December 15, 2017

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

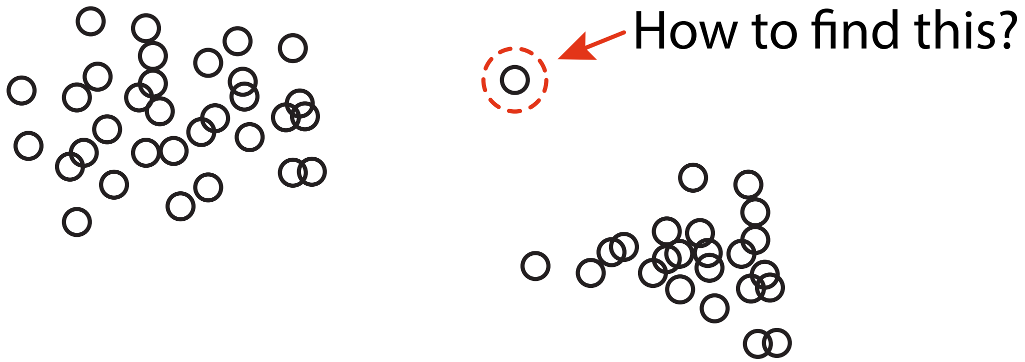
Outlier Detection

Data Mining 08 (データマイニング)

Mahito Sugiyama (杉山麿人)

Today's Outline

- Today's topic is outlier detection
  - studied in statistics, machine learning & data mining (unsupervised learning)
- **Problem:** How can we find outliers efficiently (from massive data) ?



What is an Outlier (Anomaly) ?

- An outlier is "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (by Hawkins, 1980)
  - There is no fixed mathematical definition

- Outliers appear everywhere:
  - Intrusions in network traffic, credit card fraud,
    defective products in industry, medical diagnosis from X-ray images

- Outliers should be detected and removed

- Outliers can cause fake results in subsequent analysis

Distance-Based Outlier Detection

- The modern distance-based approach
  - A data point is an outlier, if its locality is sparsely populated
  - One of the most popular approaches in outlier detection
    - Distribution-free
    - Easily applicable for various types of data

- See the following for other traditional model-based approaches, e.g., statistical tests or changes of variances
  - Aggarwal, C. C., Outlier Analysis, Springer (2013)
  - Kriegel, H.-P., Kröger, P., Zimak, A., Outlier Detection Techniques, Tutorial at SIGKDD2010 [Link]
  - 井手剛，入門 機械学習による異常検知，コロナ社，(2015)

The First Distance-Based Method

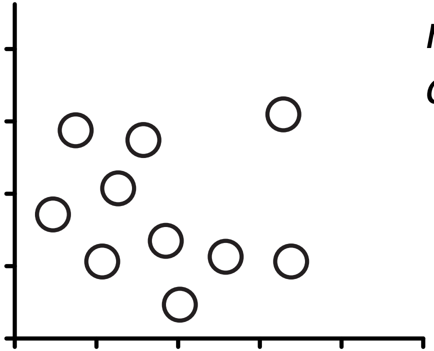
- Knorr and Ng were the first to formalize a distance-based outlier detection scheme
  - "Algorithms for mining distance-based outliers in large datasets", VLDB 1998

The First Distance-Based Method

- Knorr and Ng were the first to formalize a distance-based outlier detection scheme

- Given a dataset X, an object $x \in X$ is a DB(α, δ)-outlier if
  $|\{ x' \in X \mid d(x, x') > \delta \}| \geq \alpha n$

- $n = |X|$ (number of objects)
- $\alpha, \delta \in \mathbb{R}$ ($0 \leq \alpha \leq 1$) are parameters



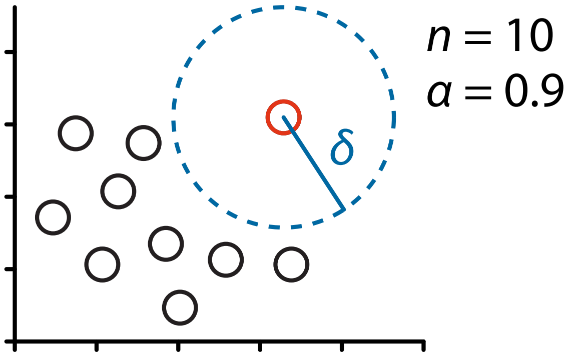
$n = 10$
$\alpha = 0.9$

The First Distance-Based Method

- Knorr and Ng were the first to formalize a distance-based outlier detection scheme

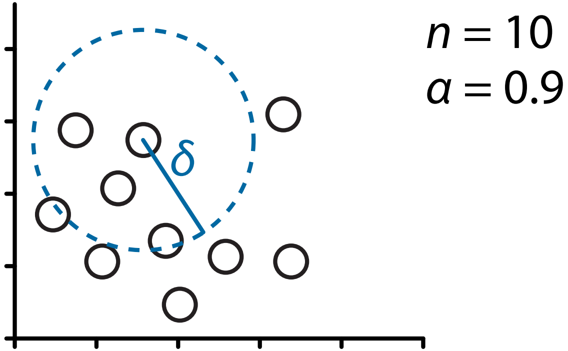
- Given a dataset X, an object $x \in X$ is a DB(α, δ)-outlier if
  $|\{ x' \in X \mid d(x, x') > \delta \}| \geq \alpha n$

- $n = |X|$ (number of objects)
- $\alpha, \delta \in \mathbb{R}$ $(0 \leq \alpha \leq 1)$ are parameters



The First Distance-Based Method

- Knorr and Ng were the first to formalize a distance-based outlier detection scheme
- Given a dataset X, an object $x \in X$ is a DB(α, δ)-outlier if $|\{ x' \in X \mid d(x, x') > \delta \}| \geq \alpha n$
- $n = |X|$ (number of objects)
- $\alpha, \delta \in \mathbb{R}$ $(0 \leq \alpha \leq 1)$ are parameters



From Classification to Ranking

- Two drawbacks of DB(α, δ)-outliers
  - (i) Setting the distance threshold δ is difficult in practice
    - Setting α is not so difficult since it is always close to 1
  - (ii) The lack of a ranking of outliers

- Ramaswamy *et al.* proposed to measure the outlierness by the *k*th-nearest neighbor (*k*th-NN) distance
  - Ramaswamy, S., Rastogi, R., Shim, K., "Efficient algorithms for mining outliers from large data sets", SIGMOD 2000

From Classification to Ranking

- Two drawbacks of DB(α, δ)-outliers
  (i) Setting the distance threshold δ is difficult in practice
    - Setting α is not so difficult since it is always close to 1
  (ii) The lack of a ranking of outliers

