The 2010 SIAM International Conference on Data Mining

Outlier Detection Techniques

Hans-Peter Kriegel, Peer Kröger, Arthur Zimek

Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de
{kriegel,kroegerp,zimek}@dbs.ifi.lmu.de

General Issues

1. Please feel free to ask questions at any time during the presentation
2. Aim of the tutorial: get the big picture
   - NOT in terms of a long list of methods and algorithms
   - BUT in terms of the basic algorithmic approaches
   - Sample algorithms for these basic approaches will be sketched
     - The selection of the presented algorithms is somewhat arbitrary
     - Please don’t mind if your favorite algorithm is missing
     - Anyway you should be able to classify any other algorithm not covered here by means of which of the basic approaches is implemented
3. The revised version of tutorial notes will soon be available on our websites

What is an outlier?

Definition of Hawkins [Hawkins 1980]:

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

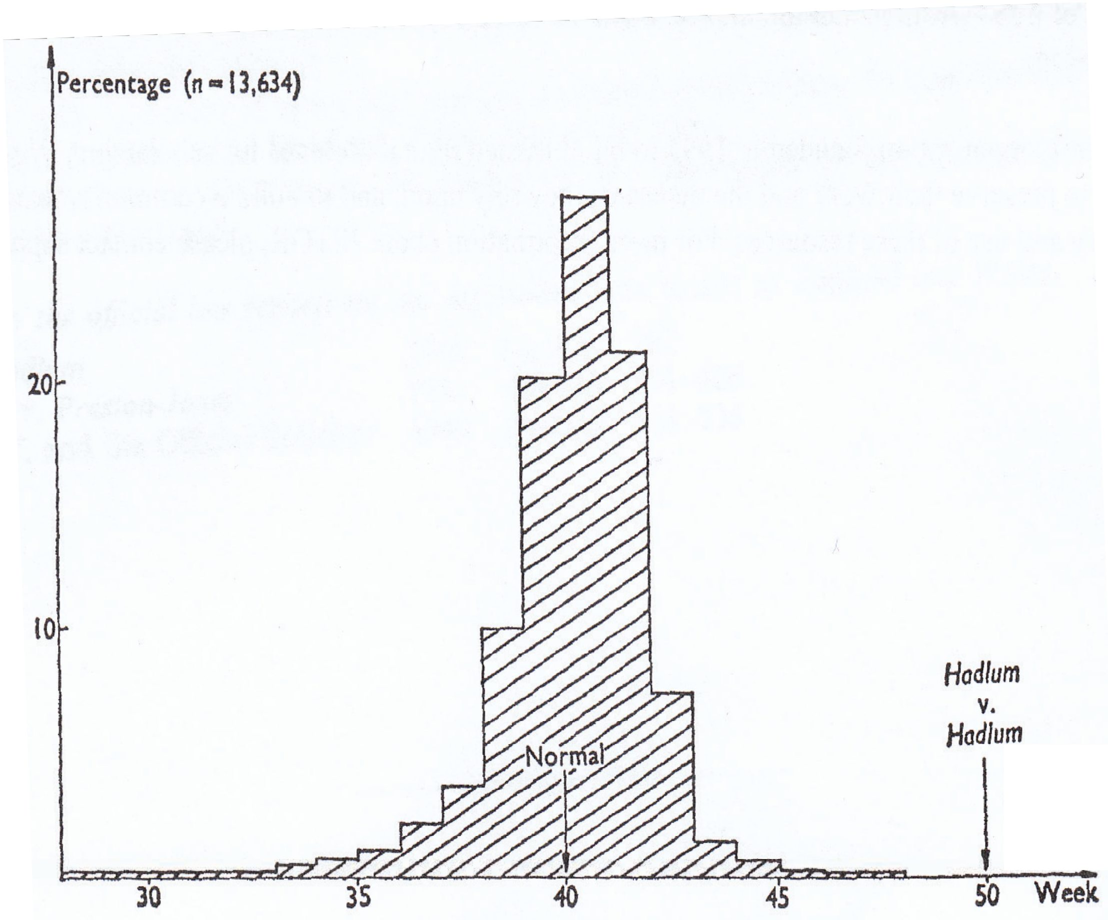
Statistics-based intuition

- Normal data objects follow a "generating mechanism", e.g. some given statistical process
- Abnormal objects deviate from this generating mechanism

Introduction

- Example: Hadlum vs. Hadlum (1949) [Barnett 1978]

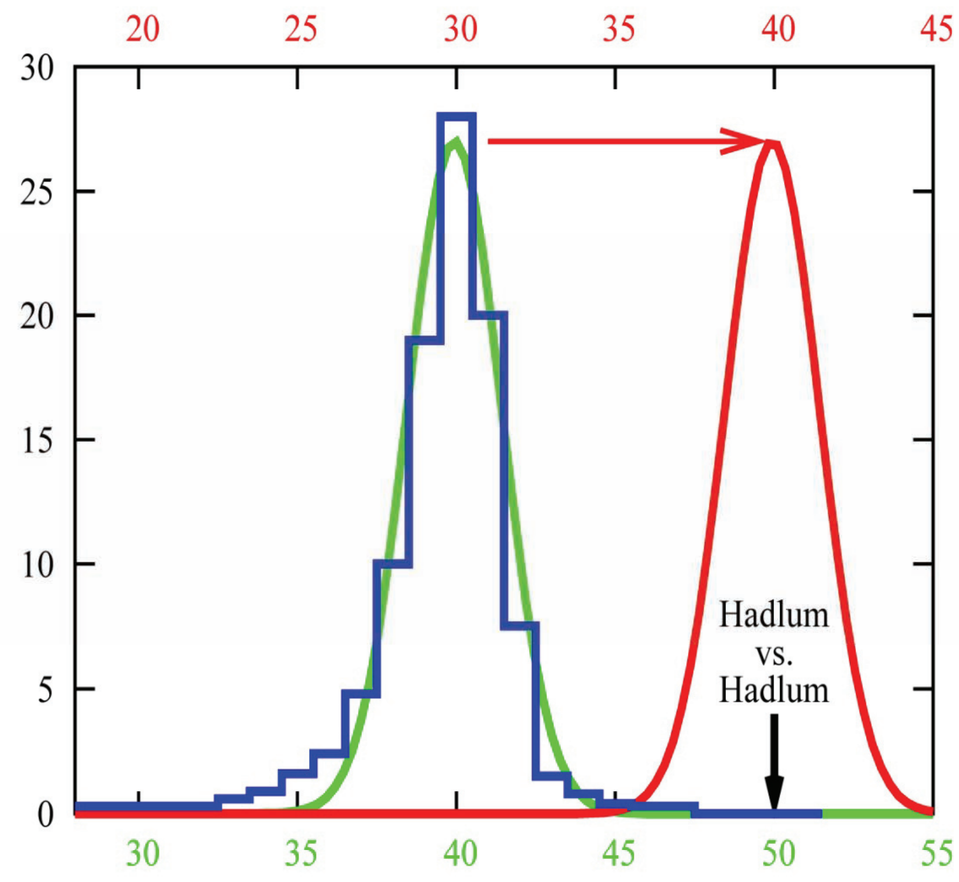
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.



