


Überblick

ABOD – angle-based outlier degree [Kriegel et al. 2008]

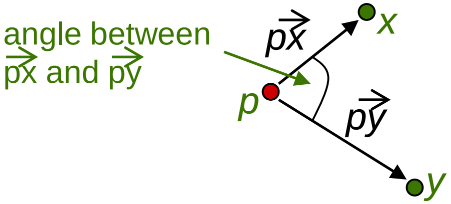
– Rational
  - Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
  - Object o is an outlier if most other objects are located in similar directions
  - Object o is no outlier if many other objects are located in varying directions

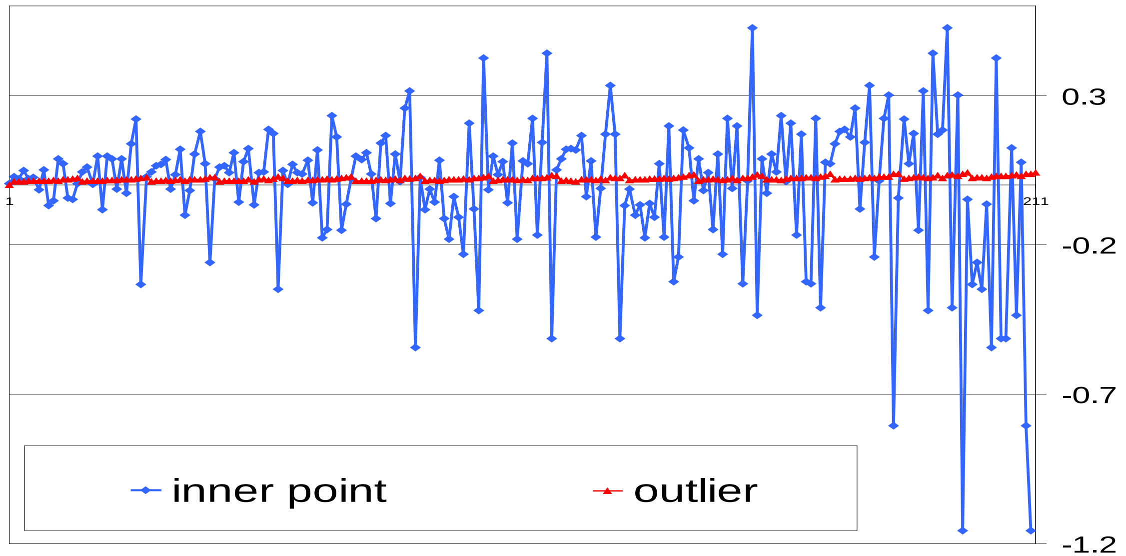




Angle-based Approach

- Basic assumption
  - Outliers are at the border of the data distribution
  - Normal points are in the center of the data distribution
- Model
  - Consider for a given point p the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two x,y from the database
  - Consider the spectrum of all these angles
  - The broadness of this spectrum is a score for the outlierness of a point





Angle-based Approach

- Model (cont.)
  - Measure the variance of the angle spectrum
  - Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)

$$ABOD(p) = \underset{x,y \in DB}{VAR}\left(\frac{\left\langle \vec{xp}, \vec{yp} \right\rangle}{\left\| \vec{xp} \right\|^2 \cdot \left\| \vec{yp} \right\|^2} \right)$$

  - Properties
    - Small ABOD => outlier
    - High ABOD => no outlier

Angle-based Approach

- Algorithms
  - Naïve algorithm is in O(n^3)
  - Approximate algorithm based on random sampling for mining top-*n* outliers
    - Do not consider all pairs of other points *x*,*y* in the database to compute the angles
    - Compute ABOD based on samples => lower bound of the real ABOD
    - Filter out points that have a high lower bound
    - Refine (compute the exact ABOD value) only for a small number of points
- Discussion
  - Global approach to outlier detection
  - Outputs an outlier score

    (inversely scaled:

    high ABOD score => inlier,

    low ABOD score => outlier)

Überblick

- Einleitung
- Statistische Modellierung
- Depth-based Outliers
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Zusammenfassung

Zusammenfassung

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"

Zusammenfassung

- 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

Zusammenfassung

- 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

Zusammenfassung

- 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

Zusammenfassung

- 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

Zusammenfassung

- 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

Literature

Achtert, E., Kriegel, H.-P., Reichert, L., Schubert, E., Wojdanowski, R., Zimek, A. 2010. Visual Evaluation of Outlier Detection Models. In Proc. International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan.