- Ramaswamy *et al.* proposed to measure the outlierness by the *k*th-nearest neighbor (*k*th-NN) distance
  – Ramaswamy, S., Rastogi, R., Shim, K., "Efficient algorithms for mining outliers from large data sets", SIGMOD 2000

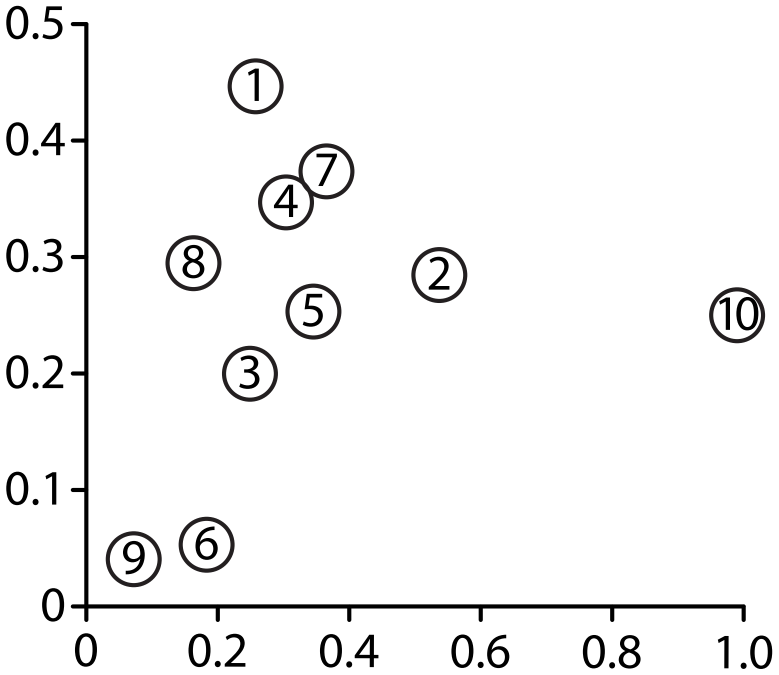
- ***From this study, the task of DB outlier detection becomes a ranking problem*** (without binary classification)

The *k*th-Nearest Neighbor Distance

- The *k*th-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X

The *k*th-Nearest Neighbor Distance

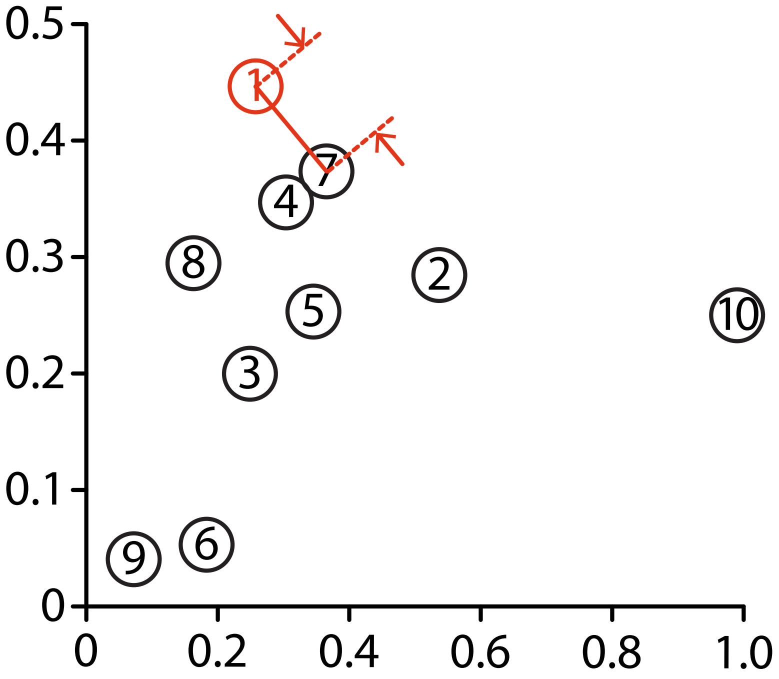
- The *k*th-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$



id	score

The *k*th-Nearest Neighbor Distance

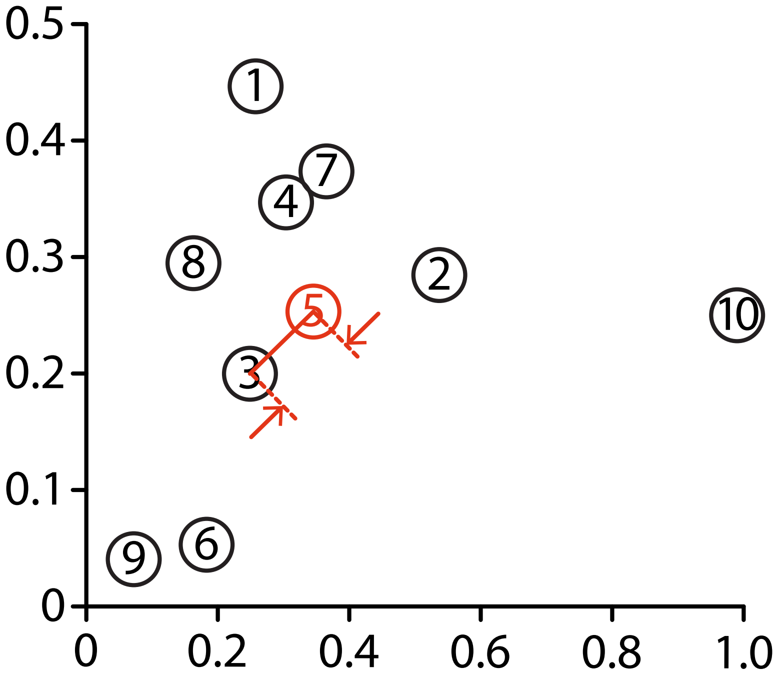
- The *k*th-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$



id	score
1	0.109

The *k*th-Nearest Neighbor Distance

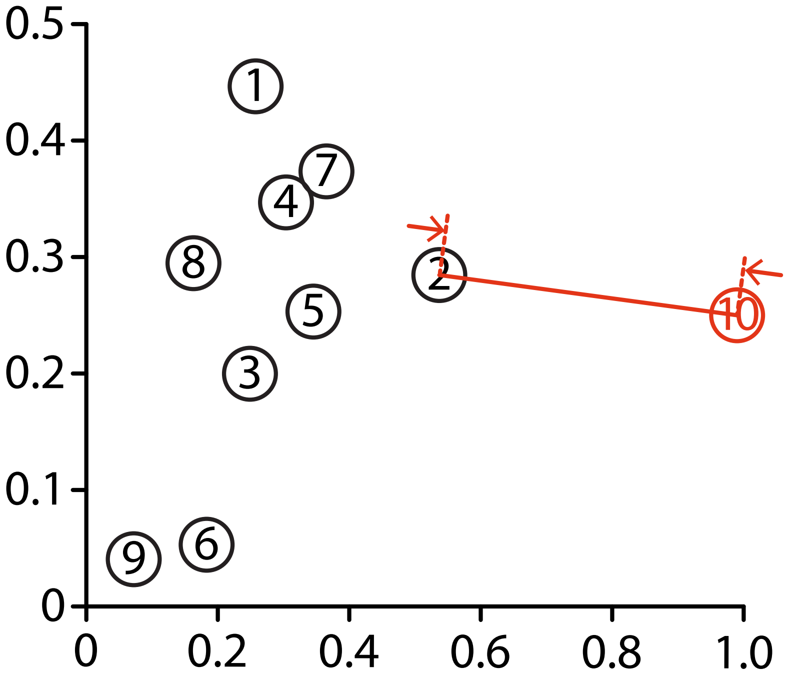
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$