Distribution of human gestation periods.

Introduction

- Example: Hadlum vs. Hadlum (1949) [Barnett 1978]
  - blue: statistical basis (13634 observations of gestation periods)
  - green: assumed underlying Gaussian process
    - Very low probability for the birth of Mrs. Hadlums child for being generated by this process
  - red: assumption of Mr. Hadlum (another Gaussian process responsible for the observed birth, where the gestation period responsible)
    - Under this assumption the gestation period has an average duration and highest-possible probability



Introduction

- Sample applications of outlier detection
  - Fraud detection
    - Purchasing behavior of a credit card owner usually changes when the card is stolen
    - Abnormal buying patterns can characterize credit card abuse
  - Medicine
    - Unusual symptoms or test results may indicate potential health problems of a patient
    - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
  - Public health
    - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
    - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Introduction

- Sample applications of outlier detection (cont.)
  - Sports statistics
    - In many sports, various parameters are recorded for players in order to evaluate the players' performances
    - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
    - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
  - Detecting measurement errors
    - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
    - Abnormal values could provide an indication of a measurement error
    - Removing such errors can be important in other data mining and data analysis tasks
    - “One person‘s noise could be another person‘s signal.”
  - …

Introduction

- Discussion of the basic intuition based on Hawkins
  – Data is usually multivariate, i.e., multi-dimensional
    => basic model is univariate, i.e., 1-dimensional
  – There is usually more than one generating mechanism/statistical process underlying the data
    => basic model assumes only one “normal” generating mechanism
  – Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
    => basic model assumes that outliers are rare observations
- Consequence: a lot of models and approaches have evolved in the past years in order to exceed these assumptions and it is not easy to keep track with this evolution.
- New models often involve typical, new, though usually hidden assumptions and restrictions.

Introduction

- General application scenarios
  - Supervised scenario
    - In some applications, training data with normal and abnormal data objects are provided
    - There may be multiple normal and/or abnormal classes
    - Often, the classification problem is highly unbalanced
  - Semi-supervised Scenario
    - In some applications, only training data for the normal class(es) (or only the abnormal class(es)) are provided
  - Unsupervised Scenario
    - In most applications there are no training data available
- In this tutorial, we focus on the unsupervised scenario

Introduction

- Are outliers just a side product of some clustering algorithms?
  - Many clustering algorithms do not assign all points to clusters but account for noise objects
  - Look for outliers by applying one of those algorithms and retrieve the noise set
  - Problem:
    - Clustering algorithms are optimized to find clusters rather than outliers
    - Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters
    - A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers

Introduction

- We will focus on three different classification approaches
  - Global versus local outlier detection

    Considers the set of reference objects relative to which each point's "outlierness" is judged

  - Labeling versus scoring outliers

    Considers the output of an algorithm

  - Modeling properties

    Considers the concepts based on which "outlierness" is modeled

NOTE: we focus on models and methods for Euclidean data but many of those can be also used for other data types (because they only require a distance measure)

Introduction

- Global versus local approaches
  - Considers the resolution of the reference set w.r.t. which the "outlierness" of a particular data object is determined
  - Global approaches
    - The reference set contains all other data objects
    - Basic assumption: there is only one normal mechanism
    - Basic problem: other outliers are also in the reference set and may falsify the results
  - Local approaches
    - The reference contains a (small) subset of data objects
    - No assumption on the number of normal mechanisms
    - Basic problem: how to choose a proper reference set
  - NOTE: Some approaches are somewhat in between
    - The resolution of the reference set is varied e.g. from only a single object (local) to the entire database (global) automatically or by a user-defined input parameter

Introduction

- Labeling versus scoring
  - Considers the output of an outlier detection algorithm
  - Labeling approaches
    - Binary output
    - Data objects are labeled either as normal or outlier
  - Scoring approaches
    - Continuous output
    - For each object an outlier score is computed (e.g. the probability for being an outlier)
    - Data objects can be sorted according to their scores
  - Notes
    - Many scoring approaches focus on determining the top-*n* outliers (parameter *n* is usually given by the user)
    - Scoring approaches can usually also produce binary output if necessary (e.g. by defining a suitable threshold on the scoring values)

Introduction

- Approaches classified by the properties of the underlying modeling approach
  - Model-based Approaches
    - Rational
      - Apply a model to represent normal data points
      - Outliers are points that do not fit to that model
    - Sample approaches
      - Probabilistic tests based on statistical models
      - Depth-based approaches
      - Deviation-based approaches
      - Some subspace outlier detection approaches

Introduction

- Proximity-based Approaches
  - Rational
    - Examine the spatial proximity of each object in the data space
    - If the proximity of an object considerably deviates from the proximity of other objects it is considered an outlier
  - Sample approaches
    - Distance-based approaches
    - Density-based approaches
    - Some subspace outlier detection approaches
- Angle-based approaches
  - Rational
    - Examine the spectrum of pairwise angles between a given point and all other points
    - Outliers are points that have a spectrum featuring high fluctuation

DATABASE SYSTEMS GROUP

Outline

1. Introduction √
2. Statistical Tests
3. Depth-based Approaches
4. Deviation-based Approaches
5. Distance-based Approaches
6. Density-based Approaches
7. High-dimensional Approaches
8. Summary

statistical model

model based on spatial proximity

adaptation of different models to a special problem

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

Statistical Tests

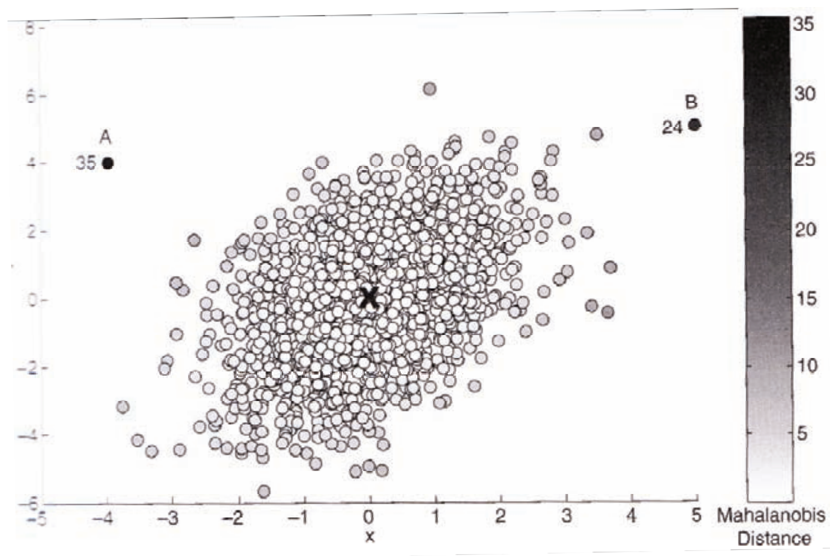
- Probability density function of a multivariate normal distribution

