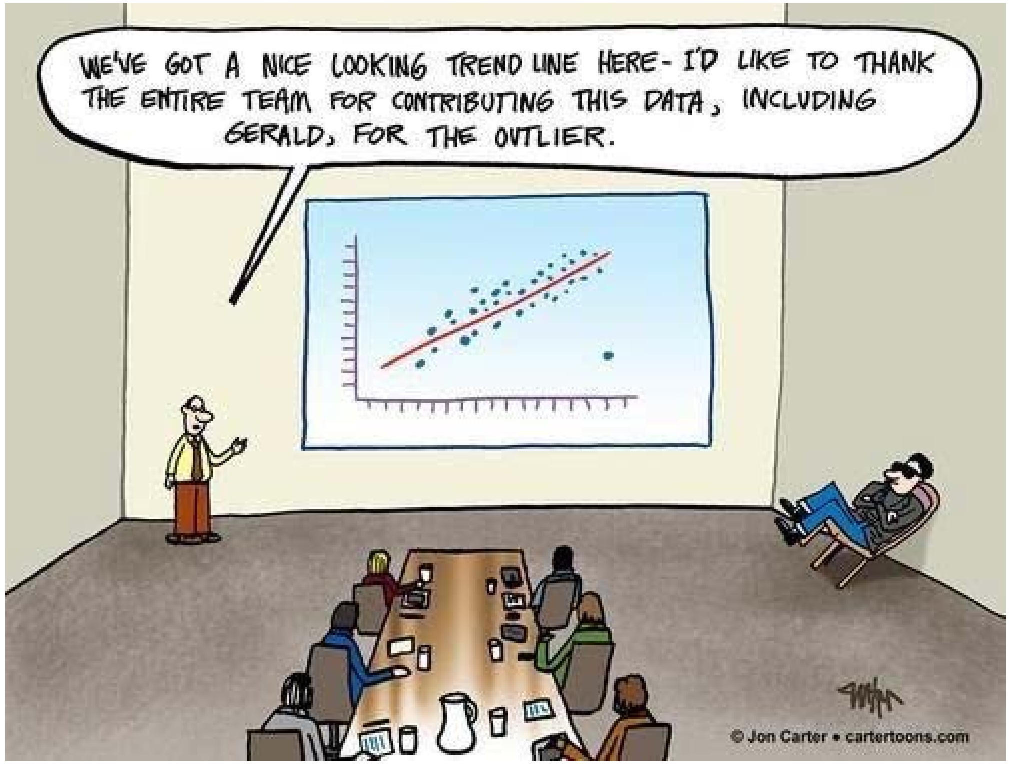
Outlier Detection

Xike Xie


WE'VE GOT A NICE LOOKING TREND LINE HERE- I'D LIKE TO THANK THE ENTIRE TEAM FOR CONTRIBUTING THIS DATA, INCLUDING GERALD, FOR THE OUTLIER.
© Jon Carter • cartertoons.com

Slides are based on Prof. Hans-Peter Kriegel's work

Introduction

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”