id	score
1	0.109
5	0.103

The *k*th-Nearest Neighbor Distance

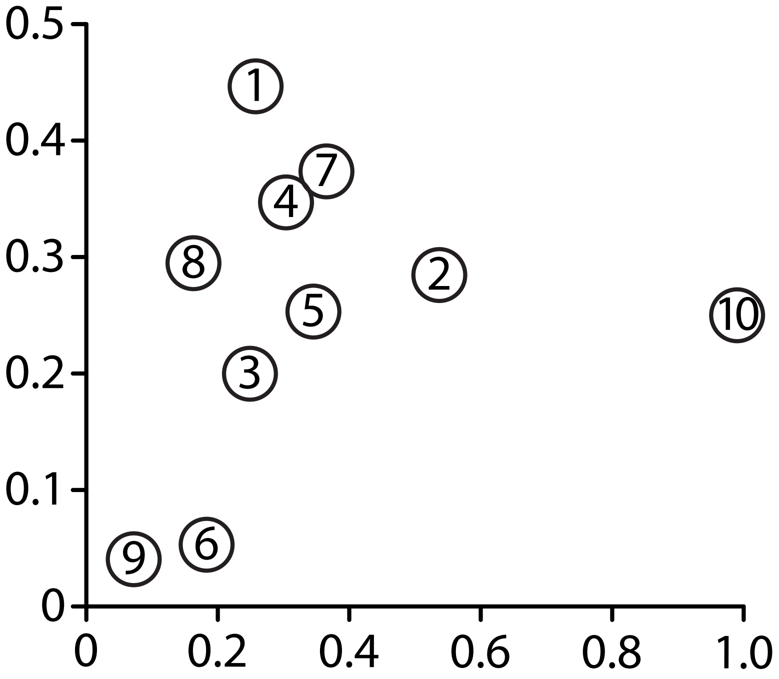
- The *k*th-NN score $q_{\mathrm{kthNN}}(x) := d^k(x; X)$



id	score
10	0.454
1	0.109
5	0.103

The *k*th-Nearest Neighbor Distance

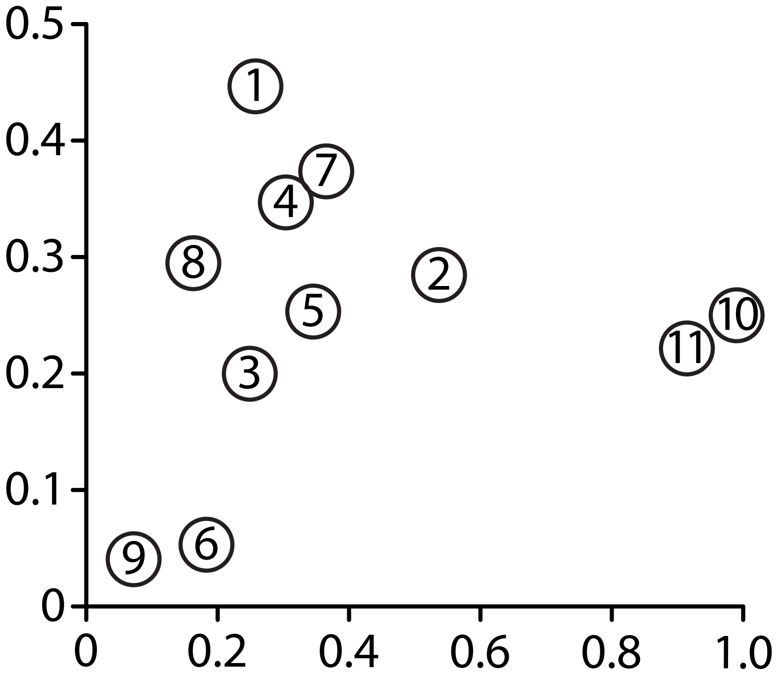
- The *k*th-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$



id	score
10	0.454
2	0.193
8	0.128
6	0.112
9	0.112
3	0.110
1	0.109
5	0.103
4	0.067
7	0.067

The *k*th-Nearest Neighbor Distance

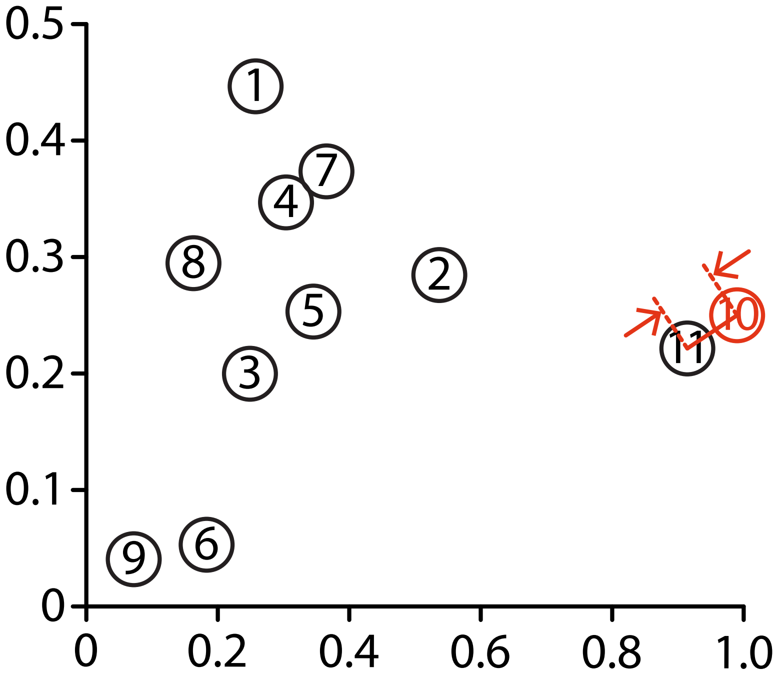
- The *k*th-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$



id	score
10	0.454
2	0.193
8	0.128
6	0.112
9	0.112
3	0.110
1	0.109
5	0.103
4	0.067
7	0.067