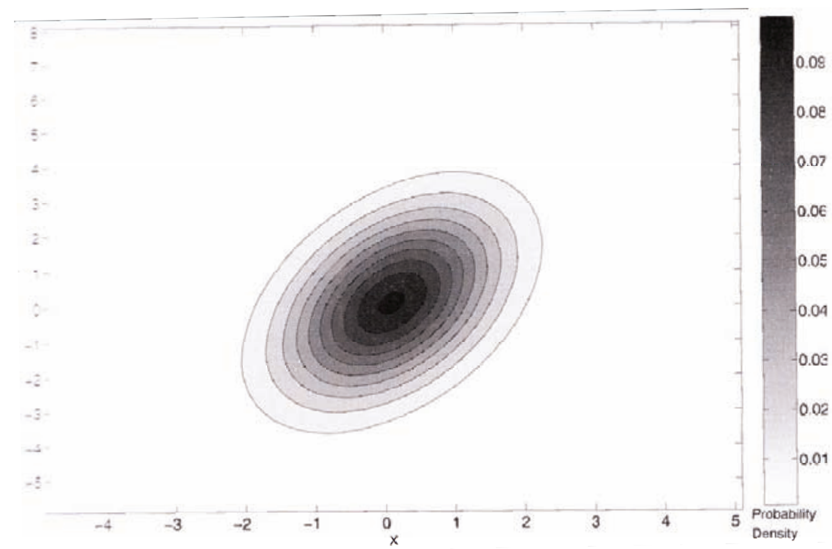
$$N(x)=\frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}e^{-\frac{(x-\mu)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)}{2}}$$

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

Statistical Tests

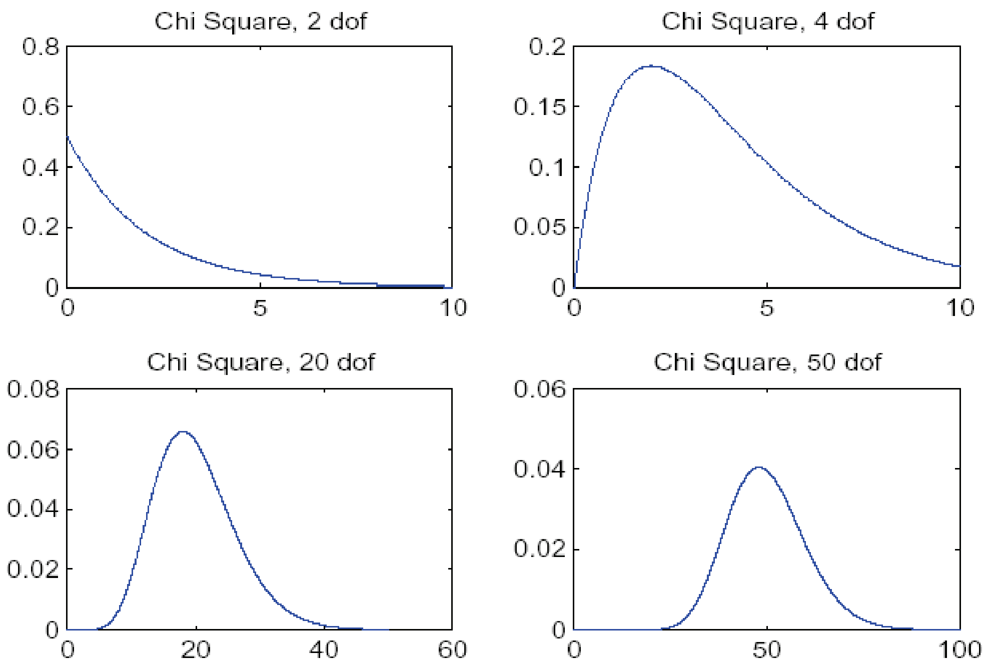
- Visualization (2D) [Tan et al. 2006]





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

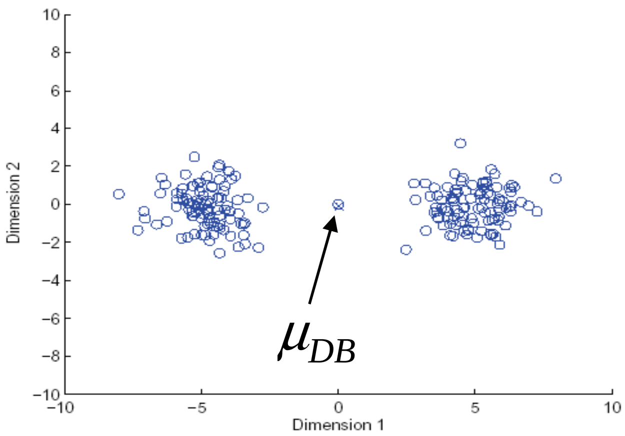
y-axis: frequency of observation

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

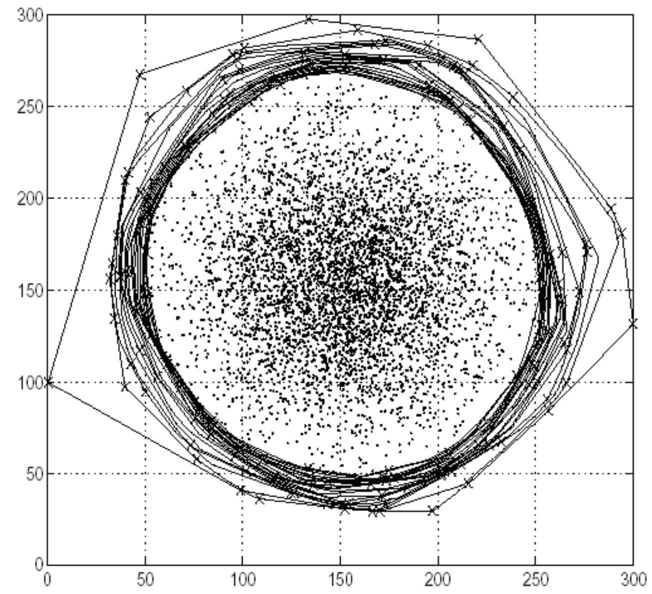


Outline

1. Introduction √
2. Statistical Tests √
3. Depth-based Approaches
4. Deviation-based Approaches
5. Distance-based Approaches
6. Density-based Approaches
7. High-dimensional Approaches
8. Summary