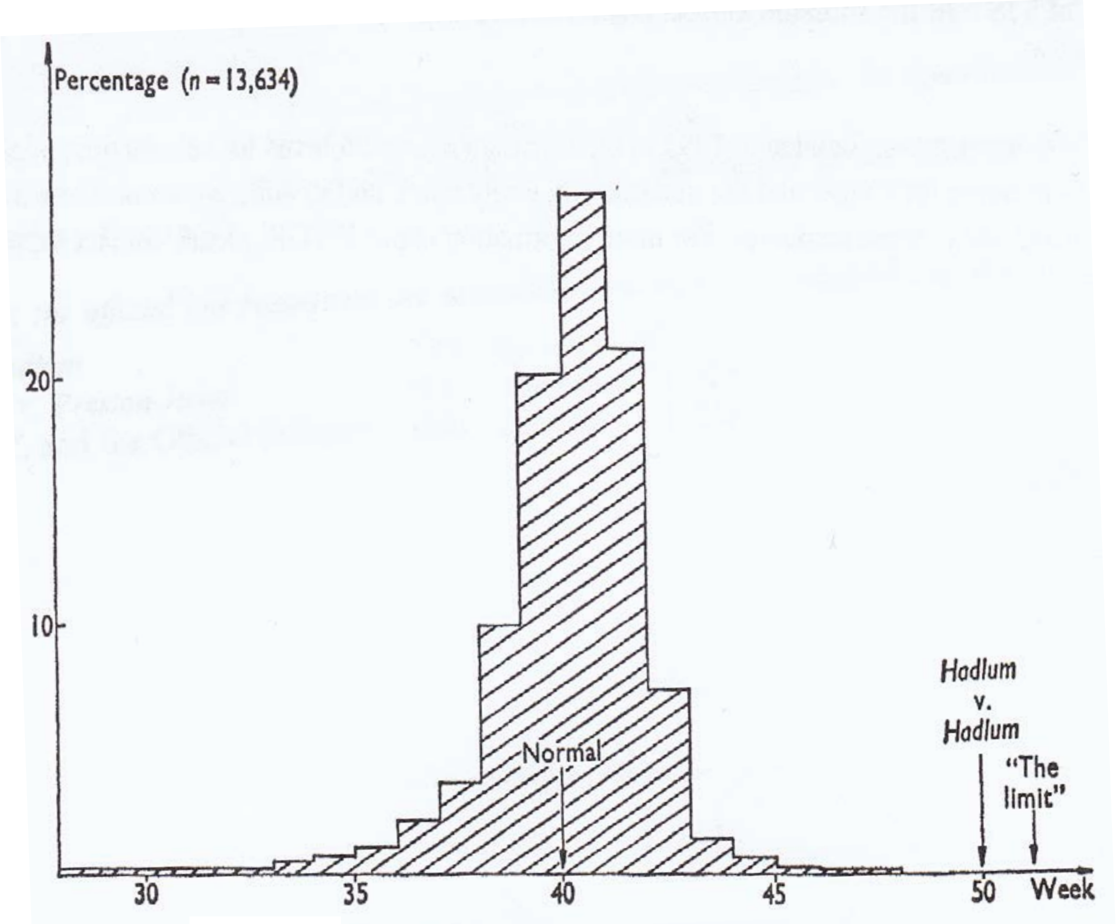
Statistics-based intuition

- 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]

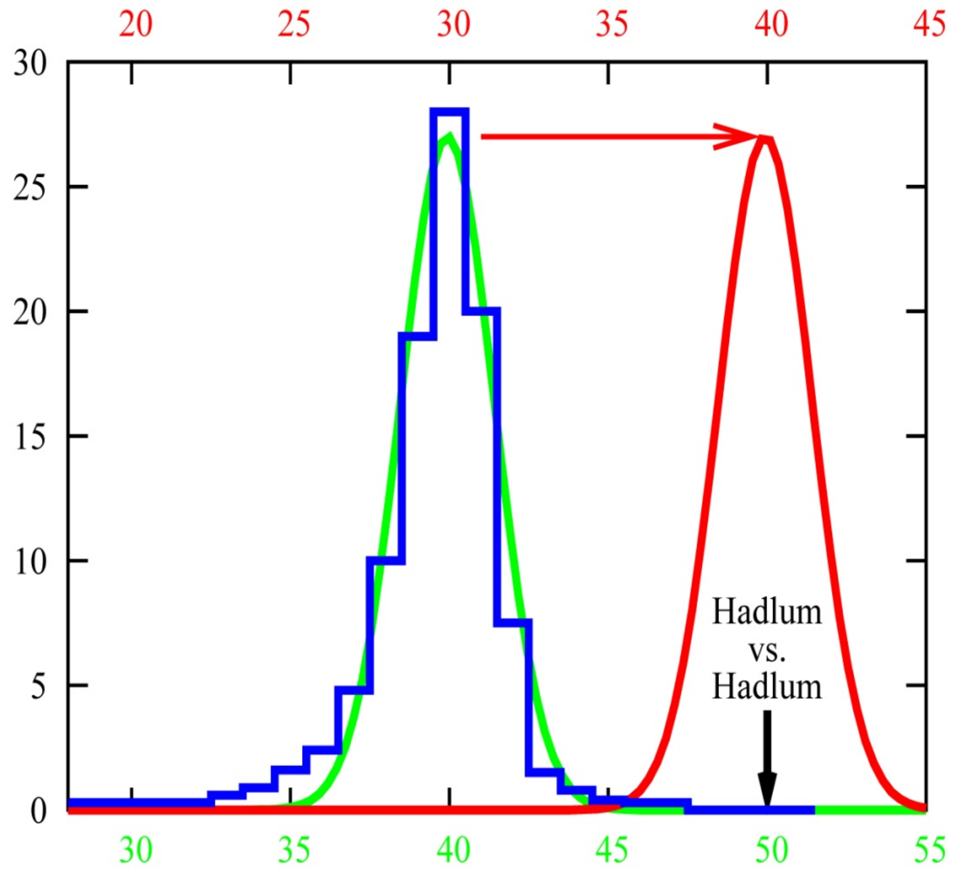
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.



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 starts later)
    - Under this assumption the gestation period has an average duration and the specific birthday has 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.