The *k*th-Nearest Neighbor Distance

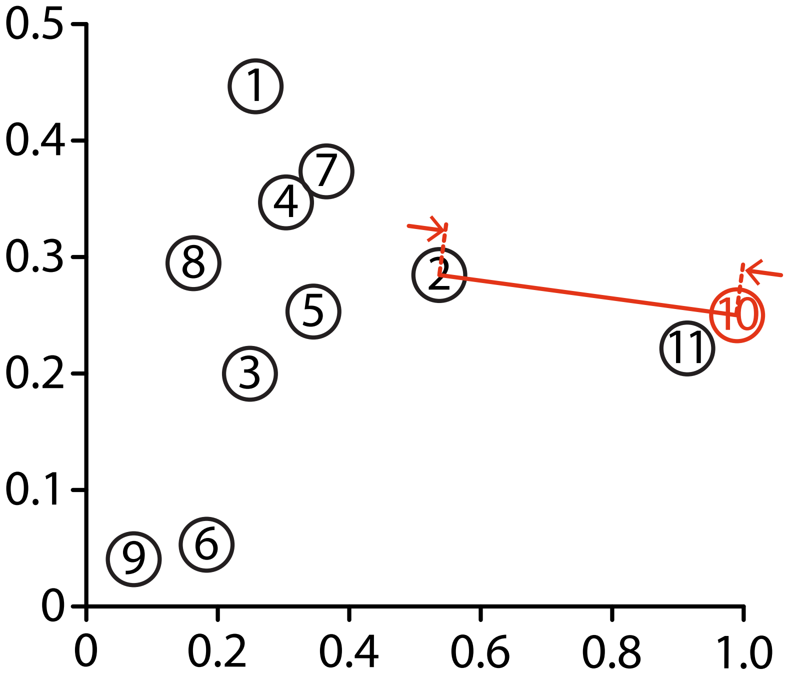
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$



id	score
2	0.193
8	0.128
6	0.112
9	0.112
3	0.110
1	0.109
5	0.103
4	0.067
7	0.067
10	0.028
11	0.028

The *k*th-Nearest Neighbor Distance

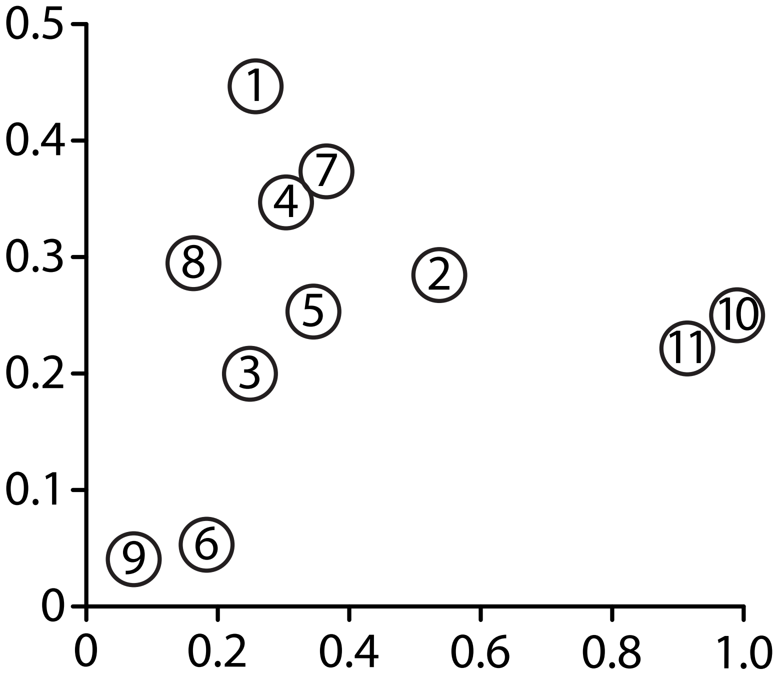
- The *k*th-NN score $q_{\mathrm{kthNN}}(x) := d^k(x; X)$



id	score
2	0.193
8	0.128
6	0.112
9	0.112
3	0.110
1	0.109
5	0.103
4	0.067
7	0.067
10	0.028
11	0.028

The *k*th-Nearest Neighbor Distance

- The *k*th-NN score $q_{\mathrm{kthNN}}(x) := d^k(x; X)$



id	score
10	0.454
11	0.436
9	0.238
2	0.194
6	0.161
8	0.150
1	0.130
3	0.128
7	0.122
5	0.110
4	0.103

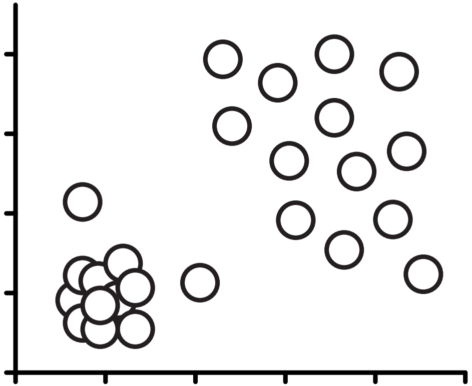
Connection with DB(α, δ)-Outliers

- The kth-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X
- Let $\alpha = (n - k)/n$
- For any threshold δ,
  the set of DB(α, δ)-outliers = $\{x \in X \mid q_{k\text{thNN}}(x) \geq \delta\}$

Two Drawbacks of the *k*th-NN Approach

1. Scalability; $O(n^2)$
   - **Solution:** Partial computation of the pairwise distances to compute scores only for the top-t outliers
     - ORCA [Bay & Schwabacher, KDD 2003], iORCA [Bhaduri et al., KDD 2011]

2. Detection ability
   - **Solution:** Introduce other definitions of the outlierness
     - Density-based (LOF) [Breunig et al. KDD 2000]
     - Angle-based (ABOD) [Kriegel et al. KDD 2008]



Partial Computation for Efficiency

- The key technique in retrieving top-t outliers: Approximate Nearest Neighbor Search (ANNS) principle
  - During computing $q_{k\text{thNN}}(x)$ within a `for` loop:

    $q_{k\text{thNN}}(x) = \infty \quad (k = 1 \text{ for simplicity})$
    `for each` $x' \in X \setminus \{x\}$
    `if` $d(x, x') < q_{k\text{thNN}}(x)$ then $q_{k\text{thNN}}(x) = d(x, x')$

    the current value $q_{k\text{thNN}}(x)$ is monotonically decreasing

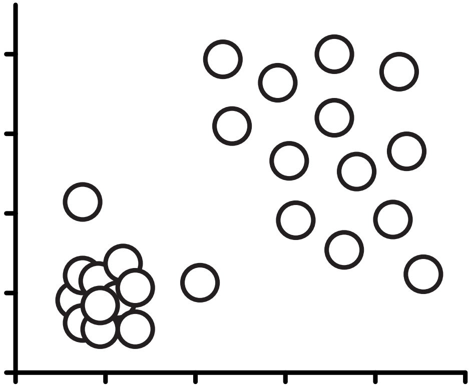
- In the `for` loop, if $q_{k\text{thNN}}(x)$ becomes smaller than the mth largest score so far, x never becomes an outlier
  - The `for` loop can be terminated earlier

