Outlier Detection in High-Dimensional Data

Tutorial

Arthur Zimek[1], Erich Schubert[2], Hans-Peter Kriegel[2]

[1]University of Alberta
Edmonton, AB, Canada

[2]Ludwig-Maximilians-Universität München
Munich, Germany

ICDM 2012, Brussels, Belgium

LMU

UNIVERSITY OF ALBERTA
FACULTY OF SCIENCE

Coverage and Objective of the Tutorial

- We assume that you know in general what outlier detection is about and have a rough idea of how classic approaches (e.g., LOF [Breunig et al., 2000]) work.
- We focus on unsupervised methods for numerical vector data (Euclidean space).
- We discuss the specific problems in high-dimensional data.
- We discuss strategies as well as the strengths and weaknesses of methods that specialize in high-dimensional data in order to
  - enhance efficiency or effectiveness and stability
  - search for outliers in subspaces

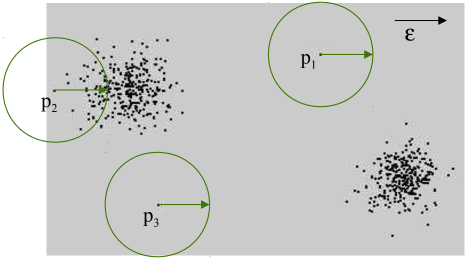
Coverage and Objective of the Tutorial

- These slides are available at:
  `http://www.dbs.ifi.lmu.de/cms/Publications/OutlierHighDimensional`
- This tutorial is closely related to our survey article [Zimek et al., 2012], where you find more details.
- Feel free to ask questions at any time!

Reminder: Distance-based Outliers

DB(ε, π)-outlier [Knorr and Ng, 1997]

- given ε, 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

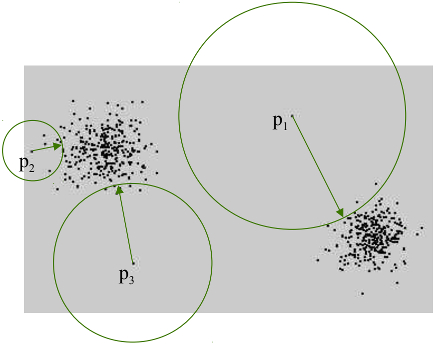


$$OutlierSet(\varepsilon, \pi) = \left\{ p \,\middle|\, \frac{Cardinality(q \in \mathcal{DB} | dist(q,p) < \varepsilon)}{Cardinality(\mathcal{DB})} \leq \pi \right\}$$

Reminder: Distance-based Outliers

Outlier scoring based on kNN distances:

- Take the kNN distance of a point as its outlier score [Ramaswamy et al., 2000]
- Aggregate the distances for the 1-NN, 2-NN, ..., kNN (sum, average) [Angiulli and Pizzuti, 2002]



Reminder: Density-based Local Outliers

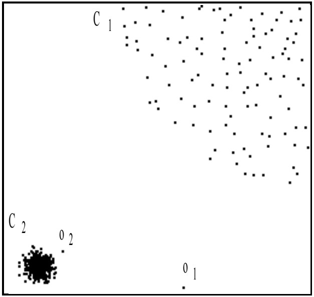


Figure from Breunig et al. [2000].

- DB-outlier model: no parameters ε, π such that o_2 is an outlier but none of the points of C_1 is an outlier
- kNN-outlier model: kNN-distances of points in C_1 are larger than kNN-distances of o_2

Reminder: Density-based Local Outliers

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

- reachability distance (smoothing factor):
  $reachdist_k(p,o) = \max\{kdist(o), dist(p,o)\}$
- local reachability distance (lrd)
  $lrd_k(p) = 1/\frac{\sum_{o \in kNN(p)} reachdist_k(p,o)}{Cardinality(kNN(p))}$
- Local outlier factor (LOF) of point p: average ratio of lrds of neighbors of p and lrd of p

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

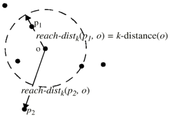


Figure from [Breunig et al., 2000]

- LOF ≈ 1: homogeneous density
- LOF $\gg 1$: point is an outlier (meaning of "$\gg$" ?)

Outline

Introduction

The "Curse of Dimensionality"

Efficiency and Effectiveness

Subspace Outlier Detection

Discussion and Conclusion

Outline

Introduction

The "Curse of Dimensionality"

- Concentration of Distances and of Outlier Scores
- Relevant and Irrelevant Attributes
- Discrimination vs. Ranking of Values
- Combinatorial Issues and Subspace Selection
- Hubness
- Consequences

Efficiency and Effectiveness

Subspace Outlier Detection

Outline

Concentration of Distances

Theorem 1 (Beyer et al. [1999])

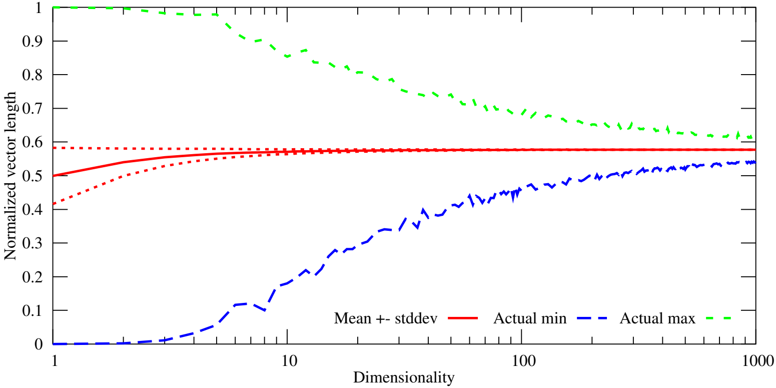
Assumption *The ratio of the variance of the length of any point vector (denoted by $\|X_d\|$) with the length of the mean point vector (denoted by $E[\|X_d\|]$) converges to zero with increasing data dimensionality.*

Consequence *The proportional difference between the farthest-point distance D_{max} and the closest-point distance D_{min} (*the relative contrast*) vanishes.*

$$\textit{If} \lim_{d\to\infty} \textit{var}\left(\frac{\|X_d\|}{E[\|X_d\|]}\right) = 0, \textit{ then } \frac{D_{max} - D_{min}}{D_{min}} \to 0.$$

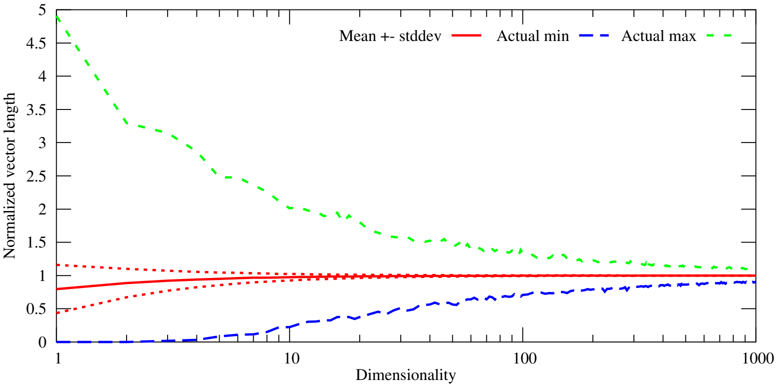
Vector Length: Loss of Contrast

Sample of 10^5 instances drawn from a uniform $[0, 1]$ distribution, normalized ($1/\sqrt{d}$).



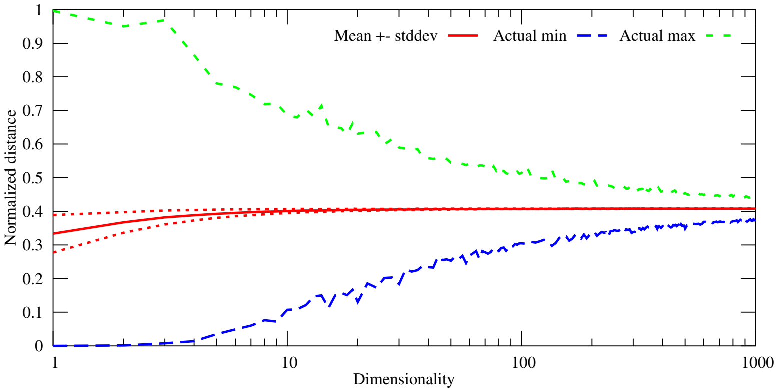
Vector Length: Loss of Contrast

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution, normalized ($1/\sqrt{d}$).



Pairwise Distances

Sample of 10^5 instances drawn from a uniform $[0, 1]$ distribution, normalized ($1/\sqrt{d}$).



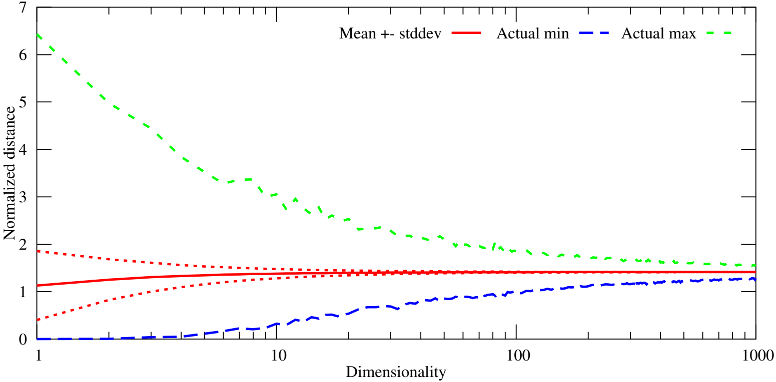
UNIVERSITY OF ALBERTA
FACULTY OF SCIENCE

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Pairwise Distances

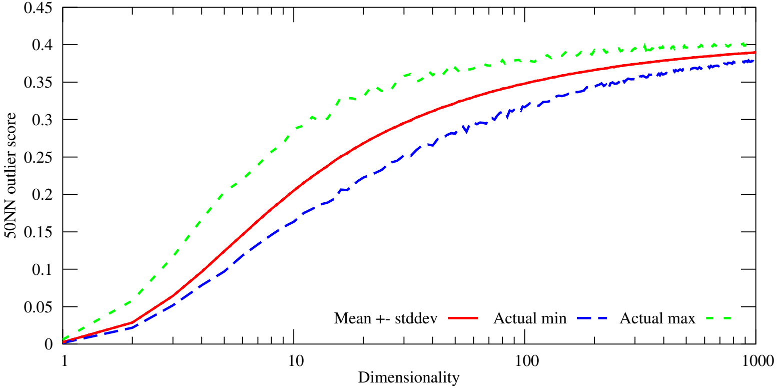
A. Zimek, E. Schubert, H.-P. Kriegel

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution, normalized $(1/\sqrt{d})$.