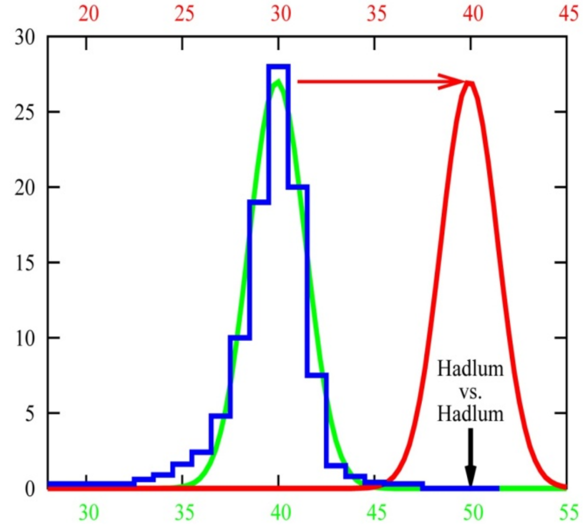
Introduction

- 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, within a specific range

EAST
WEST
KIA
PLAYER OF THE MONTH
GIANNIS ANTETOKOUNMPO
30.9 PTS 13.4 REB 5.9 AST 1.5 STL 17-3 W-L
LUKA DONCIC
30.8 PTS 9.9 REB 9.6 AST 1.4 STL 12-6 W-L
@NBA

Introduction

- Discussion of the basic intuition based on Hawki
  - 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 “normal” 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



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 imbalanced
  - 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 lecture, 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

    - 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

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

Model-based

Proximity-based

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)
  - See e.g. Barnett’s discussion of Hadlum vs. Hadlum

- 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

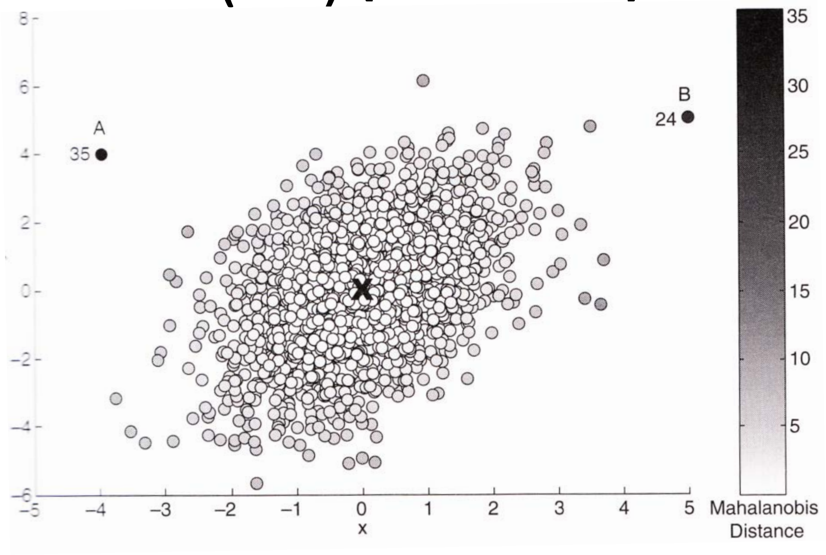
- Probability density function of a multivariate normal distribution

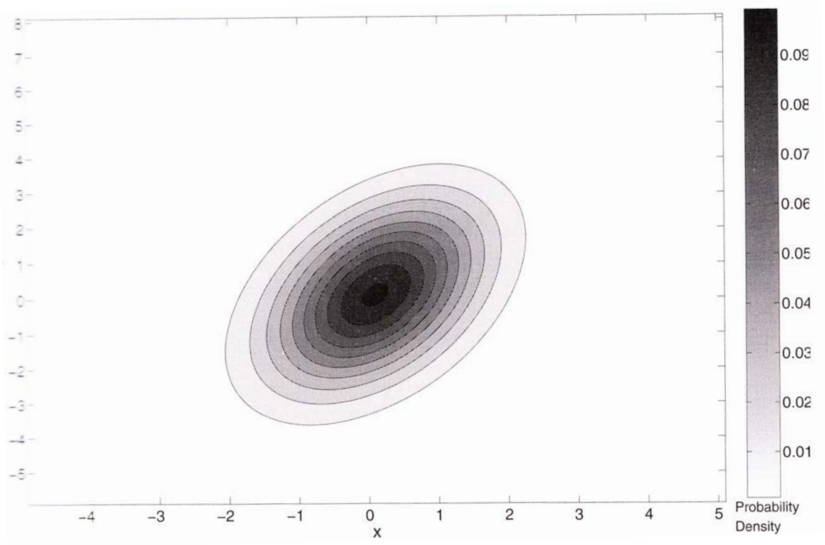
$$N(x) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}{2}}$$

  - 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) > χ^2(0,975) [$\approx 3\cdot\sigma$]