Further Pruning with Indexing

- iORCA employed an indexing technique
  - Bhaduri, K., Matthews, B.L., Giannella, C.R., "Algorithms for speeding up distance-based outlier detection", SIGKDD 2011
- Select a point $r \in X$ randomly (reference point)
- Re-order the dataset X with increasing distance from r
- ***If*** $d(x, r) + q_{\mathrm{kthNN}}(r) < c$***, x never be an outlier***
  - c is the cutoff, the m-th largest score so far
- Drawback: the efficiency strongly depends on m

Two Drawbacks of the *k*th-NN Approach

1. Scalability; $O(n^2)$
   - **Solution:** Partial computation of the pairwise distances to compute scores only for the top-*t* outliers
     - ORCA [Bay & Schwabacher, KDD 2003], iORCA [Bhaduri et al., KDD 2011]
2. Detection ability
   - **Solution:** Introduce other definitions of the outlierness
     - Density-based (LOF) [Breunig et al. KDD 2000]
     - Angle-based (ABOD) [Kriegel et al. KDD 2008]



LOF (Local Outlier Factor) (1/2)

- $N^k(x)$: the set of kNNs of x
- Reachability distance $\mathrm{Rd}(x; x') = \max\left\{d^k(x', X), d(x, x')\right\}$

LOF (Local Outlier Factor) (2/2)

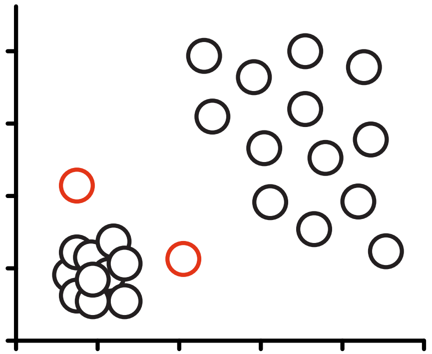
- Local reachability density is

$$\Delta(x) := \left(\frac{1}{|N^k(x)|} \sum_{x' \in N^k(x)} \mathrm{Rd}(x; x') \right)^{-1}$$

- The LOF of x is defined as

$$\mathrm{LOF}(x) := \frac{\left(1/|N^k(x)|\right) \sum_{y \in N^k(x)} \Delta(y)}{\Delta(x)}$$

- The ratio of the local reachability density of x and the average of the local reachability densities of its kNNs

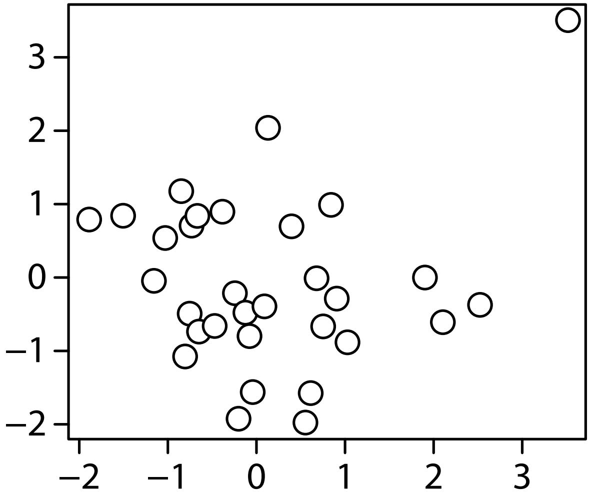


LOF is Popular

- LOF is one of the most popular outlier detection methods
  - Easy to use (only one parameter k)
  - Higher detection ability than kth-NN
- The main drawback: scalability
  - $O(n^2)$ is needed for neighbor search
  - Same as kth-NN

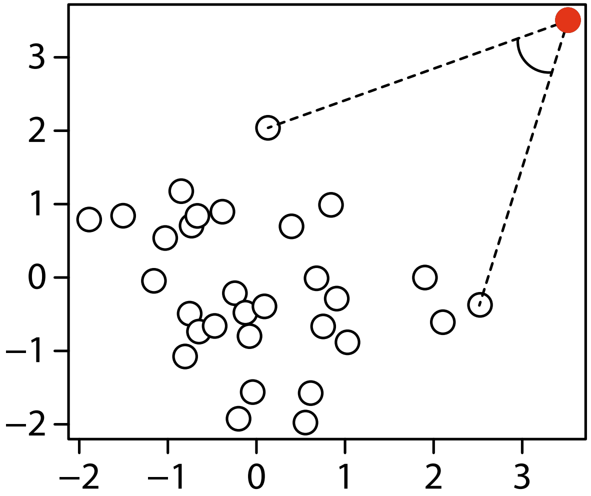
ABOD (Angle-Based Outlier Detection)

- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small



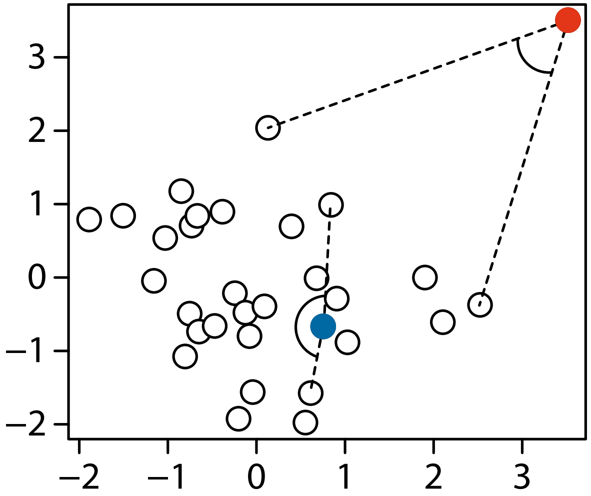
ABOD (Angle-Based Outlier Detection)

- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small



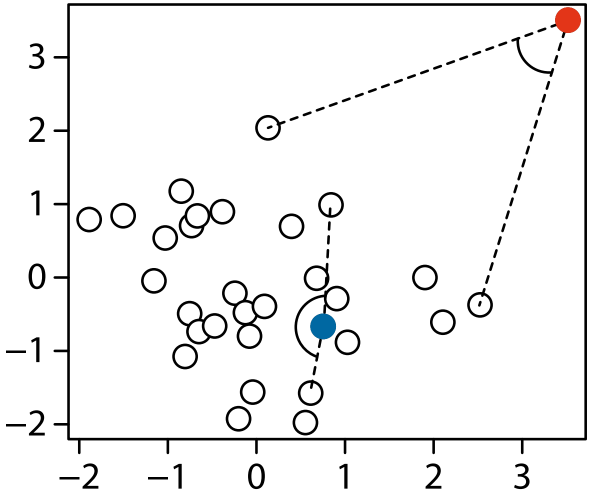
ABOD (Angle-Based Outlier Detection)