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

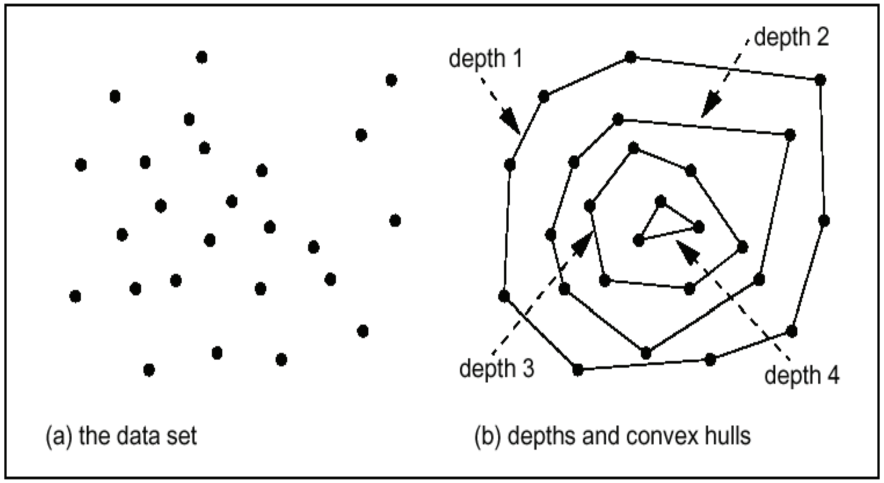


Picture taken from [Johnson et al. 1998]

- 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

Outline

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

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:

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

- 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

Outline

Distance-based Approaches

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

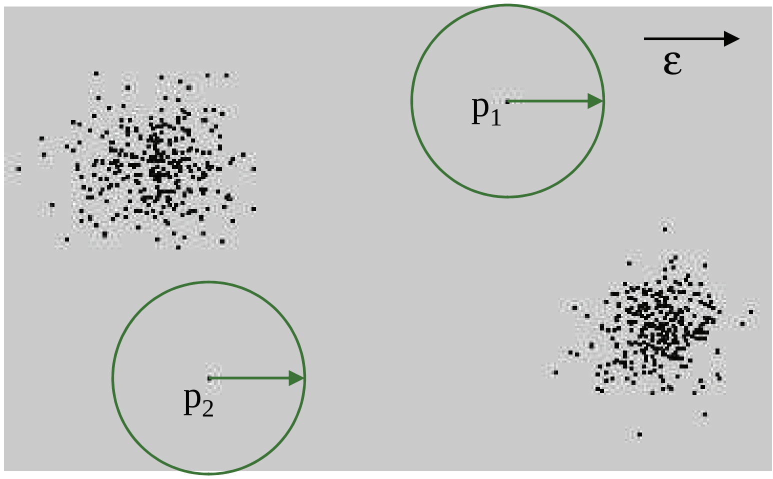
- Basic Assumption
  – Normal data objects have a dense neighborhood
  – Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