Statistical Tests

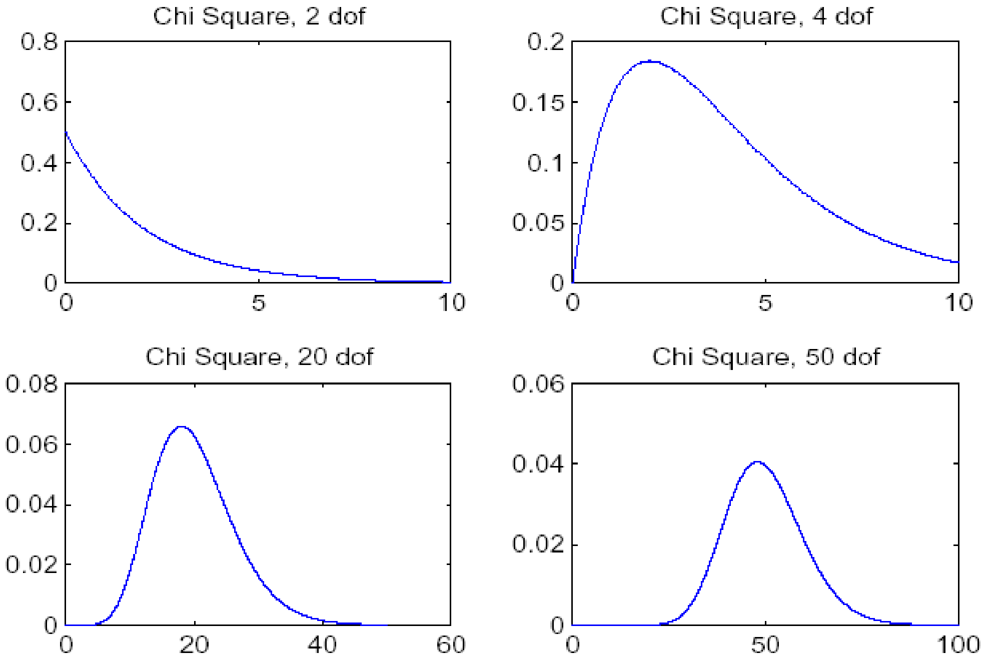
- Visualization (2D) [Tan et al. 2006]





Statistical Tests

- Problems
  - Curse of dimensionality
    - The la *MDist* values for all points