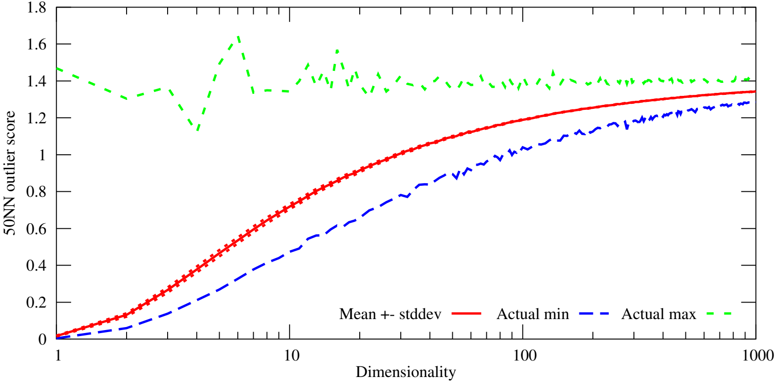
50-NN Outlier Score

Sample of 10^5 instances drawn from a uniform $[0,1]$ distribution, normalized ($1/\sqrt{d}$). kNN outlier score [Ramaswamy et al., 2000] for $k = 50$.



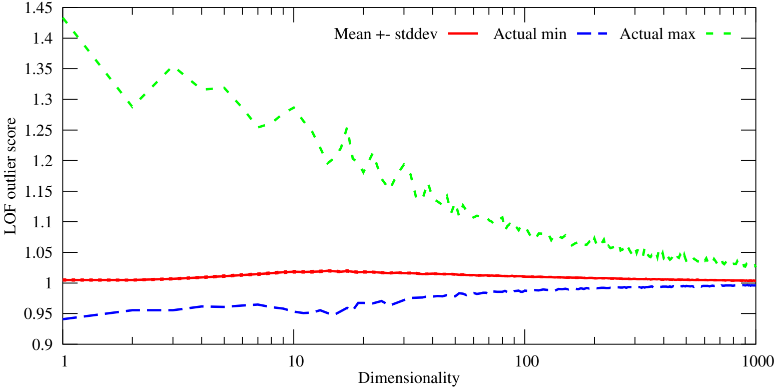
50-NN Outlier Score

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution, normalized $(1/\sqrt{d})$. kNN outlier score [Ramaswamy et al., 2000] for $k = 50$.



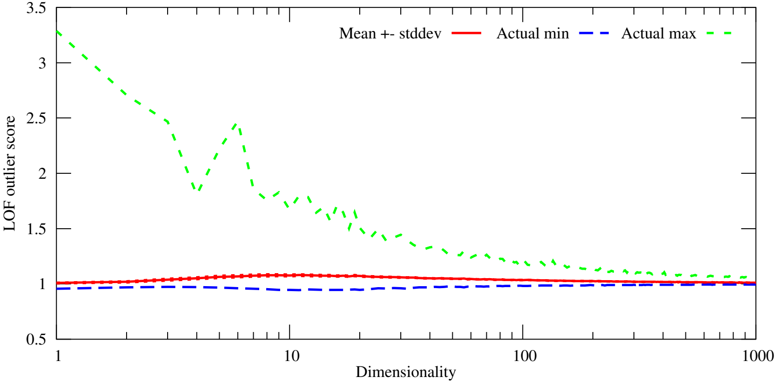
LOF Outlier Score

Sample of 10^5 instances drawn from a uniform $[0, 1]$ distribution, LOF [Breunig et al., 2000] score for neighborhood size 50.



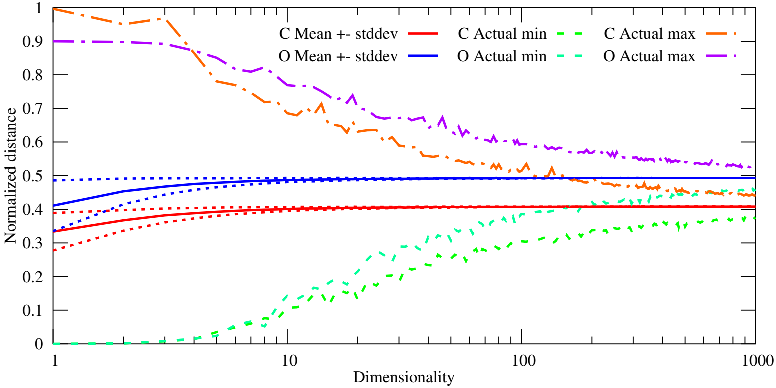
LOF Outlier Score

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution, LOF [Breunig et al., 2000] score for neighborhood size 50.



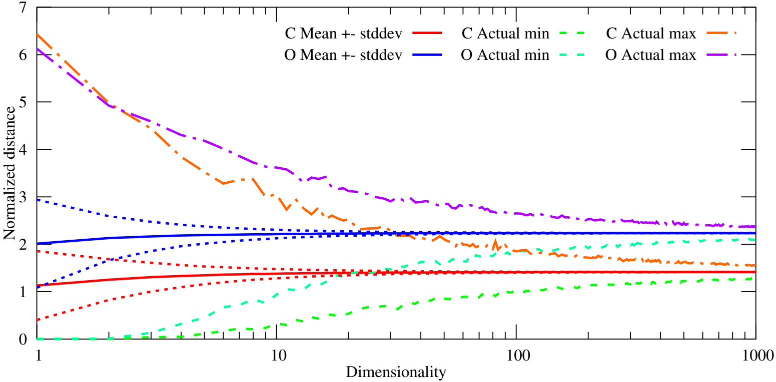
Pairwise Distances with Outlier

Sample of 10^5 instances drawn from a uniform $[0,1]$ distribution, normalized $(1/\sqrt{d})$.
Outlier manually placed at 0.9 in every dimension.



Pairwise Distances with Outlier

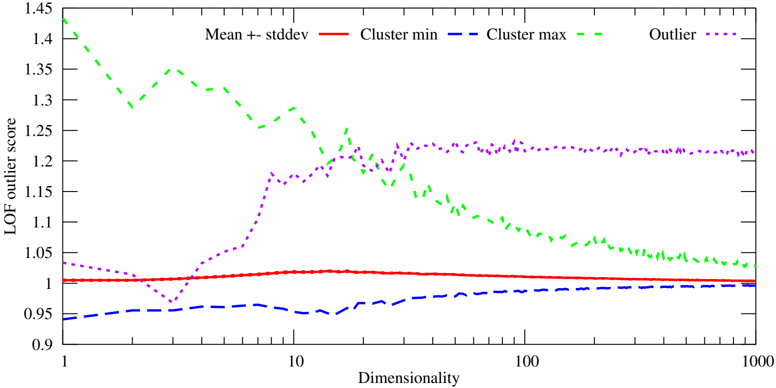
Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution, normalized $(1/\sqrt{d})$.
Outlier manually placed at 2σ in every dimension.



LOF Outlier Score with Outlier

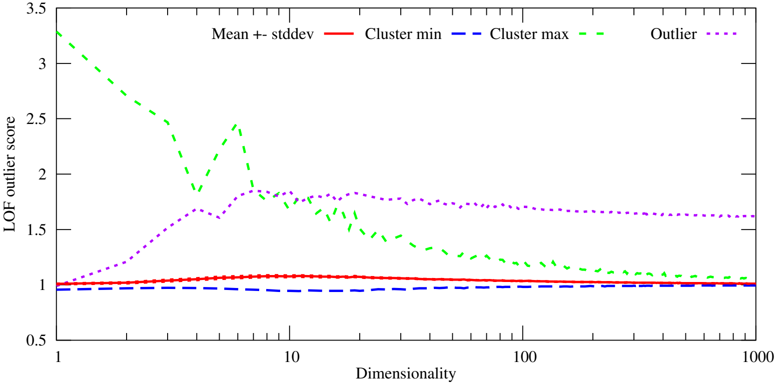
Sample of 10^5 instances drawn from a uniform $[0, 1]$ distribution.

Outlier manually placed at 0.9 in every dimension, LOF [Breunig et al., 2000] score for neighborhood size 50.



LOF Outlier Score with Outlier

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution.
Outlier manually placed at 2σ in every dimension, LOF [Breunig et al., 2000] score for neighborhood size 50.



Conclusion

- The concentration effect *per se* is *not* the main problem for mining high-dimensional data.
- If points deviate in every attribute from the usual data distribution, the outlier characteristics will become more pronounced with increasing dimensionality.
- More dimensions add more information for discriminating the different characteristics.