Distance-based Approaches

- 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

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

range-query with radius ε



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

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
```

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

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

Distance-based Approaches

– Sample Algorithms (cont.)
- RBRP [Ghoting et al. 2006],
  - Idea: try to increase the cut-off as uick 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*)

Distance-based Approaches

- 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 than equal to *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*)

Distance-based Approaches

- 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* so that at each step at least one point changes from being outlier to being a member of a cluster
    - Cluster is defined similar as in DBSCAN [Ester et al 1996] 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

Outline

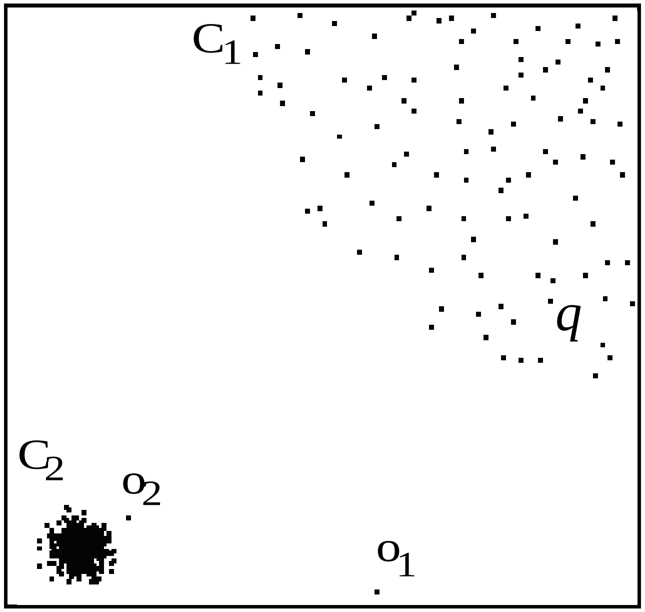
Density-based Approaches

- 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

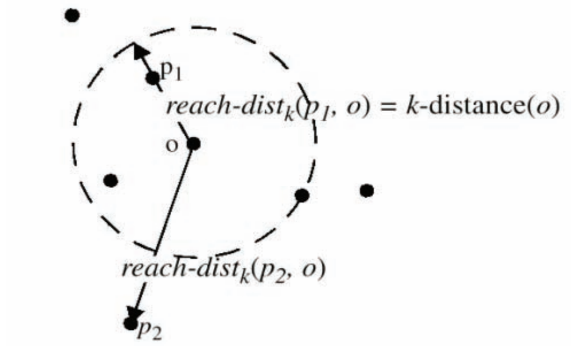
- Basic assumption
  - The density around a normal data object is similar to the density around its neighbors
  - The density around an outlier is considerably different to the density around its neighbors

Density-based Approaches

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



– Model
  - Reachability distance
    – Introduces a smoothing factor

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

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

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

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

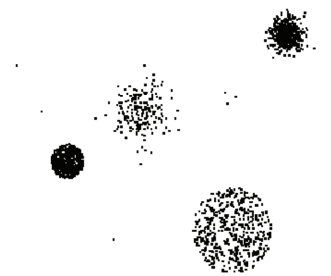
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card\left(kNN(p)\right)}$$

DATABASE SYSTEMS GROUP

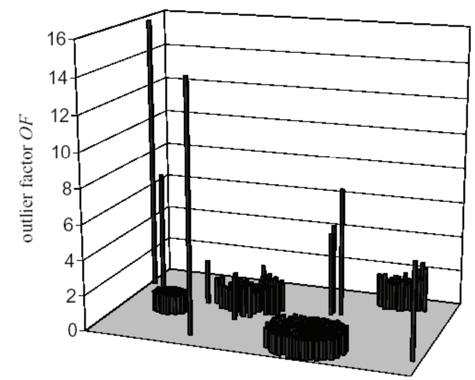
Density-based Approaches

LMU

– 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



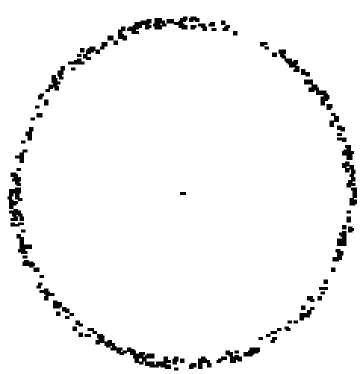
LOFs (*MinPts* = 40)

– 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)

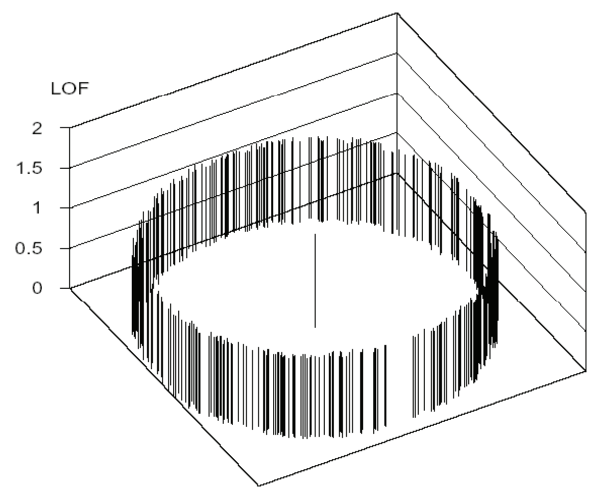
Density-based Approaches

- 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

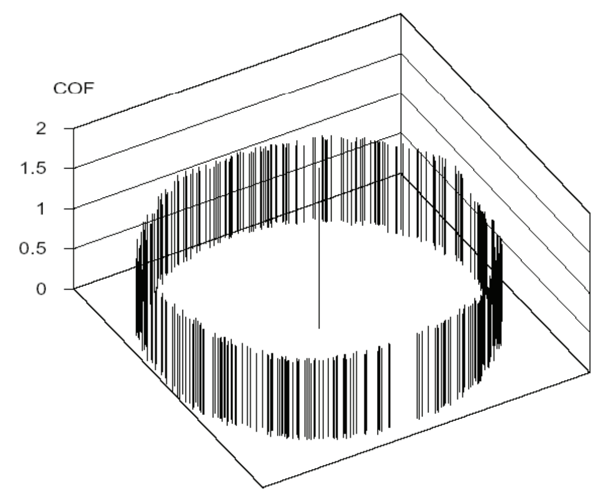
- Variants of LOF (cont.)
  - Connectivity-based outlier factor (COF) [Tang et al. 2002]
    - Motivation
      - In regions of low density, it may be hard to detect outliers
      - Choose a low value for *k* is often not appropriate
    - Solution
      - Treat “low density” and “isolation” differently
    - Example



Data set



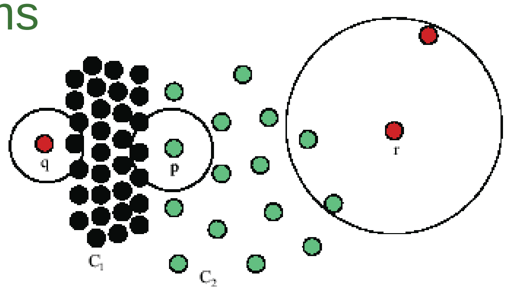
LOF



COF

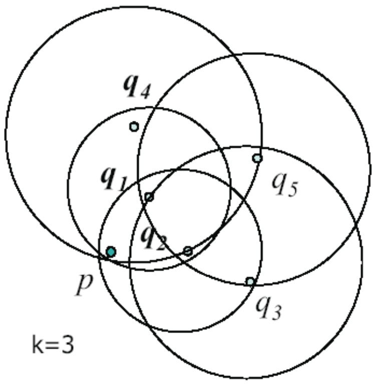
Density-based Approaches

- 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 (kIS(p)) of a point p includes its kNNs (kNN(p)) and its reverse kNNs (RkNN(p))