x-axis: observed *MDist* values

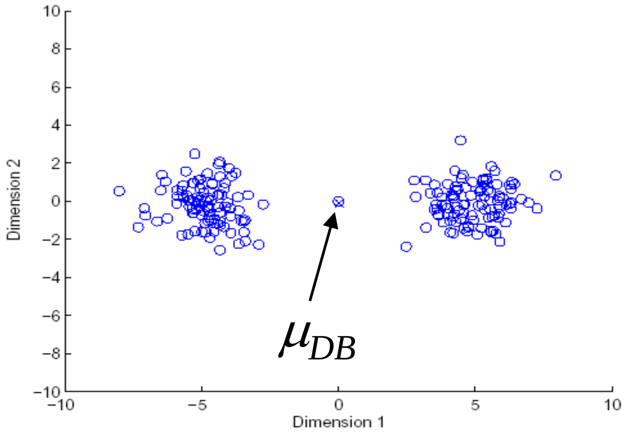
y-axis: frequency of observation

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

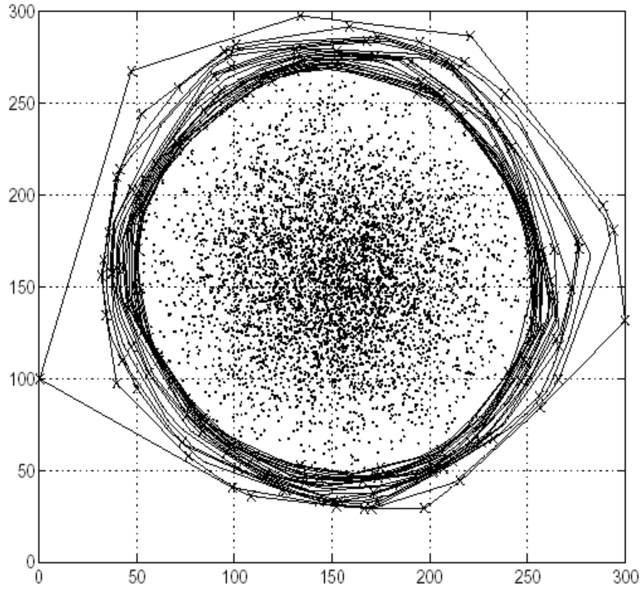


Outline

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

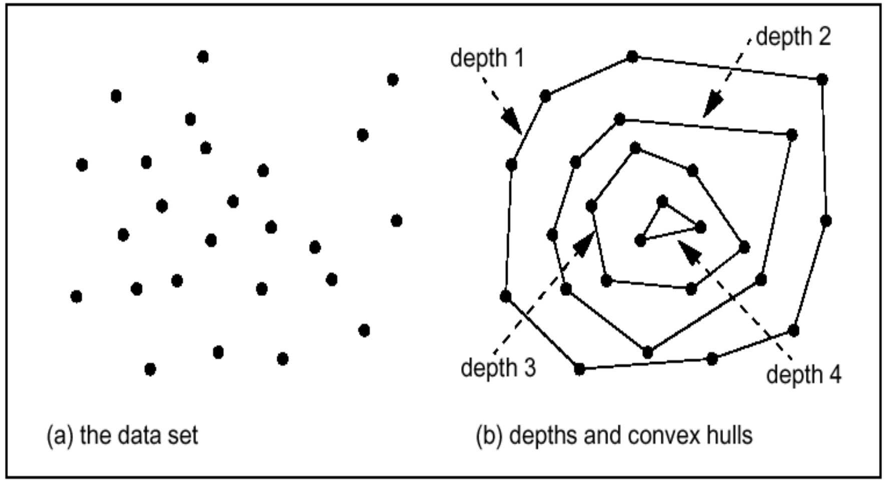


Picture taken from [Johnson et al. 1998]

- 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 (e.g. take depth as scoring value)
  - Uses a global reference set for outlier detection

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

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*
  - If two sets have an equal *SF* value, take the smaller set
  - The outliers are the elements of the **exception set** $E \subseteq DB$ for which the following holds:

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

- 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

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

Distance-based Approaches

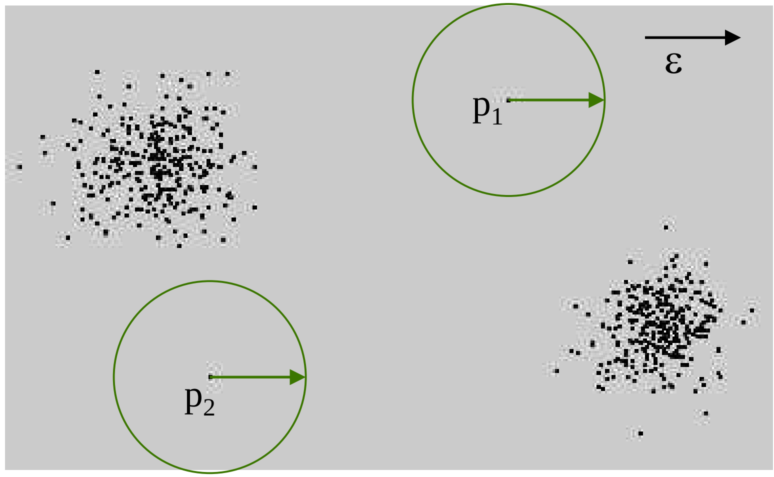
- General Idea
  - Judge a point based on the distance(s) to its neighbors
  - Several variants proposed

- 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

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



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

- 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 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*)

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 $q \in k\mathrm{NN}(p)$ 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

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

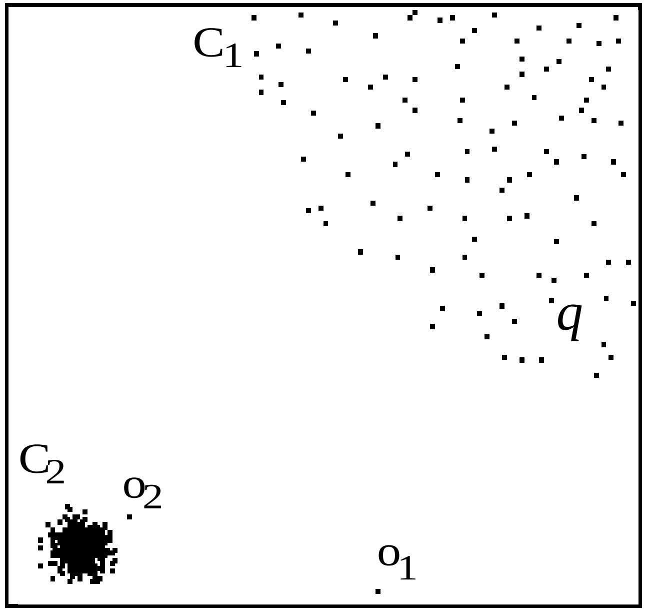
Density-based Approaches

- 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 essentially differ in how to estimate density