Outline

Introduction

The "Curse of Dimensionality"

- Concentration of Distances and of Outlier Scores
- **Relevant and Irrelevant Attributes**
- Discrimination vs. Ranking of Values
- Combinatorial Issues and Subspace Selection
- Hubness
- Consequences

Efficiency and Effectiveness

Subspace Outlier Detection

Separation of Clusters – “Meaningful” Nearest Neighbors

Theorem 2 (Bennett et al. [1999])

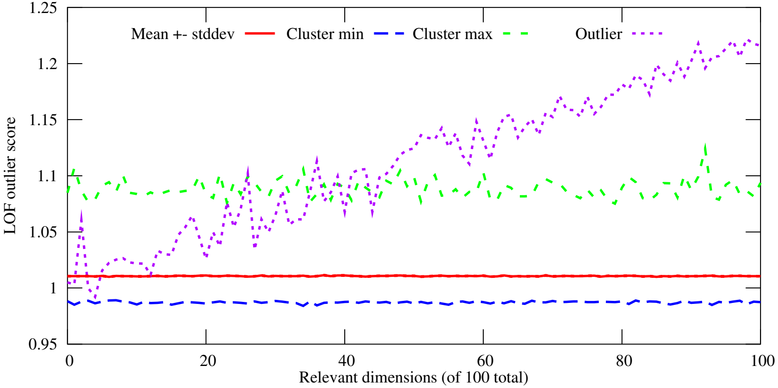
Assumption *Two clusters are pairwise stable, i.e., the between cluster distance dominates the within cluster distance.*

Consequence *We can meaningfully discern “near” neighbors (members of the same cluster) from “far” neighbors (members of the other cluster).*

- This is the case if enough information (relevant attributes) is provided to separate different distributions.
- Irrelevant attributes can mask the separation of clusters or outliers.

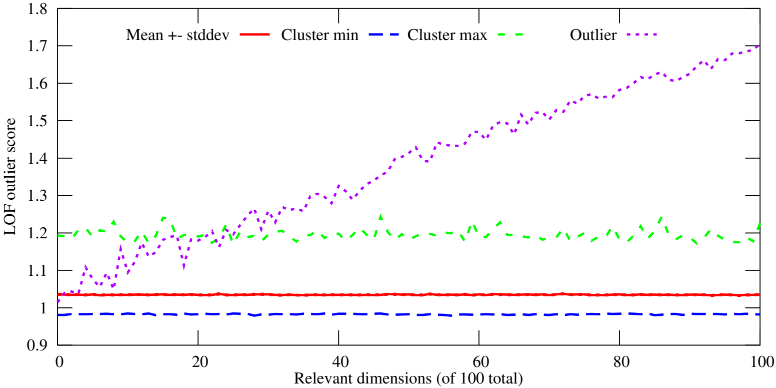
Relevant and Irrelevant Attributes

Sample of 10^5 instances drawn from a uniform $[0, 1]$ distribution.
Fixed dimensionality $d = 100$.
Outlier manually placed at 0.9 in relevant dimensions, in irrelevant dimensions the attribute values for the outlier are drawn from the usual random distribution.



Relevant and Irrelevant Attributes

Sample of 10^5 instances drawn from a Gaussian $(0, 1)$ distribution.
Fixed dimensionality $d = 100$.
Outlier manually placed at 2σ in relevant dimensions, in irrelevant dimensions the attribute values for the outlier are drawn from the usual random distribution.



Conclusion

- Motivation for subspace outlier detection: find outliers in the relevant subspaces
- Challenge of identifying relevant attributes
- even more: *different* attributes may be relevant for identifying *different* outliers

Outline

Introduction

The "Curse of Dimensionality"

- Concentration of Distances and of Outlier Scores
- Relevant and Irrelevant Attributes
- **Discrimination vs. Ranking of Values**
- Combinatorial Issues and Subspace Selection
- Hubness
- Consequences

Efficiency and Effectiveness

Subspace Outlier Detection

Discrimination of Distance Values

- distance concentration: low contrast of distance values of points from the same distribution
- other side of the coin: hard to choose a distance threshold to distinguish between near and far points (e.g. for distance queries)

Illustration: “Shrinking” (?) Hyperspheres

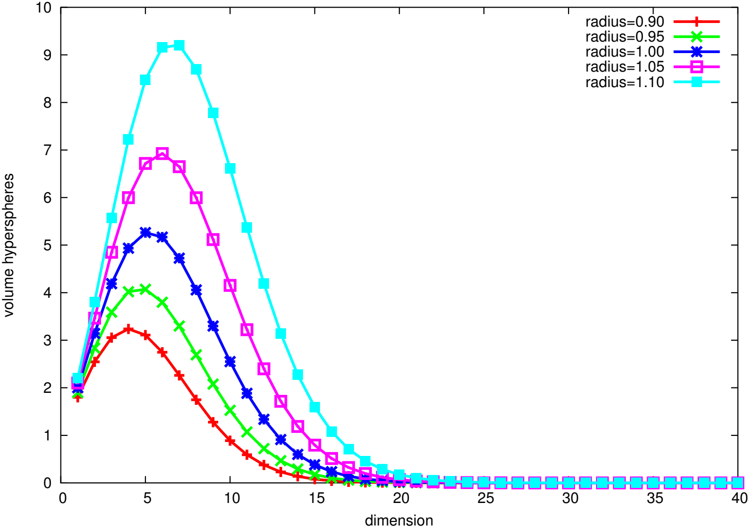
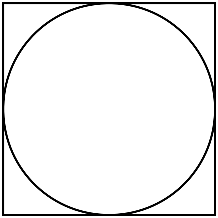


Illustration: “Shrinking” (?) Hyperspheres



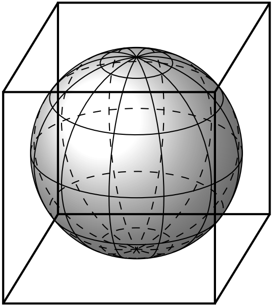
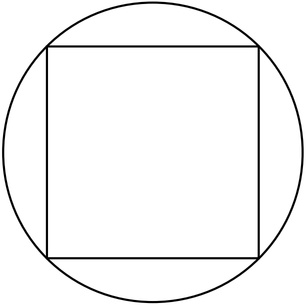


Illustration: “Shrinking” (?) Hyperspheres



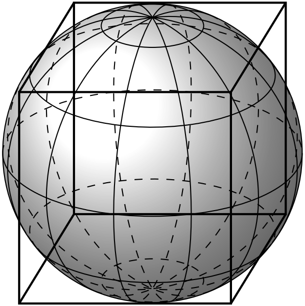


Illustration: “Shrinking” (?) Hyperspheres

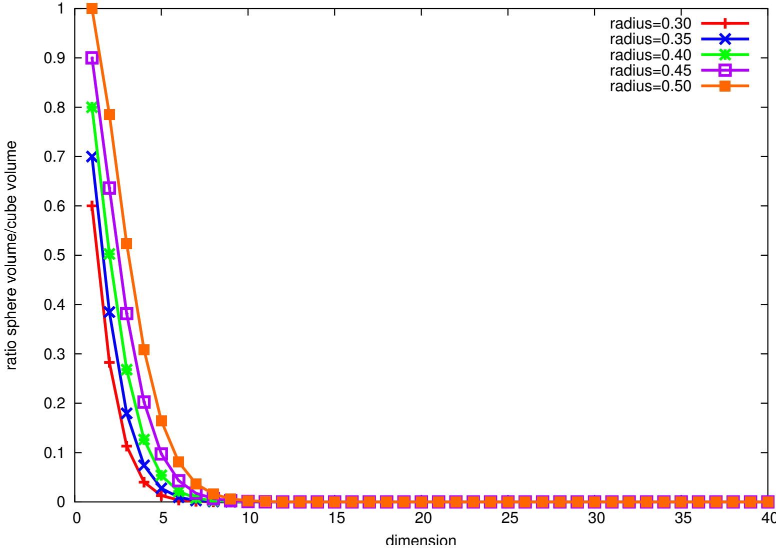
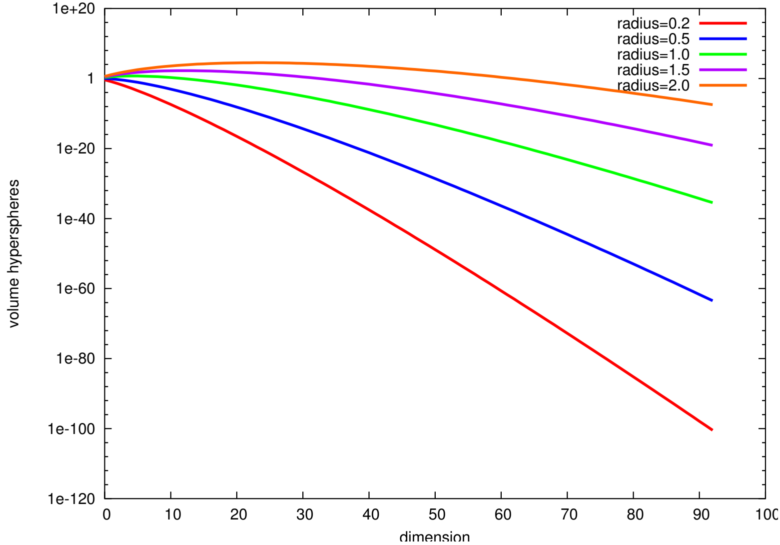


Illustration: “Shrinking” (?) Hyperspheres



Meaningful Choice of Distance Thresholds?

- distance values are not comparable over data (sub-)spaces of different dimensionality
- ε-range queries for high-dimensional data are hard to parameterize
- some change of ε may have no effect in some dimensionality and may decide whether nothing or everything is retrieved in some other dimensionality
- density-thresholds are in the same way notoriously sensitive to dimensionality

Distance Rankings – "Meaningful" Nearest Neighbors

- Even if absolute distance *values* are not helpful, distance *rankings* can be.
- Shared-neighbor information is based on these findings [Houle et al., 2010].
- In the same way, often
  - outlier rankings are good but
  - the absolute values of the outlier scores are not helpful.

Conclusion

a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between outliers and the group of inliers, as measured on some suitably standardized scale

[Hawkins, 1980]

- outlier rankings may be still good while the underlying outlier scores do not allow to separate between outliers and inliers
- outlier scores are in many models influenced by distance values, that substantially vary over different dimensionality – how can these scores be compared?