Angiulli, F. and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In Proc. European Conf. on Principles of Knowledge Discovery and Data Mining, Helsinki, Finland.

Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.

Barnett, V. 1978. The study of outliers: purpose and model. Applied Statistics, 27(3), 242–250.

Bay, S.D. and Schwabacher, M. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 1999. OPTICS-OF: identifying local outliers. In Proc. European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD), Prague, Czech Republic.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 2000. LOF: identifying density-based local outliers. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

Literature

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.

Fan, H., Zaïane, O., Foss, A., and Wu, J. 2006. A nonparametric outlier detection for efficiently discovering top-n outliers from engineering data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.

Ghoting, A., Parthasarathy, S., and Otey, M. 2006. Fast mining of distance-based outliers in high dimensional spaces. In Proc. SIAM Int. Conf. on Data Mining (SDM), Bethesda, ML.

Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK.

Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.

Jin, W., Tung, A., and Han, J. 2001. Mining top-*n* local outliers in large databases. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA.

Jin, W., Tung, A., Han, J., and Wang, W. 2006. Ranking outliers using symmetric neighborhood relationship. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.

Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY.

Knorr, E.M. and Ng, R.T. 1997. A unified approach for mining outliers. In Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada.

Literature

Knorr, E.M. and NG, R.T. 1998. Algorithms for mining distance-based outliers in large datasets. In Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY.

Knorr, E.M. and Ng, R.T. 1999. Finding intensional knowledge of distance-based outliers. In Proc. Int. Conf. on Very Large Data Bases (VLDB), Edinburgh, Scotland.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009a. LoOP: Local Outlier Probabilities. In Proc. ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2011. Interpreting and Unifying Outlier Scores. In Proc. 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ.

Kriegel, H.-P., Schubert, M., and Zimek, A. 2008. Angle-based outlier detection, In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV.

McCallum, A., Nigam, K., and Ungar, L.H. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA.

Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. 2003. LOCI: Fast outlier detection using the local correlation integral. In Proc. IEEE Int. Conf. on Data Engineering (ICDE), Hong Kong, China.

Preparata, F. and Shamos, M. 1988. Computational Geometry: an Introduction. Springer Verlag.

Ramaswamy, S. Rastogi, R. and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

Literature

Rousseeuw, P.J. and Leroy, A.M. 1987. Robust Regression and Outlier Detection. John Wiley.

Ruts, I. and Rousseeuw, P.J. 1996. Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis, 23, 153–168.

Schubert E., Zimek A., Kriegel H.-P. 2012. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery, online first, DOI: 10.1007/s10618-012-0300-z.

Tao Y., Xiao, X. and Zhou, S. 2006. Mining distance-based outliers from large databases in any metric space. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), New York, NY.

Tan, P.-N., Steinbach, M., and Kumar, V. 2006. Introduction to Data Mining. Addison Wesley.

Tang, J., Chen, Z., Fu, A.W.-C., and Cheung, D.W. 2002. Enhancing effectiveness of outlier detections for low density patterns. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan.

Tukey, J. 1977. Exploratory Data Analysis. Addison-Wesley.

Zhang, T., Ramakrishnan, R., Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Montreal, Canada.

Was haben Sie gelernt?

- Outlier: Intuition, aber auch Vagheit des Konzepts
- Kategorien, Eigenschaften von Outlier-Modellen
- Probabilistisches Modell
- Tiefen-basierte Modelle
- Distanz-basierte Modelle
  - DB-Outlier
  - kNN-basierte Modelle
- Dichte-basierte Modelle
  - LOF: Motivation, Modell
  - Varianten von LOF (top-*n*, connectivity, influence set)
- Lokalität
- Winkel-basiertes Modell