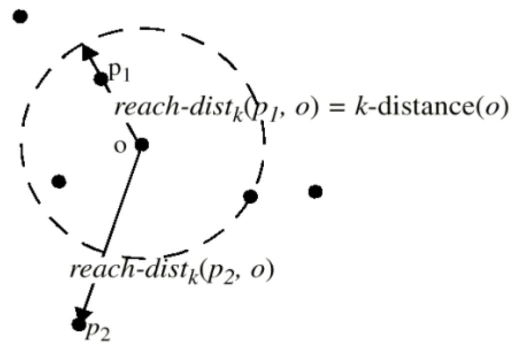
- Basic assumption
  - The density around a normal data object is similar to the density around its neighbors
  - The density around an outlier is considerably different to the density around its neighbors

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



Density-based Approaches



- 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*

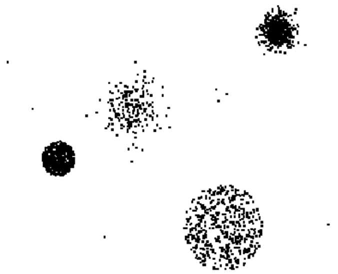
$$lrd_k(p) = 1 / \left(\frac{\sum_{o \in kNN(p)} reach{-}dist_k(p,o)}{Card\big(kNN(p)\big)} \right)$$

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

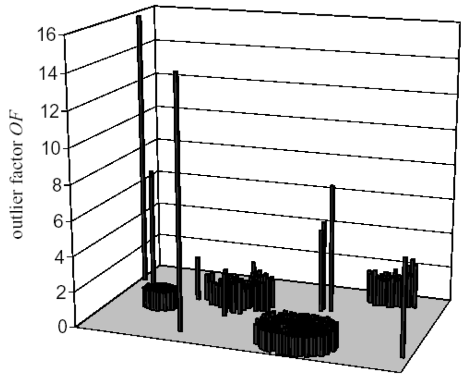
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card\big(kNN(p)\big)}$$

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)

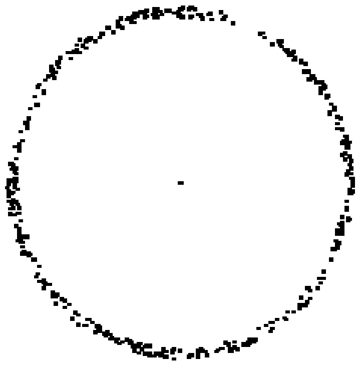
- Discussion
  - Choice of *k* (*MinPts* in the original paper) specifies the reference set
  - Originally implements a local approach (resolution depends on the user's choice for *k*)
  - Outputs a scoring (assigns an LOF value to each point)

Density-based Approaches

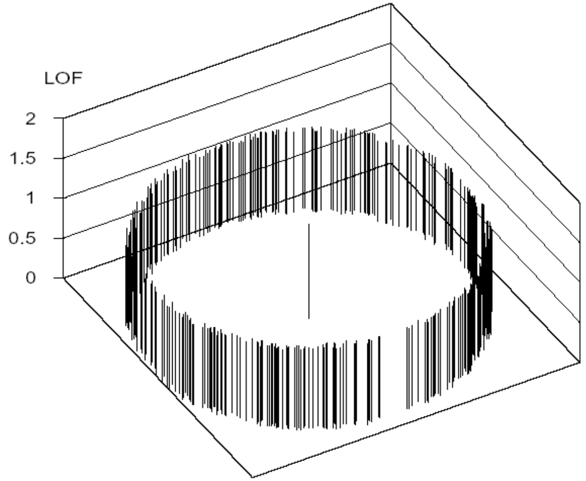
- Variants of LOF
  - Mining top-*n* local outliers [Jin et al. 2001]
    - Idea:
      - Usually, a user is only interested in the top-*n* outliers
      - Do not compute the LOF for all data objects => save runtime
    - Method
      - Compress data points into micro clusters using the CFs of BIRCH [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

Density-based Approaches

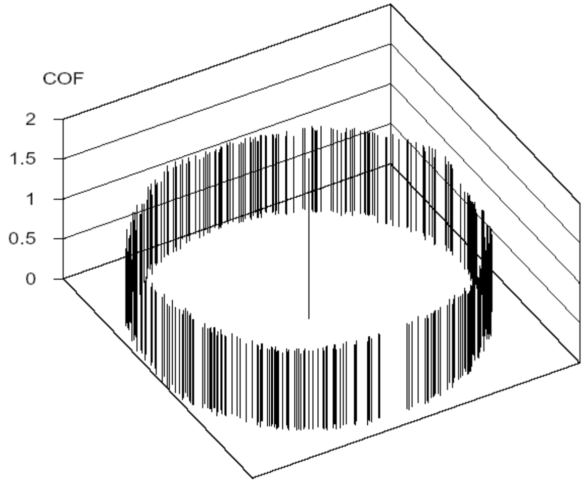
- Variants of LOF (cont.)
  - Connectivity-based outlier factor (COF) [Tang et al. 2002]
    - Motivation
      - In regions of low density, it may be hard to detect outliers
      - Choose a low value for *k* is often not appropriate
    - Solution
      - Treat “low density” and “isolation” differently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