$k\text{IS}(p) = k\text{NN}(p) \cup \text{R}k\text{NN}(p))$

$= \{q_1, q_2, q_4\}$

Density-based Approaches

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

    den(*p*) = 1/*k*-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

– Proposed algorithms for mining top-*n* outliers
  - Index-based
  - Two-way approach
  - Micro cluster based approach

Density-based Approaches

- Properties
  - Similar to LOF
  - INFLO ≈ 1: point is in a cluster
  - INFLO >> 1: point is an outlier

- Discussion
  - Outputs an outlier score
  - Originally proposed as a local approach (resolution of the reference set *k*IS can be adjusted by the user setting parameter *k*)

Density-based Approaches

- Local outlier correlation integral (LOCI) [Papadimitriou et al. 2003]
  - Idea is similar to LOF and variants
  - Differences to LOF
    - Take the ε-neighborhood instead of *k*NNs as reference set
    - Test multiple resolutions (here called "granularities") of the reference set to get rid of any input parameter
  - Model
    - ε-neighborhood of a point p: $N(p,\varepsilon) = \{q \mid dist(p,q) \leq \varepsilon\}$
    - Local density of an object p: number of objects in $N(p,\varepsilon)$
    - Average density of the neighborhood

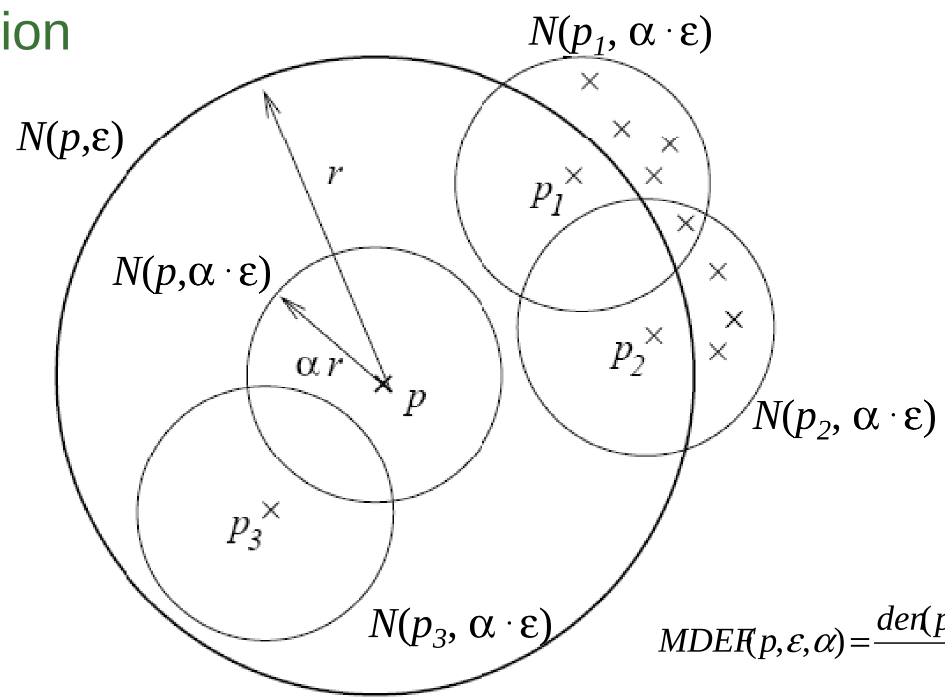
$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

    - Multi-granularity Deviation Factor (MDEF)

$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

Density-based Approaches

– Intuition



$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

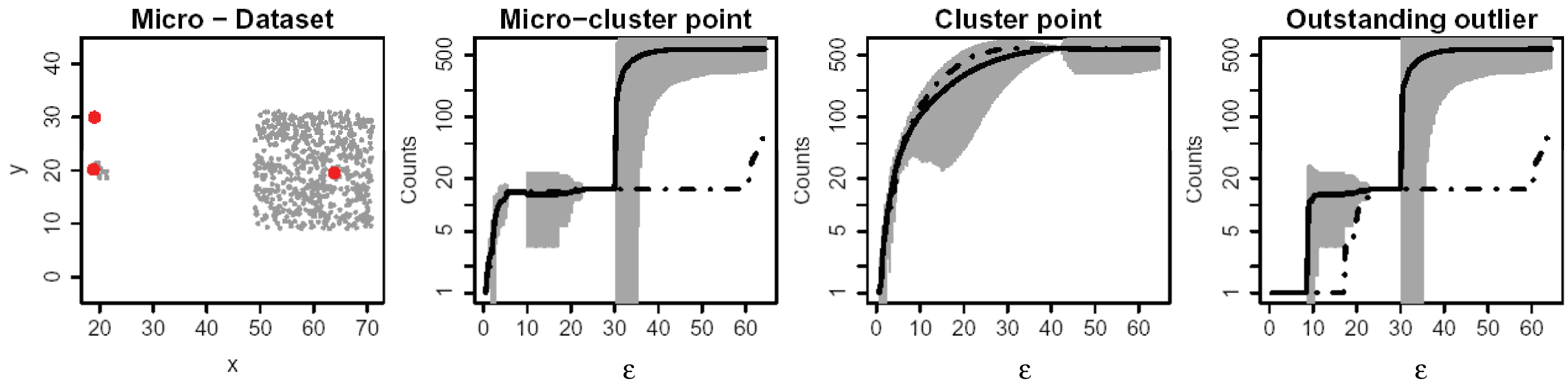
– σMDEF(p,ε,α) is the normalized standard deviation of the densities of all points from $N(p,\varepsilon)$

– Properties

- MDEF = 0 for points within a cluster
- MDEF > 0 for outliers   or MDEF > 3·σMDEF => outlier

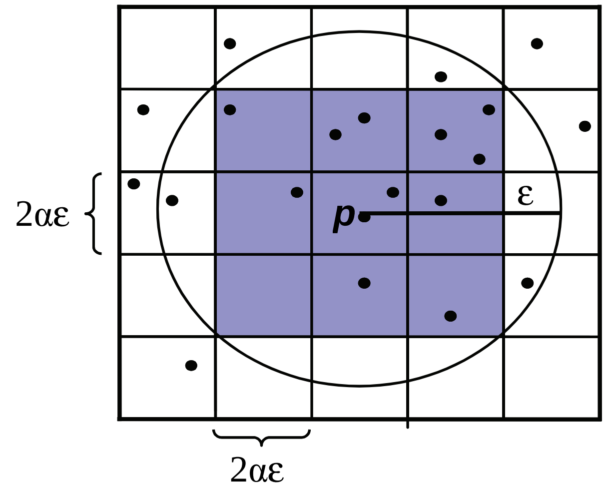
Density-based Approaches

– Features
  - Parameters ε and α are automatically determined
  - In fact, all possible values for ε are tested
  - LOCI plot displays for a given point *p* the following values w.r.t. ε
    – *Card*(*N*(*p*, α·ε))
    – *den*(*p*, ε, α) with a border of ± 3·σ*den*(*p*, ε, α)



Density-based Approaches

– Algorithms
  - Exact solution is rather expensive (compute MDEF values for all possible ε values)
  - aLOCI: fast, approximate solution
    – Discretize data space using a grid with side length 2αε
    – Approximate range queries trough grid cells
    – ε - neighborhood of point p: ζ(p,ε) all cells that are completely covered by ε-sphere around *p*
    – Then,



$$Card(N(q,\alpha\cdot\varepsilon)) = \frac{\sum_{c_j\in\zeta(p,\varepsilon)} c_j^{\,2}}{\sum_{c_j\in\zeta(p,\varepsilon)} c_j}$$

where c_j is the object count the corresponding cell

    – Since different ε values are needed, different grids are constructed with varying resolution
    – These different grids can be managed efficiently using a Quad-tree