Outline

Combinatorial Explosion: Statistics

- for a normal distribution, an object is farther away from the mean than $3 \times \sigma$ *in a single dimension* with a probability of $\approx 0.27\% = 1 - 0.9973$
- for d independently normally distributed dimensions, the combined probability of an object appearing to be *normal in every single dimension* is $\approx 0.9973^d$

$$d = 10 : 97.33\%$$
$$d = 100 : 76.31\%$$
$$d = 1000 : \quad 6.696\%$$

- in high-dimensional distributions, every object is extreme in at least one dimension
- selected subspaces for outliers need to be tested independently

Combinatorial Explosion: Subspace Selection

- 2^d axis-parallel subspaces of a d-dimensional space
- grid-based approaches: 10 bins in each dimension

$$d = 2 : 10^2 \text{ cells (i.e., one hundred)}$$
$$d = 100 : 10^{100} \text{ cells (i.e., one googol)}$$

- need at least as many objects for the cells to not already be empty *on average*
- need even more to draw statistically valid conclusions

Conclusion

- exploding model search space requires improved search heuristics, many established approaches (thresholds, grids, distance functions) no longer work
- evaluating an object against many possible subspaces can introduce a statistical bias ("data snooping")
- Try to do proper statistical hypothesis testing!
- Example: choose *few* candidate subspaces without knowing the candidate object!

Outline

Introduction

The "Curse of Dimensionality"

- Concentration of Distances and of Outlier Scores
- Relevant and Irrelevant Attributes
- Discrimination vs. Ranking of Values
- Combinatorial Issues and Subspace Selection
- **Hubness**
- Consequences

Efficiency and Effectiveness

Subspace Outlier Detection

Hubness

- k-hubness of an object o: $N_k(o)$: the number of times a point o is counted as one of the k nearest neighbors of any other point in the data set
- with increasing dimensionality, many points show a small or intermediate hubness while some points exhibit a very high hubness [Radovanović et al., 2009, 2010]
- related to Zipf's law on word frequencies
- Zipfian distributions frequently seen in social networks
- interpreting the kNN graph as social network, 'hubs' as very popular neighbors
- "Fact or Artifact?" [Low et al., 2013] – not *necessarily* present in high-dimensional data, and can also occur in low-dimensional data

Conclusion

- what does this mean for outlier detection?
  – it is the "Hubs" which are infrequent, but central!
- the other side of the coin:
  anti-hubs might exist that are far away from most other points (i.e., qualify as kNN outliers) yet they are not unusual

Outline

Introduction

The "Curse of Dimensionality"

- Concentration of Distances and of Outlier Scores
- Relevant and Irrelevant Attributes
- Discrimination vs. Ranking of Values
- Combinatorial Issues and Subspace Selection
- Hubness
- Consequences

Efficiency and Effectiveness

Subspace Outlier Detection

Consequences

Problem 1 (Concentration of Scores)

Due to the central limit theorem, the distances of attribute-wise i.i.d. distributed objects converge to an approximately normal distribution with low variance, giving way to numerical and parametrization issues.

Consequences

Problem 2 (Noise attributes)

A high portion of irrelevant (not discriminative) attributes can mask the relevant distances.
We need a good signal-to-noise ratio.

Consequences

Problem 3 (Definition of Reference-Sets)

Common notions of locality (for local outlier detection) rely on distance-based neighborhoods, which often leads to the vicious circle of needing to know the neighbors to choose the right subspace, and needing to know the right subspace to find appropriate neighbors.

Consequences

Problem 4 (Bias of Scores)

Scores based on L_p norms are biased towards high dimensional subspaces, if they are not normalized appropriately. In particular, distances in different dimensionality (and thus distances measured in different subspaces) are not directly comparable.

Consequences

Problem 5 (Interpretation & Contrast of Scores)

Distances and distance-derived scores may still provide a reasonable ranking, while (due to concentration) the scores appear to be virtually identical. Choosing a threshold boundary between inliers and outliers based on the distance or score may be virtually impossible.

Consequences

Problem 6 (Exponential Search Space)

The number of potential subspaces grows exponentially with the dimensionality, making it increasingly hard to systematically scan through the search space.

Consequences

Problem 7 (Data-snooping Bias)

Given enough subspaces, we can find at least one subspace such that the point appears to be an outlier. Statistical principles of testing the hypothesis on a different set of objects need be employed.

Consequences

Problem 8 (Hubness)

What is the relationship of hubness and outlier degree? While antihubs may exhibit a certain affinity to also being recognized as distance-based outliers, hubs are also rare and unusual and, thus, possibly are outliers in a probabilistic sense.

Outline

Introduction

The "Curse of Dimensionality"

Efficiency and Effectiveness

- Fundamental Efficiency Techniques
- Outlier Detection Methods Enhancing Efficiency
- Outlier Detection Methods Enhancing Effectiveness and Stability

Subspace Outlier Detection

Discussion and Conclusion

Efficiency and Effectiveness

- dimensionality reduction / feature selection (e.g., Vu and Gopalkrishnan [2010]) – find all outliers in the remaining or transformed feature space
- global dimensionality reduction (e.g., by PCA) is likely to fail in the typical subspace setting [Keller et al., 2012]
- here, we discuss methods that
  - try to find outliers in the full space (present section) and
    - enhance efficiency
    - enhance effectiveness and stability
  - identify potentially different subspaces for different outliers (next section)

Outline

Approximate Neighborhoods: Random Projection

- Locality sensitive hashing (LSH) [Indyk and Motwani, 1998]: based on approximate neighborhoods in projections
- key ingredient:

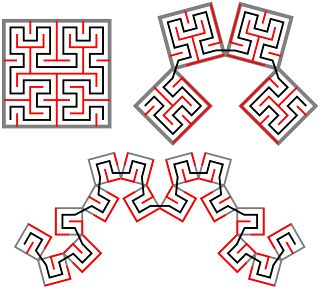
Lemma 3 (Johnson and Lindenstrauss [1984])

There exist projections of n objects into a lower dimensional space (dimensionality $\mathcal{O}(\log n/\varepsilon^2)$) such that the distances are preserved within a factor of $1+\varepsilon$.

- note: reduced dimensionality depends on number of objects and error-bounds, but *not* on the original dimensionality
- popular technique: "database-friendly" (i.e., efficient) random projections [Achlioptas, 2001]

Approximate Neighborhoods: Space-filling Curves

- space-filling curves, like Peano [1890], Hilbert [1891], or the Z-curve [Morton, 1966], do not directly preserve distances but – to a certain extend – neighborhoods
- a one-dimensional fractal curve gets arbitrarily close to every data point without intersecting itself
- intuitive interpretation: repeated cuts, opening the data space
- neighborhoods are not well preserved along these cuts
- number of cuts increases with the dimensionality



Outline

Recursive Binning and Re-projection (RBRP)

RBRP [Ghoting et al., 2008]: adaptation of ORCA [Bay and Schwabacher, 2003] to high-dimensional data, based on a combination of binning and projecting the data

- first phase: bin the data, recursively, into k clusters

  results in k bins, and again, in each bin, k bins and so forth, unless a bin does not contain a sufficient number of points

- second phase: approximate neighbors are listed following their linear order as projected onto the principal component of each bin

  within each bin (as long as necessary): a variant of the nested loop algorithm [Bay and Schwabacher, 2003] derives the top-n outliers

- resulting outliers are reported to be the same as delivered by ORCA but retrieved more efficiently in high-dimensional data

Locality Sensitive Outlier Detection (LSOD)

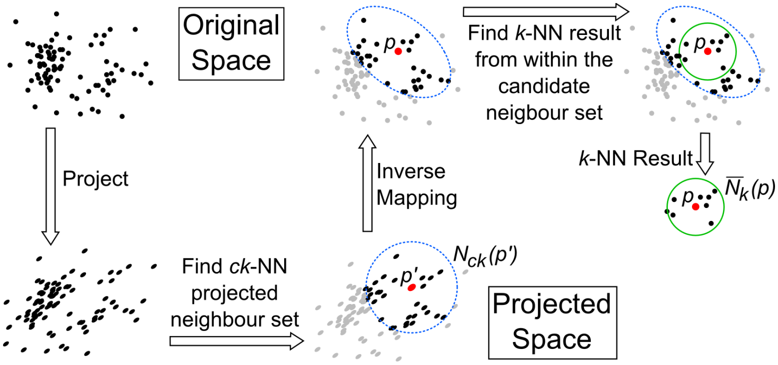
LSOD [Wang et al., 2011]: combination of approximate neighborhood search (here based on LSH) and data partitioning step using a k-means type clustering

- idea of outlierness: points in sparse buckets will probably have fewer neighbors and are therefore more likely to be (distance-based) outliers
- pruning is based on a ranking of this outlier *likelihood*, using statistics on the partitions
- the authors conjecture that their approach *"can be used in conjunction with any outlier detection algorithm"*
- actually, however, the intuition is closely tied to a distance-based notion of outlierness

Projection-indexed Nearest Neighbors (PINN)

- PINN [de Vries et al., 2010, 2012] uses Johnson and Lindenstrauss [1984] lemma
- random projections [Achlioptas, 2001] preserve *distances* approximately
- preserve also *neighborhoods* approximately [de Vries et al., 2010, 2012]
- use projected index (kd-tree, R-tree), query more neighbors than needed
- refine found neighbors to get almost-perfect neighbors

Projection-indexed Nearest Neighbors (PINN)



(Figure by M. E. Houle)

Projection-indexed Nearest Neighbors (PINN)

Advertisement

fundamentals on generalized expansion dimension:
Houle et al. [2012a] at yesterday's workshop
application to similarity search:
Houle et al. [2012b] (talk tomorrow morning)

Space-filling Curves

- Angiulli and Pizzuti [2002, 2005] find top N k-NN-outliers exactly, saves by detecting true misses
- project data to Hilbert curve [Hilbert, 1891]
- sort data, process via sliding window
- multiple scans with shifted curves, refining top candidates and skipping true misses
- good for large data sets in *low* dimensionality:
- Minkowski norms only – suffers from distance concentration: few true misses in high-dimensional data
- Hilbert curves $\sim$ grid based approaches: l^d bits when l bits per dimension