Data set



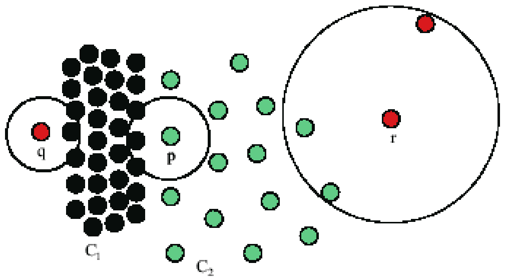
LOF



COF

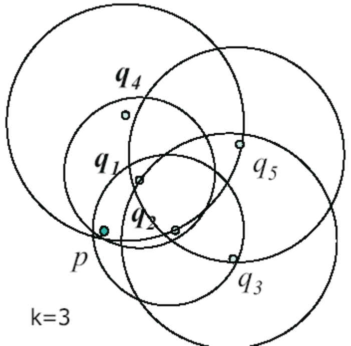
Density-based Approaches

- 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 (*k*IS(*p*)) of a point *p* includes its *k*NNs (*k*NN(*p*)) and its reverse *k*NNs (R*k*NN(*p*))



*k*IS(*p*) = kNN(p) ∪ R*k*NN(*p*))

$= \{q_1, q_2, q_3\} \cup \{q_1, q_2, q_4\}$

$= \{q_1, q_2, q_3, q_4\}$

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 $\approx$ 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*)

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

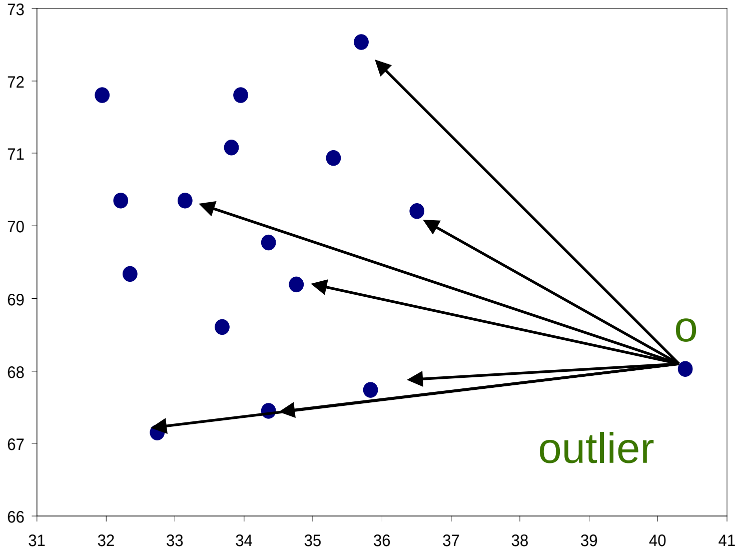
- Motivation
    - One sample class of adaptions of existing models to a specific problem (high dimensional data)
    - Why is that problem important?
        - Some (ten) years ago:
            - Data recording was expensive
            - Variables (attributes) where carefully evaluated if they are relevant for the analysis task
            - Data sets usually contain only a few number of relevant dimensions
        - Nowadays:
            - Data recording is easy and cheap
            - “Everyone measures everything”, attributes are not evaluated just measured
            - Data sets usually contain a large number of features
                - Molecular biology: gene expression data with >1,000 of genes per patient
                - Customer recommendation: ratings of 10-100 of products per person
                - …

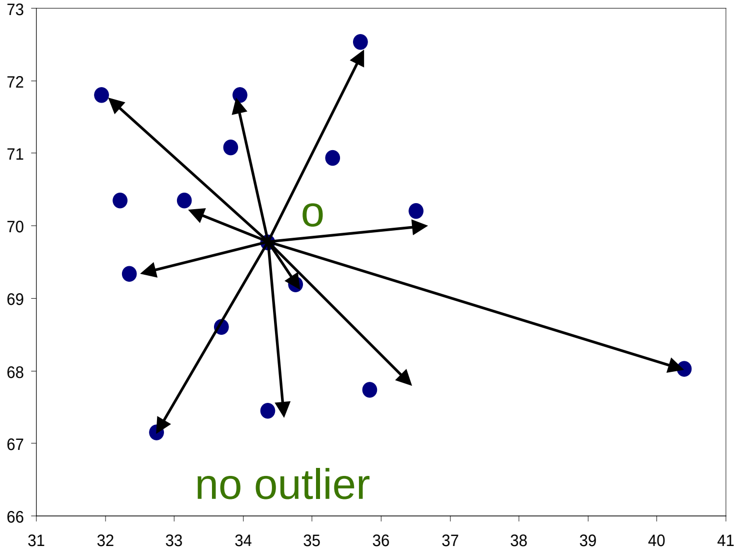
High-dimensional Approaches

- Challenges
    - Curse of dimensionality
        - Relative contrast between distances decreases with increasing dimensionality
        - Data are 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