Density-based Approaches

– Discussion
  - Exponential runtime w.r.t. data dimensionality
  - Output:
    – Label: it MDEF of a point > 3·σMDEF then this point is marked as outlier
    – LOCI plot
      » At which resolution is a point an outlier (if any)
      » Additional information such as diameter of clusters, distances to clusters, etc.
  - All interesting resolutions, i.e., possible values for ε, (from local to global) are tested

Outline

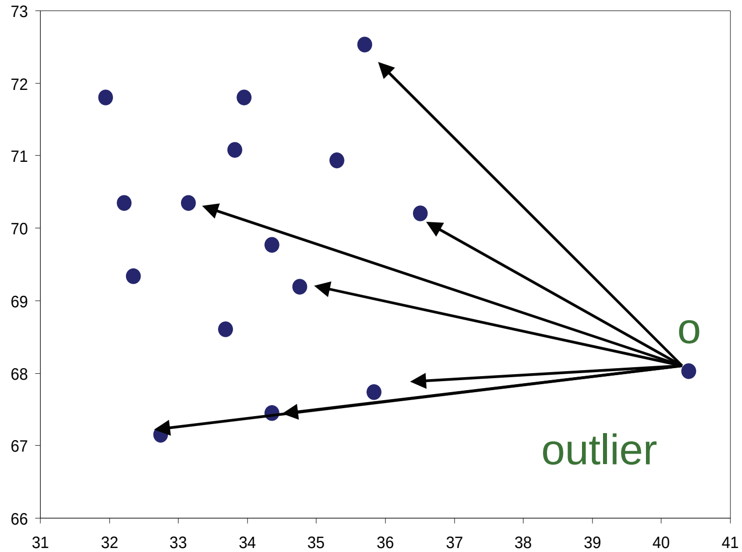
1. Introduction √
2. Statistical Tests √
3. Depth-based Approaches √
4. Deviation-based Approaches √
5. Distance-based Approaches √
6. Density-based Approaches √
7. High-dimensional Approaches
8. Summary

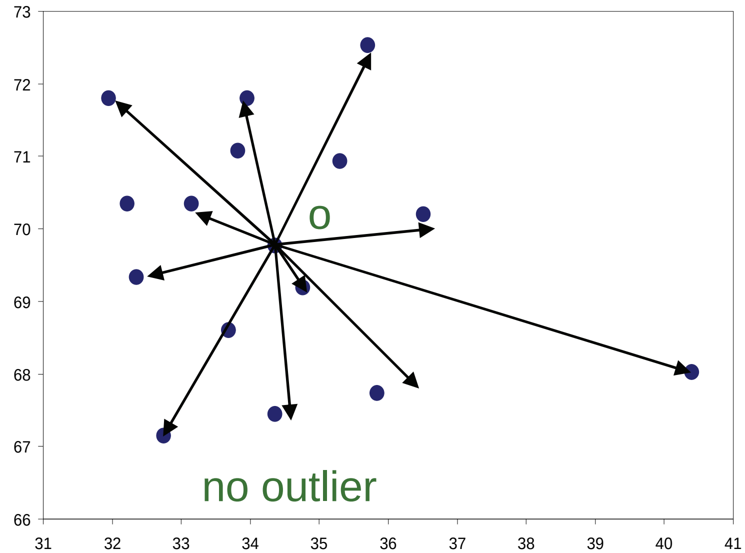
High-dimensional Approaches

- Challenges
  - Curse of dimensionality
    - Relative contrast between distances decreases with increasing dimensionality
    - Data is very sparse, almost all points are outliers
    - Concept of neighborhood becomes meaningless
  - Solutions
    - Use more robust distance functions and find full-dimensional outliers
    - Find outliers in projections (subspaces) of the original feature space

High-dimensional Approaches

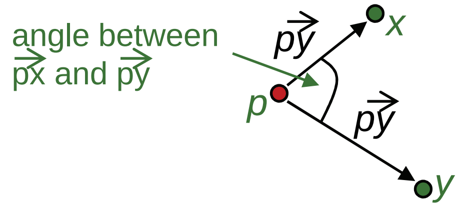
- ABOD – angle-based outlier degree [Kriegel et al. 2008]
  - Rational
    - Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
    - Object o is an outlier if most other objects are located in similar directions
    - Object o is no outlier if many other objects are located in varying directions

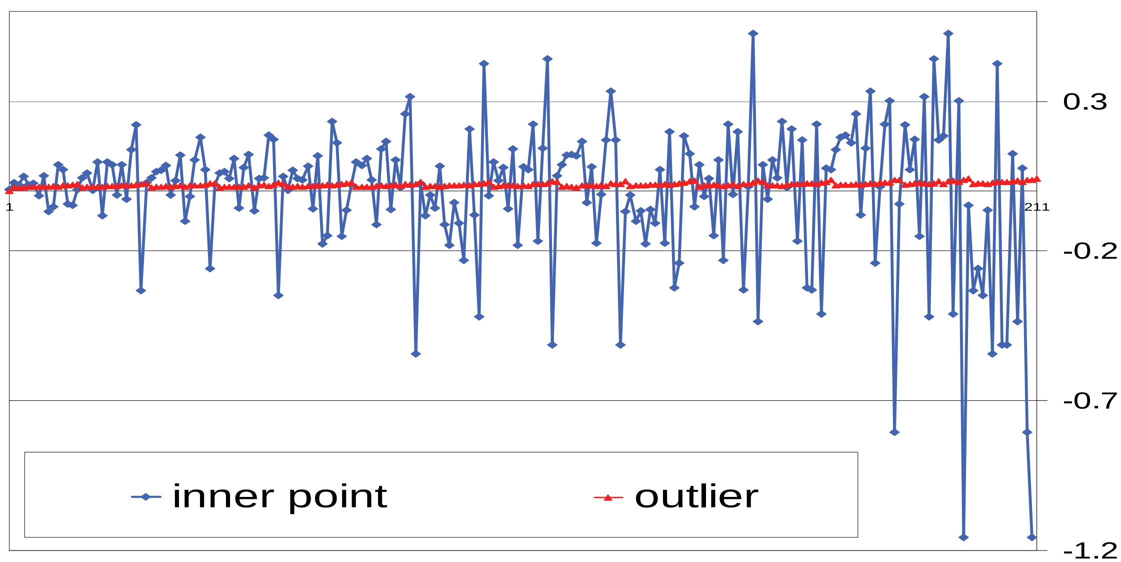




High-dimensional Approaches

- 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





High-dimensional Approaches

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

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

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

High-dimensional Approaches

- 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

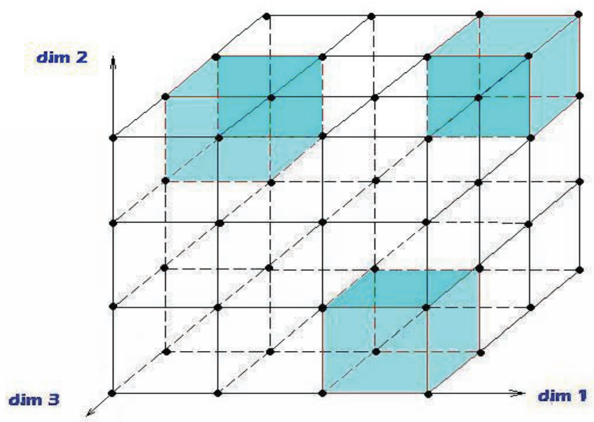
High-dimensional Approaches

- Grid-based subspace outlier detection [Aggarwal and Yu 2000]
  - Model
    - Partition data space by an equi-depth grid (Φ = number of cells in each dimension)
    - Sparsity coefficient *S(C)* for a *k*-dimensional grid cell *C*

$$S(C) = \frac{count(C) - n \cdot (1/\Phi)^k}{\sqrt{n \cdot (1/\Phi)^k \cdot (1 - (1/\Phi)^k)}}$$

where *count*(*C*) is the number of data objects in C

- *S(C)* < 0 => *count*(*C*) is lower than expected
- Outliers are those objects that are located in lower-dimensional cells with negative sparsity coefficient

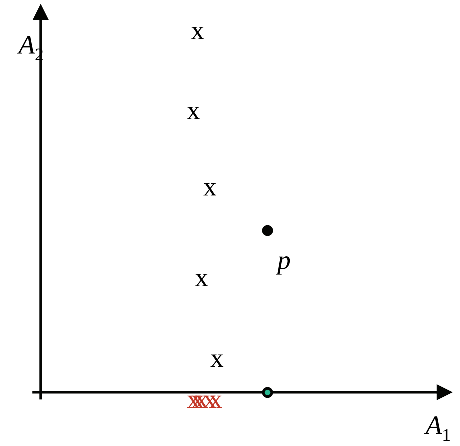


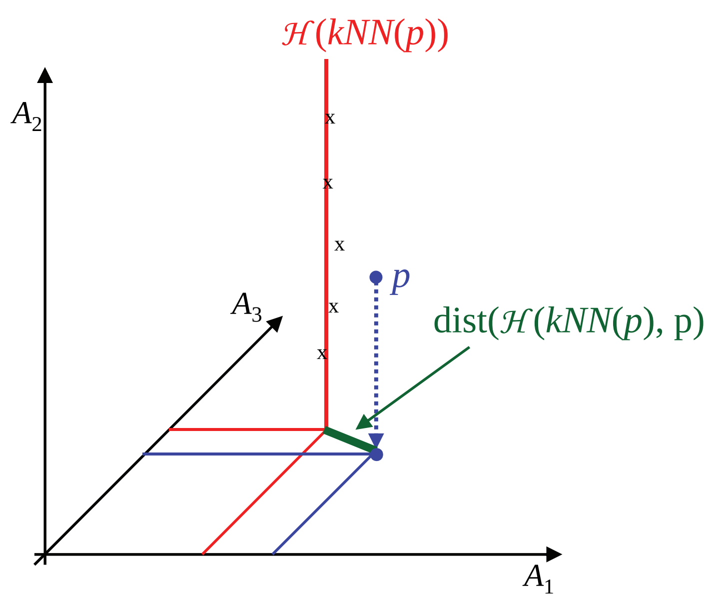
$\Phi = 3$

High-dimensional Approaches

- Algorithm
  - Find the *m* grid cells (projections) with the lowest sparsity coefficients
  - Brute-force algorithm is in $O(\Phi^d)$
  - Evolutionary algorithm (input: *m* and the dimensionality of the cells)