Outline

Introduction

The "Curse of Dimensionality"

Efficiency and Effectiveness

Fundamental Efficiency Techniques

Outlier Detection Methods Enhancing Efficiency

Outlier Detection Methods Enhancing Effectiveness and Stability

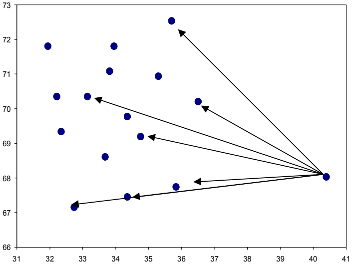
Subspace Outlier Detection

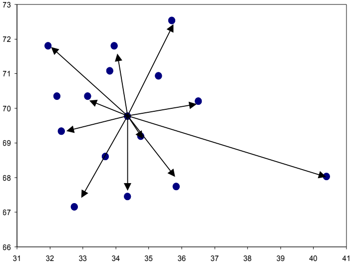
Discussion and Conclusion

Angle-based Outlier Detection

ABOD [Kriegel et al., 2008] uses the variance of angles between points as an outlier degree

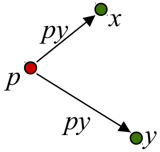
- angles more stable than distances
- outlier: other objects are clustered ⇒ some directions
- inlier: other objects are surrounding ⇒ many directions

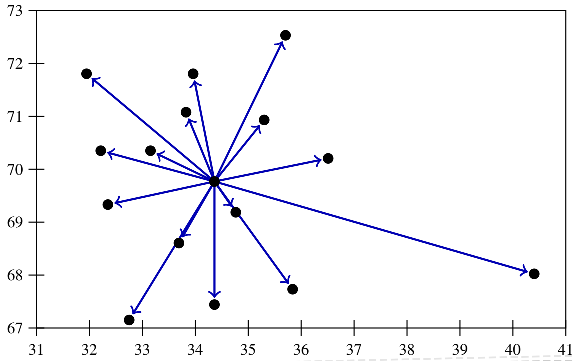




Angle-based Outlier Detection

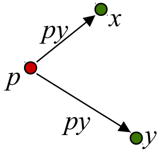
- consider for a given point p the angle between $\vec{px}$ and $\vec{py}$ for any two x, y from the database
- for each point, a measure of variance of all these angles is an outlier score

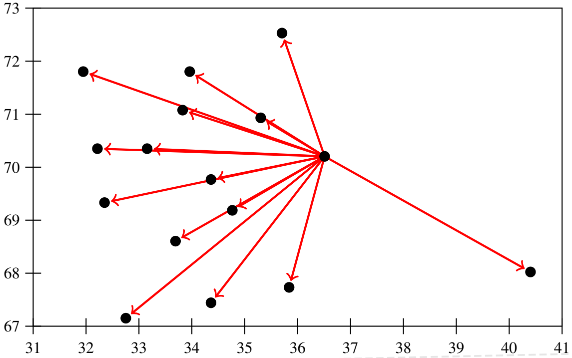




Angle-based Outlier Detection

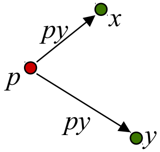
- consider for a given point p the angle between $\vec{px}$ and $\vec{py}$ for any two x, y from the database
- for each point, a measure of variance of all these angles is an outlier score

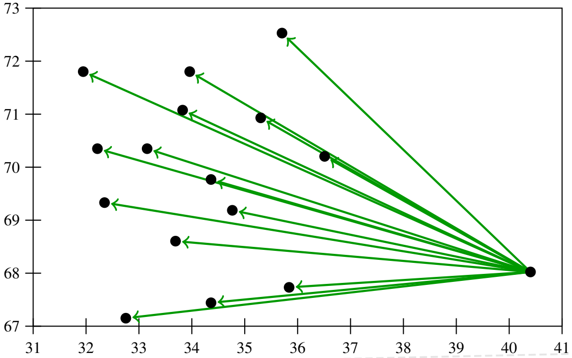




Angle-based Outlier Detection

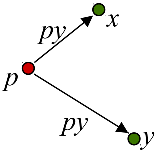
- consider for a given point p the angle between $\vec{px}$ and $\vec{py}$ for any two x, y from the database
- for each point, a measure of variance of all these angles is an outlier score

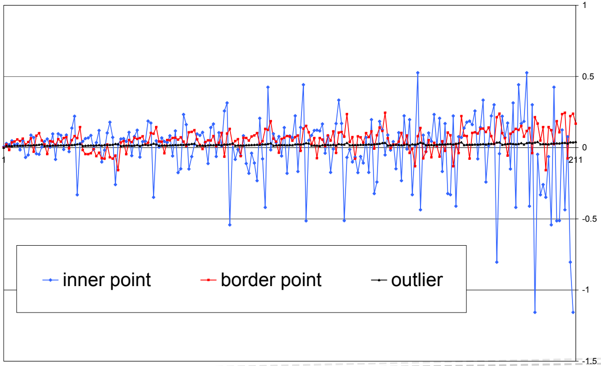




Angle-based Outlier Detection

- consider for a given point p the angle between $\vec{px}$ and $\vec{py}$ for any two x, y from the database
- for each point, a measure of variance of all these angles is an outlier score





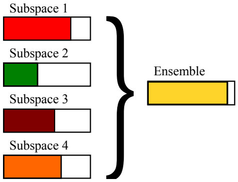
Angle-based Outlier Detection

- ABOD: cubic time complexity
- FastABOD [Kriegel et al., 2008]: approximation based on samples $\Rightarrow$ quadratic time complexity
- LB-ABOD [Kriegel et al., 2008]: approximation as filter-refinement $\Rightarrow$ quadratic time complexity
- approximation based on random-projections and a simplified model [Pham and Pagh, 2012] $\Rightarrow \mathcal{O}(n \log n)$

Feature Subset Combination

"Feature bagging" [Lazarevic and Kumar, 2005]:

- run outlier detection (e.g., LOF) in several random feature subsets (subspaces)
- combine the results to an ensemble



- not a specific approach for high-dimensional data but provides efficiency gains by computations on subspaces and effectiveness gains by ensemble technique
- application to high-dimensional data with improved combination: Nguyen et al. [2010]

Outlier Detection Ensembles

Outlier scores in different subspaces scale differently, have different meaning (Problem 4). Direct combination is problematic.

- improved reasoning about combination, normalization of scores, ensembles of different methods: Kriegel et al. [2011]
- study of the impact of diversity on ensemble outlier detection: Schubert et al. [2012a]

In general, ensemble techniques for outlier detection have potential to address problems associated with high-dimensional data. Research here has only begun.

Outline

Outliers in Subspaces

- feature bagging uses *random* subspaces to derive a full dimensional result
- "subspace outlier detection" aims at finding outliers in *relevant* subspaces that are not outliers in the full-dimensional space (where they are covered by "irrelevant" attributes)
- predominant issues are
  1. identification of subspaces:
     Which subspace is relevant and why?
     (recall data snooping bias, Problem 7)
  2. comparability of outlier scores:
     How to compare outlier results from different subspaces (of different dimensionality)?
     (cf. Problem 4: Bias of Scores)

Outline

Subspace Outlier Detection

common Apriori [Agrawal and Srikant, 1994]-like procedure for subspace *clustering* [Kriegel et al., 2009c, 2012b, Sim et al., 2012]:

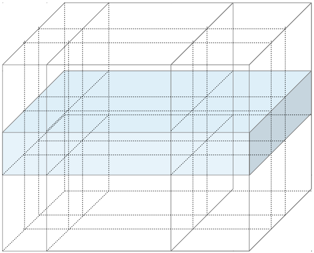
- evaluate all n-dimensional subspaces (e.g., look for clusters in the corresponding subspace)
- combine all "interesting" (e.g., containing clusters) n-dim. subspaces (i.e., "candidates") to $n+1$-dim. subspaces
- start with 1-dim. subspaces and repeat this bottom-up search until no candidate subspaces remain
- requirement: anti-monotonicity of the criterion of "interestingness" (usually the presence of clusters)

unfortunately, no meaningful outlier criterion is known so far that behaves anti-monotoneously

Subspace Outlier Detection

first approach for high-dimensional (subspace) outlier detection: Aggarwal and Yu [2001]

- resembles a grid-based subspace clustering approach but not searching dense but sparse grid cells
- report objects contained within sparse grid cells as outliers
- evolutionary search for those grid cells (Apriori-like search not possible, complete search not feasible)



- divide data space in ϕ equi-depth cells
- each 1-dim. hyper-cuboid contains $f = \frac{N}{\phi}$ objects
- expected number of objects in k-dim. hyper-cuboid: $N \cdot f^k$
- standard deviation: $\sqrt{N \cdot f^k \cdot (1 - f^k)}$
- "sparse" grid cells: contain unexpectedly few data objects

Problems of Aggarwal and Yu [2001]

- with increasing dimensionality, the expected value of a grid cell quickly becomes too low to find significantly sparse grid cells $\Rightarrow$ only small values for k meaningful (Problem 6: Exponential Search Space)
- parameter k must be fixed, as the scores are not comparable across different values of k (Problem 4)
- search space is too large even for a fixed k $\Rightarrow$ genetic search preserving the value of k across mutations (Problem 6)
- restricted computation time allows only inspection of a tiny subset of the $\binom{n}{k}$ projections (not yet to speak of individual subspaces); randomized search strategy does encourage neither fast enough convergence nor diversity $\Rightarrow$ no guarantees about the outliers detected or missed
- randomized model optimization without a statistical control $\Rightarrow$ statistical bias (Problem 7): how meaningful are the detected outliers?
- presence of clusters in the data set will skew the results considerably
- equidepth binning is likely to include outliers in the grid cell of a nearby cluster $\Rightarrow$ hide them from detection entirely
- dense areas also need to be refined to detect outliers that happen to fall into a cluster bin