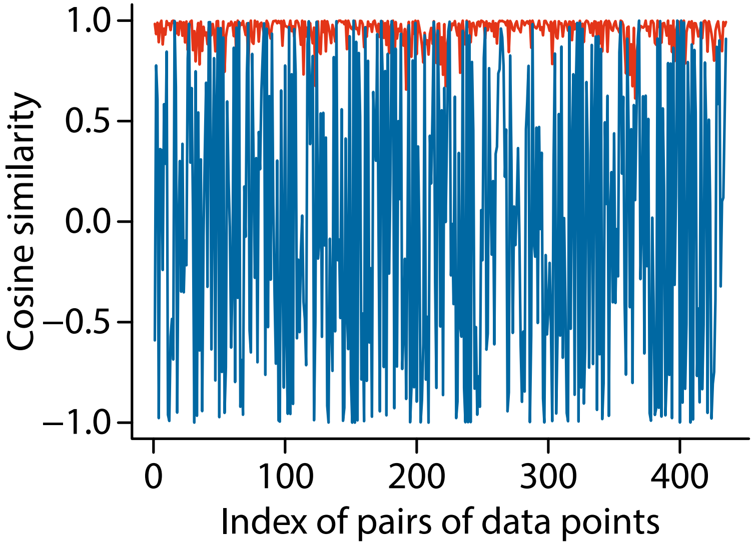
- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small



ABOD (Angle-Based Outlier Detection)

- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small





Definition of ABOD

- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small
- The score $\text{ABOF}(x) := \text{Var}_{y,z \in X}\, s(y - x, z - x)$
  - $s(x, y)$ is the similarity between vectors x and y, e.g. the cosine similarity
  - $s(z - x, y - x)$ correlates with the angle of y and z w.r.t. the coordinate origin x
- Pros: Parameter-free
- Cons: High computational cost $O(n^3)$

Speeding Up ABOD

- Pham and Pagh proposed a fast approximation algorithm FastVOA
  - Pham, N., Pagh, R., "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data", SIGKDD 2012
  - It estimates the first and the second moment of the variance $\mathrm{Var}_{y,z \in X} s(y - x, z - x)$ independently using random projections and AMS sketches
- Pros: near-linear complexity: $O(tn(m + \log n + c_1 c_2))$
  - t: the number of hyperplanes for random projections
  - c_1, c_2: the number of repetitions for AMS sketches
- Cons: Many parameters

Other Interesting Approaches

- iForest (isolation forest)
  - Liu, F.T. and Ting, K.M. and Zhou, Z.H., “Isolation forest”, ICDM 2008
  - A random forest-like method with recursive partitioning of datasets
  - An outlier tends to be easily partitioned

- One-class SVM
  - Schölkopf, B. et al., “Estimating the support of a high-dimensional distribution”, Neural computation (2001)
  - This classifies objects into inliers and outliers by introducing a hyperplane between them
  - This can be used as a ranking method by considering the signed distance to the separating hyperplane

iForest (Isolation Forest)

- Given X, we construct an *i*Tree:
  (i) X is partitioned into X_{L} and X_{R} such that:
  $X_{\mathrm{L}} = \{ x \in X \mid x_q < v \},\ X_{\mathrm{R}} = X \setminus X_{\mathrm{L}}$,
  where v and q are randomly chosen
  (ii) Recursively apply to each set until it becomes a singleton
  – Can be combined with sampling

- The outlierness score *i*Tree(x) is defined as $2^{-\overline{h(x)}/c(\mu)}$
  – $h(x)$ is the number of edges from the root to the leaf of x
  – $\overline{h(x)}$ is the average of $h(x)$ on t *i*Trees
  – $c(\mu) := 2H(\mu - 1) - 2(\mu - 1)/n$ (H is the harmonic number)

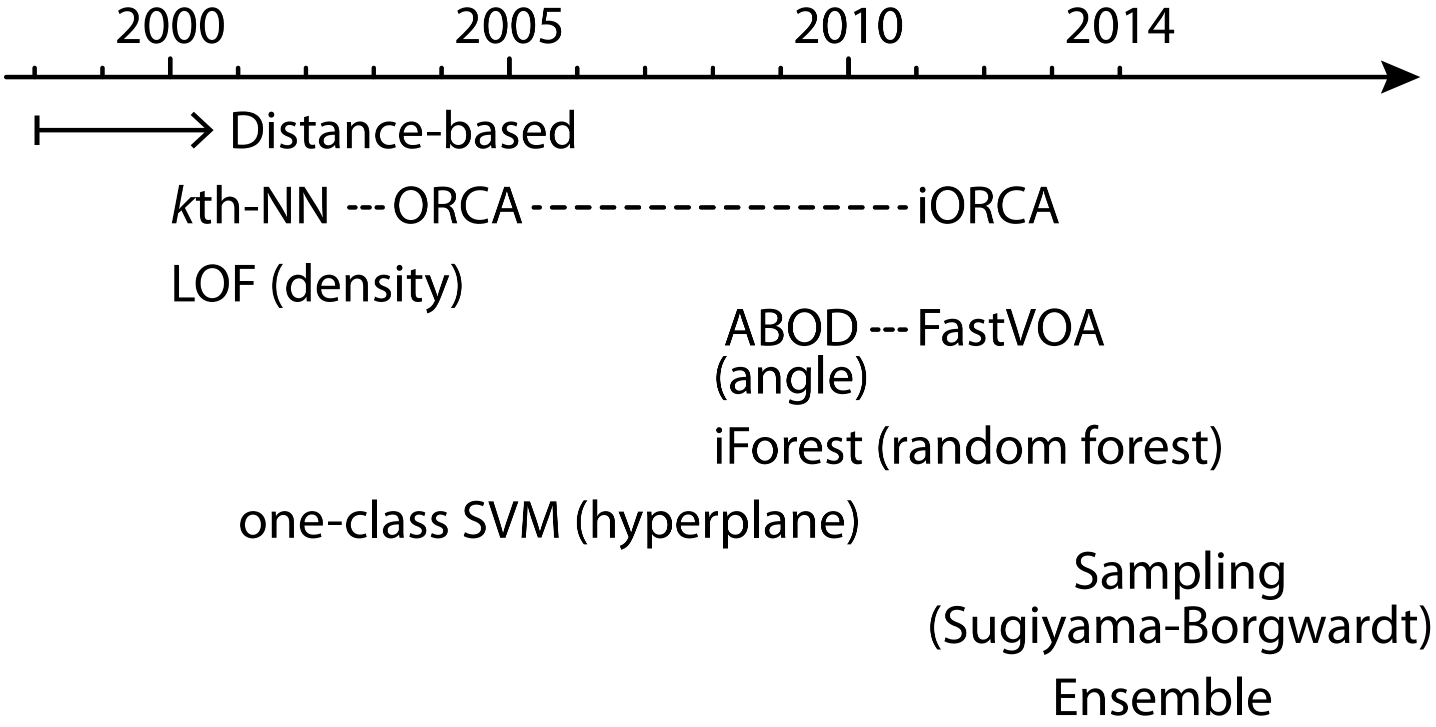
One-class SVM

- A technique via hyperplanes by Schölkopf *et al.*
- The score of a vector **x** is $\rho - (w \cdot \Phi(\mathbf{x}))$
  - Φ: a feature map
  - w and ρ are the solution of the following quadratic program:

$$\min_{w \in F, \xi \in \mathbb{R}^n, \rho \in \mathbb{R}} \frac{1}{2} ||w||^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho$$

    subject to $(w \cdot \Phi(x_i)) \geq \rho - \xi_i, \ \xi_i \geq 0$
  - The term $w \cdot \Phi(\mathbf{x})$ can be replaced with $\sum_{i=1}^{n} a_i k(\mathbf{x}_i, \mathbf{x})$ using a kernel function k

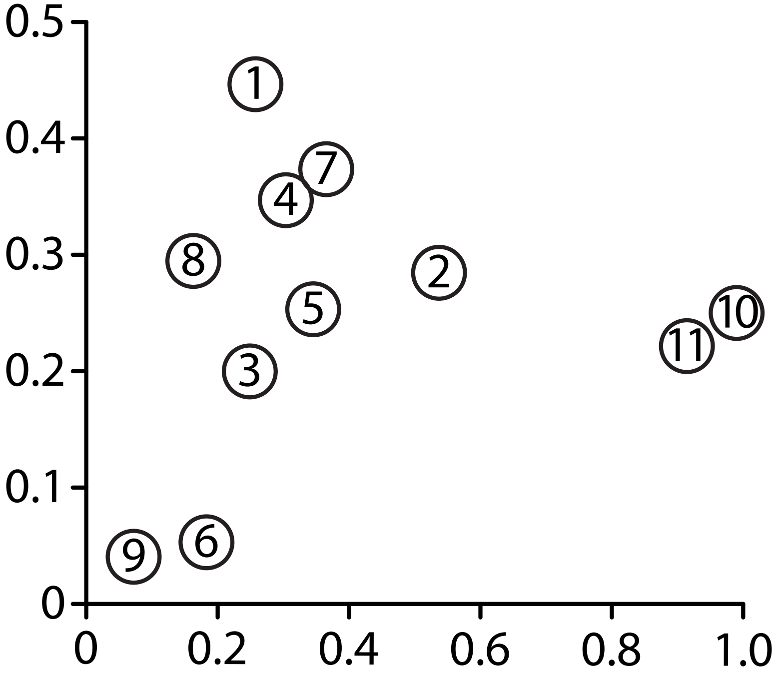
Timeline


2000
2005
2010
2014
Distance-based
*k*th-NN --- ORCA ------------------ iORCA
LOF (density)
ABOD --- FastVOA
(angle)
iForest (random forest)
one-class SVM (hyperplane)
Sampling
(Sugiyama-Borgwardt)
Ensemble

Outlier Detection via Sampling

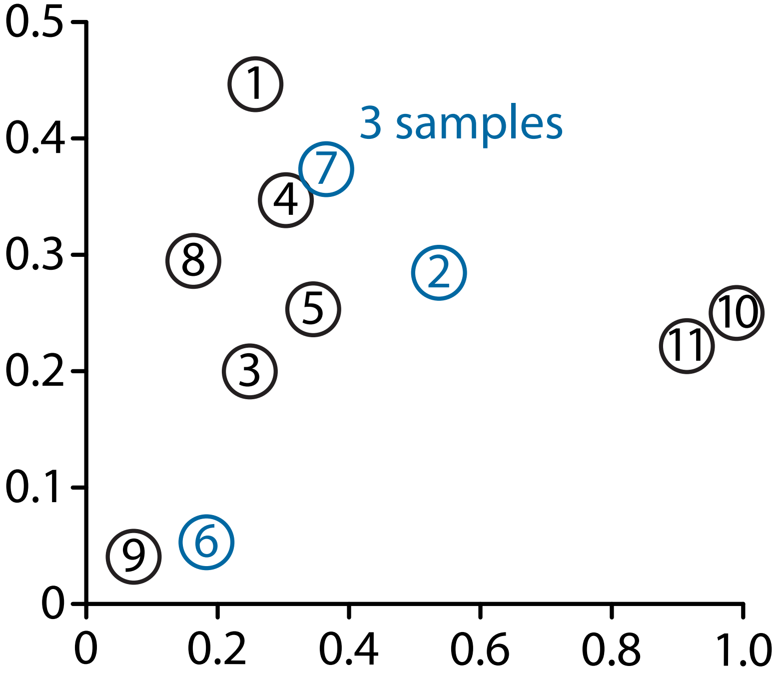
- (Sub-)Sampling was largely ignored in outlier detection
  - Find outliers from samples seems hopeless
- Use samples as a reference set
  - Sugiyama, M., Borgwardt, K.M., "Rapid Distance-Based Outlier Detection via Sampling", NIPS 2013
  - Sample size is surprisingly small, which is sometimes 0.0001% of the total number of data points
  - Accuracy is competitive with state-of-the-art methods
- Ensemble methods are recently emerging
  - Aggarwal, C.C., Outlier Ensembles: An Introduction, Springer (2017)