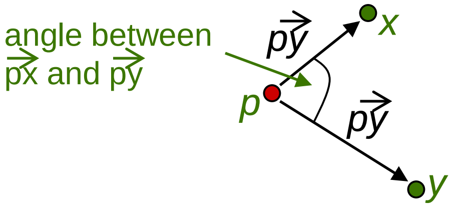
- ABOD – angle-based outlier degree [Kriegel et al. 2008]
  - Rational
    - Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
    - Object o is an outlier if most other objects are located in similar directions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

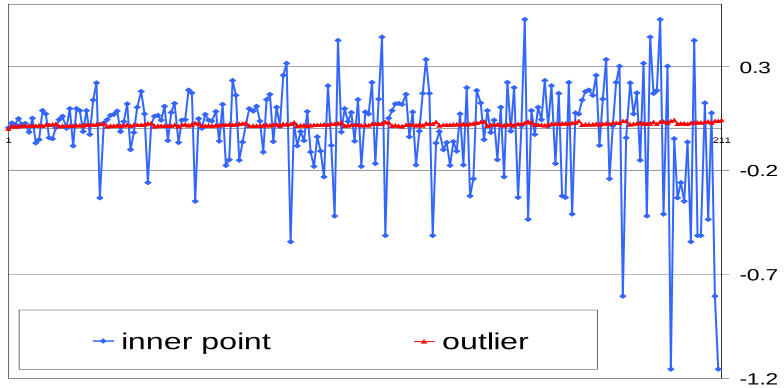




High-dimensional Approaches

- 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 *px* and *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 (inversely scaled: high ABOD => inlier, low ABOD => outlier)

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

Summary

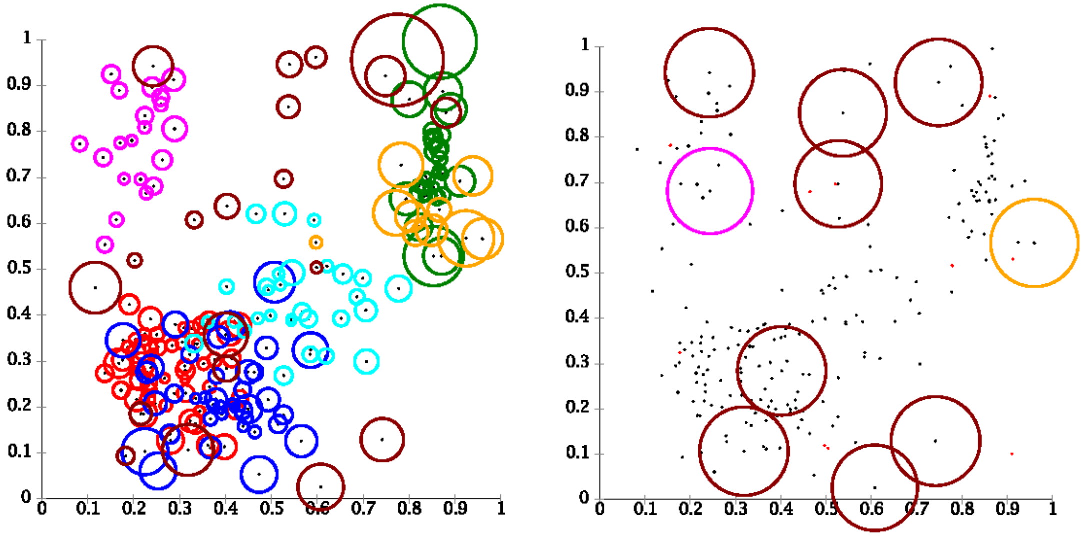
- Summary
    - Historical evolution of outlier detection methods
        - Statistical tests
            - Limited (univariate, no mixture model, outliers are rare)
            - No emphasis on computational time
        - Extensions to these tests
            - Multivariate, mixture models, …
            - Still no emphasis on computational time
        - Database-driven approaches
            - First, still statistically driven intuition of outliers
            - Emphasis on computational complexity
        - Database and data mining approaches
            - Spatial intuition of outliers
            - Even stronger focus on computational complexity
              (e.g. invention of top-k problem to propose new efficient algorithms)

Summary

- Consequence
    - 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

Summary

- 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

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 √