HOS-Miner

Zhang et al. [2004] identify the subspaces in which a given point is an outlier

- define the outlying degree of a point w.r.t. a certain space (or possibly a subspace) s in terms of the sum of distances to the k nearest neighbors in this (sub-)space s
- for a fixed subspace s, this is the outlier model of Angiulli and Pizzuti [2002]
- monotonic behavior over subspaces and superspaces of s, since the outlying degree OD is directly related to the distance-values; for L_p-norms the following property holds for any object o and subspaces s_1, s_2:
  $OD_{s_1}(o) \geq OD_{s_2}(o) \iff s_1 \supseteq s_2$
- Apriori-like search for outlying subspaces for any query point: threshold T discriminates outliers ($OD_s(o) \geq T$) from inliers ($OD_s(o) < T$) in any subspace s

Problems of HOS-Miner [Zhang et al., 2004]

- fixed threshold to discern outliers w.r.t. their score OD in subspaces of different dimensionality $\Rightarrow$ these scores are rather incomparable (Problem 4)
- the monotonicity *must* not be fulfilled for true subspace outliers (since it would imply that the outlier can be found trivially in the full-dimensional space) — as pointed out by Nguyen et al. [2011]
- systematic search for the subspace with the highest score $\Rightarrow$ data-snooping bias (Problem 7)

OutRank

Müller et al. [2008] analyse the result of some (grid-based/density-based) subspace clustering algorithm

- clusters are more stable than outliers to identify in different subspaces
- avoids statistical bias
- outlierness: how often is the object recognized as part of a cluster and what is the dimensionality and size of the corr. subspace clusters

Problems of OutRank [Müller et al., 2008]

- a strong redundancy in the clustering is implicitly assumed — result biased towards (anti-)hubs? (Problem 8)
- outliers as just a side-product of density-based clustering can result in a large set of outliers
- outlier detection based on subspace clustering relies on the subspace clusters being well separated
  - Theorem 2 (Separation of Clusters)
  - Problem 1 (Concentration Effect)
  - Problem 2 (Noise Attributes)

Problems of OutRank [Müller et al., 2008]

- a strong redundancy in the clustering is implicitly assumed — result biased towards (anti-)hubs? (Problem 8)
- outliers as just a side-product of density-based clustering can result in a large set of outliers
- outlier detection based on subspace clustering relies on the subspace clusters being well separated
  - Theorem 2 (Separation of Clusters)
  - Problem 1 (Concentration Effect)
  - Problem 2 (Noise Attributes)

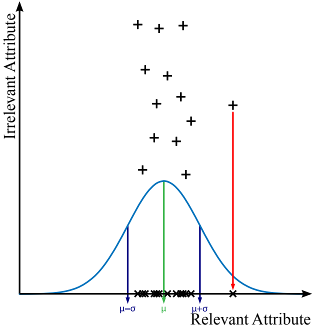
Advertisement

note a follow-up on this workshop paper at this ICDM: Müller et al. [2012] (talk tomorrow morning)

SOD

SOD (subspace outlier detection) [Kriegel et al., 2009a] finds outliers in subspaces without an *explicit* clustering

- a reference set is possibly defining (implicitly) a subspace cluster (or a part of such a cluster)
- If the query point deviates considerably from the subspace of the reference set, it is a subspace outlier w.r.t. the corresponding subspace.
- not a decision (outlier vs. inlier) but a (normalized, sort of) subspace distance outlier score



Problems of SOD [Kriegel et al., 2009a]

- how to find a good reference set (Problem 3)?
  Kriegel et al. [2009a] define the reference-set using SNN-distance [Houle et al., 2010], which introduces a second neighborhood parameter
- normalization of scores is over-simplistic
  interpretation as "probability estimates" of the (subspace) distance distribution would be a desirable post-processing to tackle Problem 5 (Interpretation and Contrast of Scores)

OUTRES [Müller et al., 2010]

- assess deviations of each object in several subspaces simultaneously
- combine ranking of the objects according to their outlier scores in all 'relevant subspaces'
- requires comparable neighborhoods (Problem 3) for each point to estimate densities
- adjust for different number of dimensions of subspaces (Problem 4): specific ε radius for each subspace
- score in a single subspace: comparing the object's density to the average density of its neighborhood
- total score of an object is the product of all its scores in all relevant subspaces

Assuming a score in $[0, 1]$ (smaller score $\propto$ stronger outlier), this should provide a good contrast for those outliers with very small scores in many relevant subspaces.

OUTRES 2 [Müller et al., 2011]

follow-up paper [Müller et al., 2011] describes selection of relevant subspaces

- reject attributes with uniformly distributed values in the neighborhood of the currently considered point o (statistical significance test)
- exclude, for this o, also any superspaces of uniformly distributed attributes
- Apriori-like search strategy can be applied to find subspaces for each point
- tackles the problem of noise attributes (Problem 2)
- based on a statistic on the neighborhood of the point $\Rightarrow$ not likely susceptible to a statistical bias (Problem 7)

Problems of OUTRES [Müller et al., 2010, 2011]

tackling many problems comes for a price:

- Apriori-like search strategy finds *subspaces for each point*, not outliers in the subspaces $\Rightarrow$ expensive approach: worst-case exponential behavior in dimensionality
- score adaptation to locally varying densities as the score of a point o is based on a comparison of the density around o vs. the average density among the neighbors of o ($\sim$ LOF [Breunig et al., 2000]) $\Rightarrow$ time complexity $O(n^3)$ for a database of n objects unless suitable data structures (e.g., precomputed neighborhoods) are used
- due to the adaptation to different dimensionality of subspaces, data structure support is not trivial

HighDOD

HighDOD (High-dimensional Distance-based Outlier Detection) [Nguyen et al., 2011]:

- motivation: the sum of distances to the k nearest neighbors as the outlier score [Angiulli and Pizzuti, 2005] is monotonic over subspaces – but a subspace search (as in HOS-Miner) is pointless as the maximum score will appear in the full-dimensional space
- modify the kNN-weight outlier score to use a normalized L_p norm
- pruning of subspaces is impossible, examine all subspaces up to a user-defined maximum dimensionality m
- use a linear-time ($\mathcal{O}(n \cdot m)$) density estimation to generate outlier candidates they compute the nearest neighbors for

Problems of HighDOD [Nguyen et al., 2011]

- examine all subspaces $\Rightarrow$ data-snooping (Problem 7)?
- no normalization to adjust different variances in different dimensionality

HiCS [Keller et al., 2012]

high contrast subspaces (HiCS) [Keller et al., 2012]

- core concept for subspaces with high contrast: correlation among the attributes of a subspace

  (deviation of the observed PDF from the expected PDF, assuming independence of the attributes)
- Monte Carlo samples to aggregate these deviations
- aggregate the LOF scores for a single object over all "high contrast" subspaces
- the authors suggest that, instead of LOF, any other outlier measure could be used
- intuition: in these subspaces, outliers are not trivial (e.g., identifiable already in 1-dimensional subspaces) but deviate from the (although probably non-linear and complex) correlation trend exhibited by the majority of data in this subspace

Problems of HiCS [Keller et al., 2012]

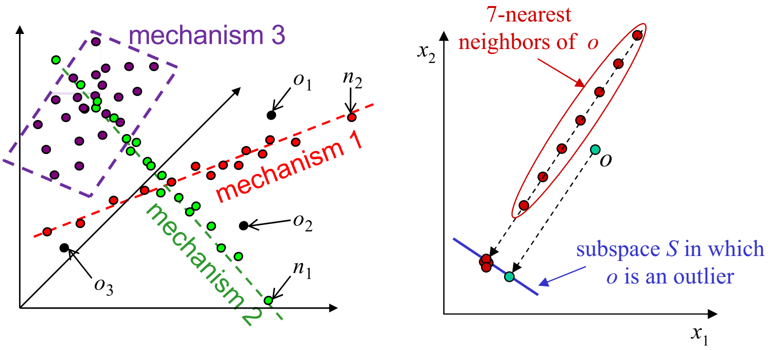
- combine LOF scores from subspaces of different dimensionality without score normalization (Problem 4: Bias of Scores)
- combination of scores is rather naïve, could benefit from ensemble reasoning
- philosophy of decoupling subspace search and outlier ranking is questionable:
  - a certain measure of contrast to identify interesting subspaces will relate quite differently to different outlier ranking measures
  - their measure of interestingness is based on an implicit notion of density, it may only be appropriate for density-based outlier scores
- however, this decoupling allows them to discuss the issue of subspace selection with great diligence as this is the focus of their study

Correlation Outlier

- so far, most algorithms for subspace outlier detection are restricted to axis-parallel subspaces

  e.g., due to grid-based approaches or to the required first step of subspace or projected clustering
- HiCS [Keller et al., 2012] is not restricted in this sense.
- earlier example for outliers in arbitrarily-oriented subspaces: COP (correlation outlier probability) [Zimek, 2008, ch. 18] (application of the correlation clustering concepts discussed by Achtert et al. [2006])
- high probability of being a "correlation outlier" if neighbors show strong linear dependencies among attributes and the point in question deviates substantially from the corresponding linear model

Correlation Outlier



Advertisement

advanced version of this earlier idea at this ICDM (tomorrow morning): [Kriegel et al., 2012a]

Outline

Introduction

The "Curse of Dimensionality"

Efficiency and Effectiveness

Subspace Outlier Detection

Identification of Subspaces

Comparability of Outlier Scores

Discussion and Conclusion

Comparability of Outlier Scores

- An outlier score provided by some outlier model should help the user to decide whether an object actually is an outlier or not.
- For many approaches even in low dimensional data the outlier score is not readily interpretable.
  - The scores provided by different methods differ widely in their scale, their range, and their meaning.
  - For many methods, the scaling of occurring values of the outlier score even differs within the same method from data set to data set.
  - Even within one data set, the identical outlier score o for two different database objects can denote actually substantially different degrees of outlierness, depending on different local data distributions.