- Discussion
  - Results need not be the points from the optimal cells
  - Very coarse model (all objects that are in cell with less points than to be expected)
  - Quality depends on grid resolution and grid position
  - Outputs a labeling
  - Implements a global approach (key criterion: globally expected number of points within a cell)

- SOD – subspace outlier degree [Kriegel et al. 2009]
  - Motivation
    - Outliers may be visible only in subspaces of the original data
  - Model
    - Compute the subspace in which the *k*NNs of a point *p* minimize the variance
    - Compute the hyperplane $\mathcal{H}(kNN(p))$ that is orthogonal to that subspace
    - Take the distance of *p* to the hyperplane as measure for its "outlierness"





High-dimensional Approaches

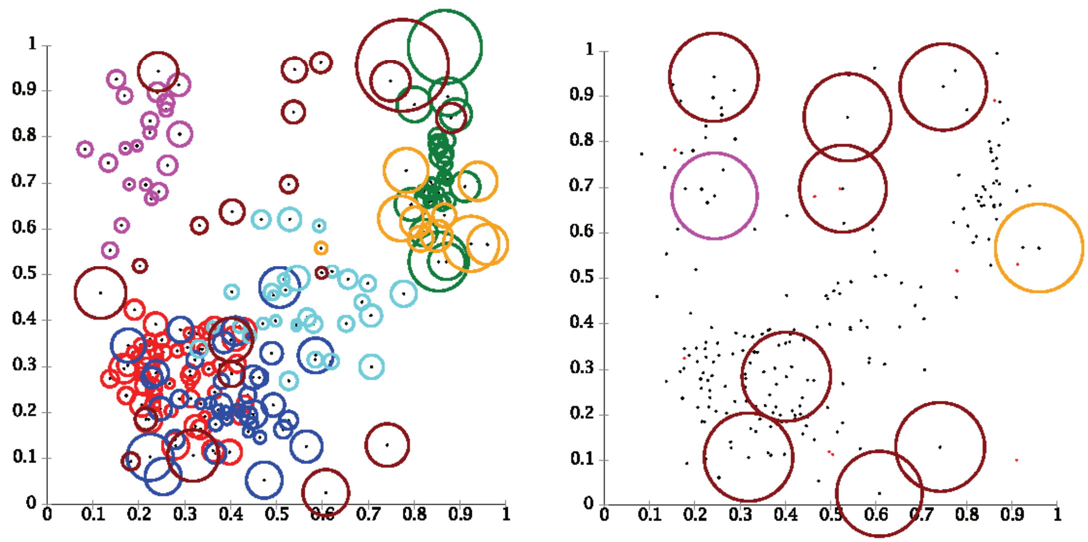
– Discussion
  - Assumes that *k*NNs of outliers have a lower-dimensional projection with small variance
  - Resolution is local (can be adjusted by the user via the parameter *k*)
  - Output is a scoring (SOD value)

Outline

Summary

- Summary
  - Different models are based on different assumptions to model outliers
  - Different models provide different types of output (labeling/scoring)
  - Different models consider outlier at different resolutions (global/local)
  - Thus, different models will produce different results
  - A thorough and comprehensive comparison between different models and approaches is still missing

- Outlook
  - Experimental evaluation of different approaches to understand and compare differences and common properties
  - A first step towards unification of the diverse approaches: providing density-based outlier scores as probability values [Kriegel et al. 2009a]: judging the deviation of the outlier score from the expected value
  - Visualization [Achtert et al. 2010]
  - New models
  - Performance issues
  - Complex data types
  - High-dimensional data
  - …



Outline

List of References

DATABASE SYSTEMS GROUP

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.

Aggarwal, C.C. and Yu, P.S. 2000. Outlier detection for high dimensional data. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

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. 2009. Outlier detection in axis-parallel subspaces of high dimensional data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand.

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., 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.

Pei, Y., Zaiane, O., and Gao, Y. 2006. An efficient reference-based approach to outlier detection in large datasets. In Proc. 6th Int. Conf. on Data Mining (ICDM), Hong Kong, China.

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

Literature

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.

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.

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.