Solutions in Low-dimensional Data

- LOF [Breunig et al., 2000] intends to level out different density values in different regions of the data, as it assesses the *local* outlier factor
- LoOP [Kriegel et al., 2009b] (a LOF variant) provides a statistical interpretation of the outlier score by translating it into a probability estimate (including a normalization to become independent from the specific data distribution in a given data set)
- Kriegel et al. [2011] proposed generalized scaling methods for a range of different outlier models
  - allows comparison and combination of different methods
  - or results from different feature subsets

More Problems in High-dimensional Data

Problems 4 (Bias of Scores) and 5 (Interpretation and Contrast of Scores)

- most outlier scorings are based on assessment of distances, usually L_p distances
- can be expected to grow with additional dimensions, while the relative variance decreases
- a numerically higher outlier score, based on a subspace of more dimensions, does not necessarily mean the corresponding object is a stronger outlier than an object with a numerically lower outlier score, based on a subspace with less dimensions
- many methods that combine multiple scores into a single score neglect to normalize the scores before the combination (e.g. using the methods discussed by Kriegel et al. [2011])

Treatment of the Comparability-Problem in Subspace Methods

- model of Aggarwal and Yu [2001] circumvents the problem since they restrict the search for outliers to subspaces of a fixed dimensionality (given by the user as input parameter)
- OutRank [Müller et al., 2008] weights the outlier scores by size and dimensionality of the corresponding reference cluster
- SOD [Kriegel et al., 2009a] uses a normalization over the dimensionality, but too simplistic

Treatment of the Comparability-Problem in Subspace Methods (contd.)

- For OUTRES [Müller et al., 2010, 2011], this problem of bias is the core motivation:
  - uses density estimates that are based on the number of objects within an ε-range in a given subspace
  - uses adaptive neighborhood (ε is increasing with dimensionality)
  - uses adaptive density by scaling the distance values accordingly
  - score is also adapted to locally varying densities as the score of a point o is based on a comparison of the density around o vs. the average density among the neighbors of o

Treatment of the Comparability-Problem in Subspace Methods (contd.)

- bias of distance-based outlier scores towards higher dimensions is also the main motivation for HighDOD [Nguyen et al., 2011]
  - kNN-weight outlier score
  - adapt the distances (L_p-norm) to the dimensionality d of the corresponding subspace by scaling the sum over attributes with $1/\sqrt[p]{d}$
  - assuming normalized (!) attributes (with a value range in $[0,1]$), this results in restricting each summand to ≤ 1 and the sum therefore to $\leq k$, irrespective of the considered dimensionality
- HiCS [Keller et al., 2012]: LOF scores retrieved in subspaces of different dimensionality are aggregated for a single object without normalization

  no problem in their experiments since the relevant subspaces vary only between 2 and 5 dimensions

Outline

Introduction

The "Curse of Dimensionality"

Efficiency and Effectiveness

Subspace Outlier Detection

Discussion and Conclusion

Tools and Implementations

Data mining framework
ELKI [Achtert et al., 2012]:
`http://elki.dbs.ifi.lmu.de/`

Environment for DeveLoping KDD-Applications Supported by Index-Structures

- Open Source: AGPL 3+
- 20+ standard (low-dim.) outlier detection methods
- 10+ spatial ("geo") outlier detection methods
- 4 subspace outlier methods: COP [Kriegel et al., 2012a], SOD [Kriegel et al., 2009a], OUTRES [Müller et al., 2010], OutRank S1 [Müller et al., 2008]
- meta outlier methods: HiCS [Keller et al., 2012], Feature Bagging [Lazarevic and Kumar, 2005], more ensemble methods . . .
- 25+ clustering algorithms (subspace, projected, . . .)
- index structures, evaluation, and visualization

Tools and Implementations

ALOI [Geusebroek et al., 2005] image data set
RGB histograms, 110,250 objects, 8 dimensions:
Same algorithm, very different performance:

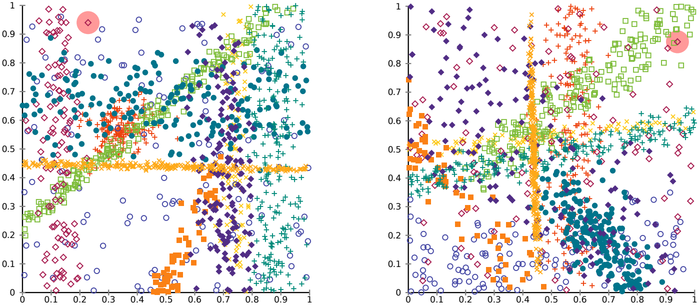
LOF, "Data Mining with R":	13402.38 sec
LOF, Weka implementation:	2611.60 sec
LOF, ELKI without index:	570.94 sec
LOF, ELKI with STR R*-Tree:	47.24 sec
LOF, ELKI STR R*, multi-core:	26.73 sec

- due to the modular architecture and high code reuse, optimizations in ELKI work very well
- ongoing efforts for *subspace indexing*

Requires some API learning, but there is a tutorial on implementing a new outlier detection algorithm:

`http://elki.dbs.ifi.lmu.de/wiki/Tutorial/Outlier`

Visualization – Scatterplots

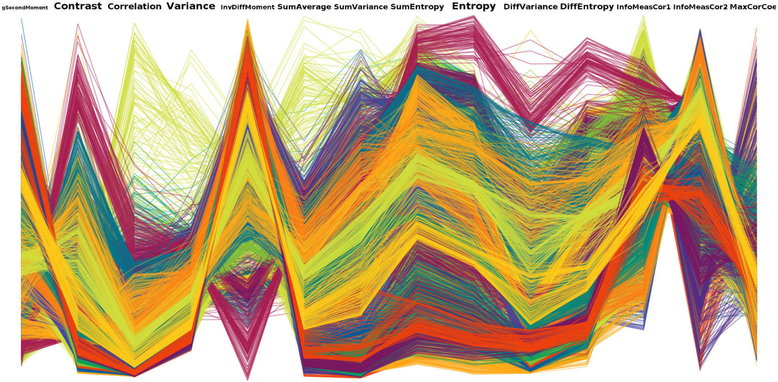


Visualized with ELKI.

- Scatterplots can only visualize low-dimensional projections of high-dimensional dataspaces.
- Nevertheless, visual inspection of several two-dimensional subspaces (if the data dimensionality is not too high) can be insightful.

Visualization – Parallel Coordinates

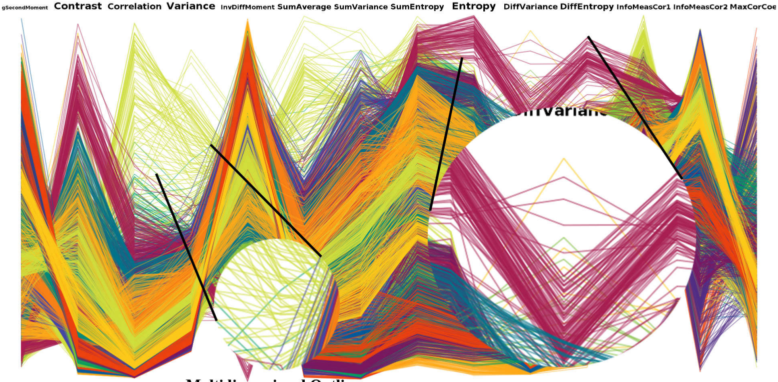
A. Zimek,
E. Schubert,
H.-P. Kriegel



Visualized with ELKI.

Parallel coordinates can visualize high-dimensional data. But every axis has only two neighbors – so actually not much more than scatterplots.

Visualization – Parallel Coordinates



Visualized with ELKI.

When arranged well, some outliers are well visible – others remain hidden.

Data Preparation, Normalization, and Bias

- preselecting attributes helps – but we would like the algorithms to do this automatically
- distance functions are heavily affected by normalization
- linear normalization $\cong$ feature weighting
  $\Rightarrow$ bad normalization $\cong$ bad feature weighting
- some algorithms are very sensitive to different preprocessing procedures
- not often discussed, the choice of a distance function can also have strong impact [Schubert et al., 2012a,b]
- subspace (or correlation) selection is influenced by the outliers that are to detect (*vicious circle*, known as "swamping" and "masking" in statistics) – requires "robust" measures of variance etc. (e.g., robust PCA)

Evaluation Measures

Classic evaluation: **Precision**@k and ROC AUC

True positive rate of known outliers in the top k.

Example: 7 out of 10 correct = 0.7

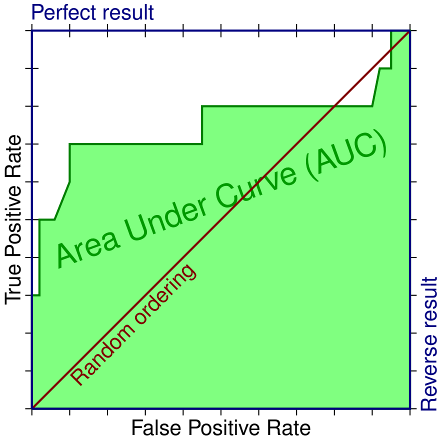
Elements past the top k are ignored. $\Rightarrow$ very crude
$k+1$ different values: $0 \dots k$ out of k correct.

Order within the top k is *ignored*.
3 false, then 7 correct $\equiv$ 7 correct, then 3 false

Average precision: same, but for $k = 1 \dots k_{\text{max}}$

Evaluation Measures

Classic evaluation: Precision@k and **ROC AUC**



Y: True positive rate
X: False positive rate

Measure: Area under Curve
Optimal: 1.000
Random: 0.500
Reverse: 0.000

Intuitive interpretation: given a pair (pos, neg): what is the chance of it being correctly ordered?

Evaluation Measures

Classic evaluation: Precision@k and ROC AUC

The popular measures

- evaluate the order of points only, *not* the scores.
- need outlier labels ...
- ...and assume that all outliers are *known*!

$\Rightarrow$ future work needed!

Evaluation: Pitfalls

- common procedure: using labeled data sets for evaluation of unsupervised methods
- highly imbalanced problem
- "ground truth" may be incomplete
- real world data may include sensible outliers that are just not yet known or were considered uninteresting during labeling
- use classification data sets assuming that some rare (or down-sampled) class contains the outliers?
  - but the rare class may be clustered
  - true outliers may occur in the frequent classes
- If a method is detecting such outliers, that should actually be rated as a good performance of the method.
- Instead, in this setup, detecting such outliers is overly punished due to class imbalance.

Evaluation of Outlier Scores?

When comparing or combining results (different subspaces, ensemble), *meaningful* score values are more informative than ranks:

- Kriegel et al. [2011], Schubert et al. [2012a] have initial attempts on evaluating score values
  - more weight on known (or estimated) outliers
  - allows non-binary ground-truth
  - improves outlier detection ensembles by combining preferably diverse (dissimilar) score vectors
- this direction of research aims in the long run to get calibrated outlier scores reflecting a notion of probability

Efficiency

- focus of research so far: identification of meaningful subspaces for outlier detection
- open problem: efficiency in subspace similarity search (many methods need to assess neighborhoods in different subspaces)
- only some preliminary approaches around: [Kriegel et al., 2006, Müller and Henrich, 2004, Lian and Chen, 2008, Bernecker et al., 2010a,b, 2011]
- HiCS uses 1 dimensional pre-sorted arrays (i.e., a very simple subspace index)
- can subspace similarity index structures help?
  — and can they be improved?
- however, first make the methods work well, then make them work fast!

Conclusion

We hope that you learned in this tutorial about

- typical problems associated with high-dimensional data (*"curse of dimensionality"*)
- the corresponding challenges and problems for outlier detection
- approaches to improve efficiency and effectiveness for outlier detection in high-dimensional data
- specialized methods for subspace outlier detection
  - how they treat some of the problems we identified
  - how they are possibly tricked by some of these problems
- tools, caveats, open issues for outlier detection (esp. in high-dimensional data)

Conclusion

More details in our survey article:
Zimek, Schubert, and Kriegel [2012]: A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387 (`http://dx.doi.org/10.1002/sam.11161`)

And we hope that you got inspired to tackle some of these open issues or known problems (or identify yet more problems) in your next (ICDM 2013?) paper!

Thank you for your attention!

LMU

UNIVERSITY OF ALBERTA
FACULTY OF SCIENCE

References I

D. Achlioptas. Database-friendly random projections. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA*, 2001.

E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. Deriving quantitative models for correlation clusters. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA*, pages 4–13, 2006. doi: 10.1145/1150402.1150408.

E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek. Evaluation of clusterings – metrics and visual support. In *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC*, pages 1285–1288, 2012. doi: 10.1109/ICDE.2012.128.

C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Santa Barbara, CA*, pages 37–46, 2001.

R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile*, pages 487–499, 1994.

References II

F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discoverys (PKDD), Helsinki, Finland*, pages 15–26, 2002. doi: 10.1007/3-540-45681-3_2.

F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203–215, 2005.

S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC*, pages 29–38, 2003. doi: 10.1145/956750.956758.

K. P. Bennett, U. Fayyad, and D. Geiger. Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA*, pages 233–243, 1999. doi: 10.1145/312129.312236.

T. Bernecker, T. Emrich, F. Graf, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek. Subspace similarity search using the ideas of ranking and top-k retrieval. In *Proceedings of the 26th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank), Long Beach, CA*, pages 4–9, 2010a. doi: 10.1109/ICDEW.2010.5452771.

References III

T. Bernecker, T. Emrich, F. Graf, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, and A. Zimek. Subspace similarity search: Efficient k-nn queries in arbitrary subspaces. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany*, pages 555–564, 2010b. doi: 10.1007/978-3-642-13818-8_38.

T. Bernecker, F. Graf, H.-P. Kriegel, C. Moennig, and A. Zimek. BeyOND – unleashing BOND. In *Proceedings of the 37th International Conference on Very Large Data Bases (VLDB) Workshop on Ranking in Databases (DBRank), Seattle, WA*, pages 34–39, 2011.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is “nearest neighbor” meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT), Jerusalem, Israel*, pages 217–235, 1999. doi: 10.1007/3-540-49257-7_15.

M. M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 93–104, 2000.

References IV

T. de Vries, S. Chawla, and M. E. Houle. Finding local anomalies in very high dimensional space. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia*, pages 128–137, 2010. doi: 10.1109/ICDM.2010.151.

T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. *Knowledge and Information Systems (KAIS)*, 32(1):25–52, 2012. doi: 10.1007/s10115-011-0430-4.

J. M. Geusebroek, G. J. Burghouts, and A.W.M. Smeulders. The Amsterdam Library of Object Images. *International Journal of Computer Vision*, 61(1): 103–112, 2005. doi: 10.1023/B:VISI.0000042993.50813.60.

A. Ghoting, S. Parthasarathy, and M. E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. *Data Mining and Knowledge Discovery*, 16(3):349–364, 2008. doi: 10.1007/s10618-008-0093-2.

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

D. Hilbert. Ueber die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38(3):459–460, 1891.

References V

M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany*, pages 482–500, 2010. doi: 10.1007/978-3-642-13818-8_34.

M. E. Houle, H. Kashima, and M. Nett. Generalized expansion dimension. In *ICDM Workshop Practical Theories for Exploratory Data Mining (PTDM)*, 2012a.

M. E. Houle, X. Ma, M. Nett, and V. Oria. Dimensional testing for multi-step similarity search. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012b.

P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing (STOC), Dallas, TX*, pages 604–613, 1998. doi: 10.1145/276698.276876.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

References VI

F. Keller, E. Müller, and K. Böhm. HiCS: high contrast subspaces for density-based outlier ranking. In *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC*, 2012.

E. M. Knorr and R. T. Ng. A unified notion of outliers: Properties and computation. In *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, CA*, pages 219–222, 1997.

H.-P. Kriegel, P. Kröger, M. Schubert, and Z. Zhu. Efficient query processing in arbitrary subspaces using vector approximations. In *Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM), Vienna, Austria*, pages 184–190, 2006. doi: 10.1109/SSDBM.2006.23.

H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, pages 444–452, 2008. doi: 10.1145/1401890.1401946.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, pages 831–838, 2009a. doi: 10.1007/978-3-642-01307-2_86.

References VII

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, pages 1649–1652, 2009b. doi: 10.1145/1645953.1646195.

H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1–58, 2009c. doi: 10.1145/1497577.1497578.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, pages 13–24, 2011.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Outlier detection in arbitrarily oriented subspaces. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012a.

H.-P. Kriegel, P. Kröger, and A. Zimek. Subspace clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4):351–364, 2012b.

A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, pages 157–166, 2005. doi: 10.1145/1081870.1081891.

References VIII

X. Lian and L. Chen. Similarity search in arbitrary subspaces under Lp-norm. In *Proceedings of the 24th International Conference on Data Engineering (ICDE), Cancun, Mexico*, pages 317–326, 2008. doi: 10.1109/ICDE.2008.4497440.

T. Low, C. Borgelt, S. Stober, and A. Nürnberger. The hubness phenomenon: Fact or artifact? In C. Borgelt, M. Á. Gil, J. M. C. Sousa, and M. Verleysen, editors, *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, Studies in Fuzziness and Soft Computing, pages 267–278. Springer Berlin / Heidelberg, 2013.

G. M. Morton. A computer oriented geodetic data base and a new technique in file sequencing. Technical report, International Business Machines Co., 1966.

E. Müller, I. Assent, U. Steinhausen, and T. Seidl. OutRank: ranking outliers in high dimensional data. In *Proceedings of the 24th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank), Cancun, Mexico*, pages 600–603, 2008. doi: 10.1109/ICDEW.2008.4498387.

E. Müller, M. Schiffer, and T. Seidl. Adaptive outlierness for subspace outlier ranking. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, ON, Canada*, pages 1629–1632, 2010. doi: 10.1145/1871437.1871690.

References IX

E. Müller, M. Schiffer, and T. Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, pages 434–445, 2011. doi: 10.1109/ICDE.2011.5767916.

E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm. Outlier ranking via subspace analysis in multiple views of the data. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012.

W. Müller and A. Henrich. Faster exact histogram intersection on large data collections using inverted VA-files. In *Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR), Dublin, Ireland*, pages 455–463, 2004. doi: 10.1007/978-3-540-27814-6_54.

H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan*, pages 368–383, 2010. doi: 10.1007/978-3-642-12026-8_29.

References X

H. V. Nguyen, V. Gopalkrishnan, and I. Assent. An unbiased distance-based outlier detection approach for high-dimensional data. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA), Hong Kong, China*, pages 138–152, 2011. doi: 10.1007/978-3-642-20149-3_12.

G. Peano. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160, 1890. doi: 10.1007/BF01199438.

N. Pham and R. Pagh. A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China*, 2012.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. In *Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, QC, Canada*, pages 865–872, 2009. doi: 10.1145/1553374.1553485.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

References XI

S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 427–438, 2000.

E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA*, pages 1047–1058, 2012a.

E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 2012b.

K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong. A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery*, 2012. doi: 10.1007/s10618-012-0258-x.

N. H. Vu and V. Gopalkrishnan. Feature extraction for outlier detection in high-dimensional spaces. *Journal of Machine Learning Research*, Proceedings Track 10:66–75, 2010.

Y. Wang, S. Parthasarathy, and S. Tatikonda. Locality sensitive outlier detection: A ranking driven approach. In *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, pages 410–421, 2011. doi: 10.1109/ICDE.2011.5767852.

References XII

J. Zhang, M. Lou, T. W. Ling, and H. Wang. HOS-miner: A system for detecting outlying subspaces of high-dimensional data. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada*, pages 1265–1268, 2004.

A. Zimek. *Correlation Clustering*. PhD thesis, Ludwig-Maximilians-Universität München, Munich, Germany, 2008.

A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012. doi: 10.1002/sam.11161.