Sugiyama-Borgwardt Method



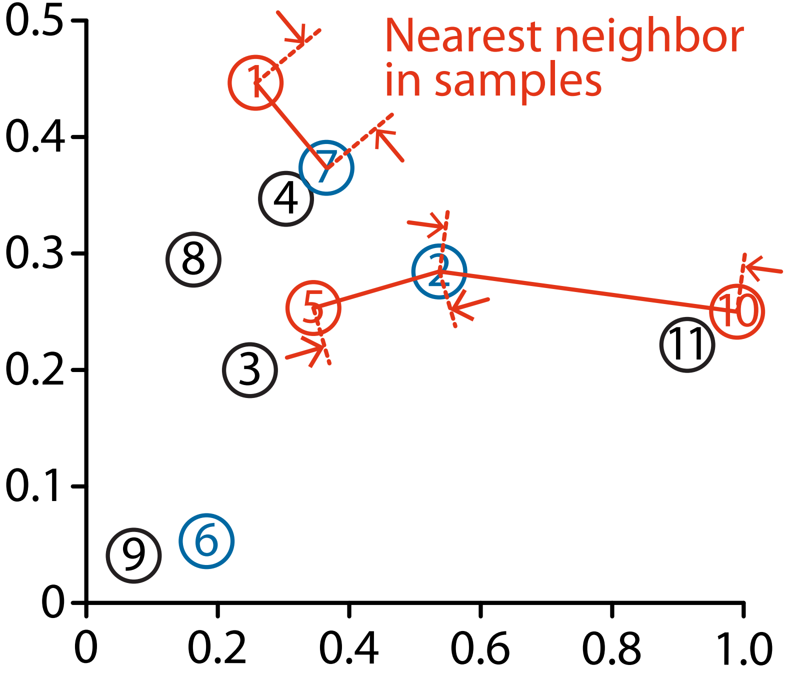
id	score

Sugiyama-Borgwardt Method



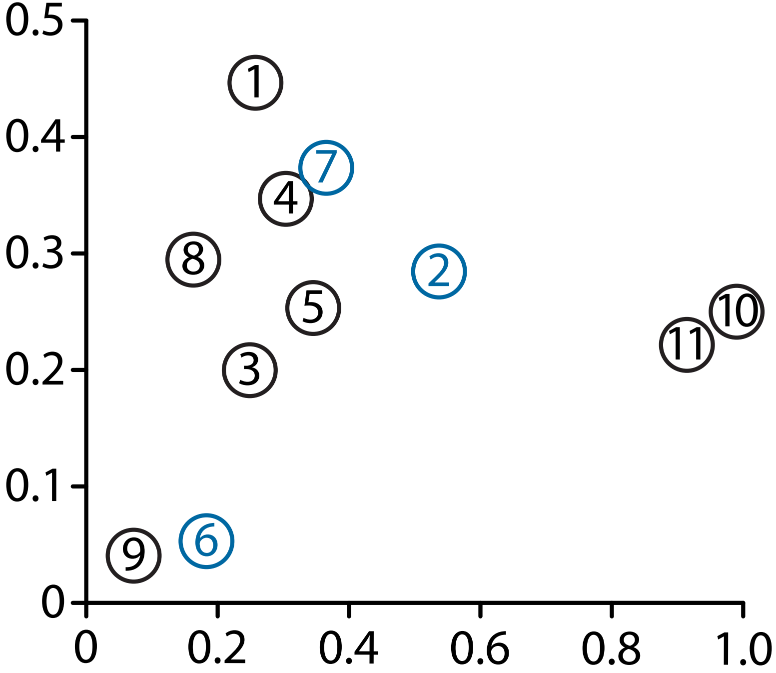
id	score

Sugiyama-Borgwardt Method



id	score
10	0.454
1	0.130
5	0.122

Sugiyama-Borgwardt Method



id	score
10	0.454
11	0.436
6	0.369
8	0.217
2	0.193
7	0.193
3	0.161
1	0.130
5	0.122
9	0.112
4	0.067

Definition

- Given a dataset X (n data points, m dimensions)
- Randomly and independently sample a subset $S(X) \subset X$
- Define the score $q_{\mathrm{Sp}}(x)$ for each object $x \in X$ as

  $$q_{\mathrm{Sp}}(x) := \min_{x' \in S(X)} d(x, x')$$

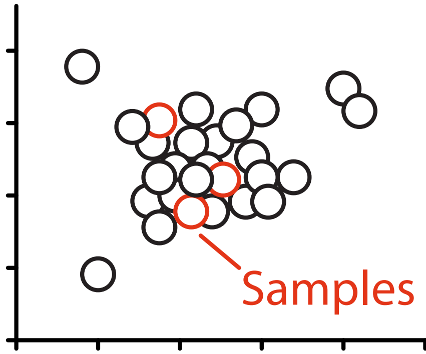
  - Input parameter: the number of samples $s = |S(X)|$
  - The time complexity is $\Theta(nms)$ and
    the space complexity is $\Theta(ms)$

Intuition

- Outliers should be significantly different from almost all inliers
  → A sample set includes only inliers with high probability
  → Outliers get high scores

- For each inlier, at least one similar data point is included in the sample set with high probability

- This scheme is expected to work with small sample sizes
  - If we pick up too many samples, some rare points, similar to an outlier, slip into the sample set



Notations

- $X(\alpha;\delta)$: the set of Knorr and Ng's DB(α,δ)-outliers
- $x \in X(\alpha;\delta)$ if $|\{\, x' \in X \mid d(x,x') > \delta \,\}| \geq \alpha n$
  - $\overline{X}(\alpha;\delta) = X \setminus X(\alpha;\delta)$: the set of inliers
  - α is expected to close to 1, meaning that an outlier is distant from almost all points
- Define β $(0 \leq \beta \leq \alpha)$ as the minimum value s.t.

  $\forall x \in \overline{X}(\alpha;\delta),\ \left|\{\, x' \in X \mid d(x,x') > \delta \,\}\right| \leq \beta n$

Theoretical Results

1. For $x \in X(\alpha;\delta)$ and $x' \in \overline{X}(\alpha;\delta)$,

   $$\Pr\left(q_{\mathrm{Sp}}(x) > q_{\mathrm{Sp}}(x')\right) \geq \alpha^{s}(1-\beta^{s})$$

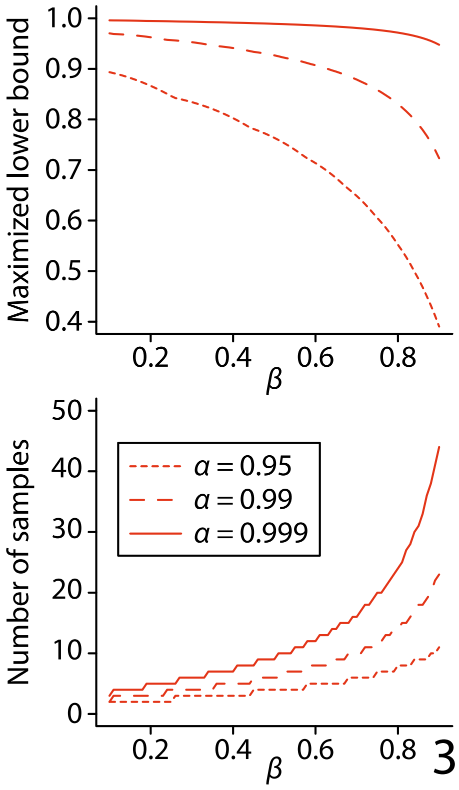
   (s is the number of samples)

   - This lower bound tends to be high in a typical setting (α is large, β is moderate)

2. This bound is maximized at

   $$s = \log_{\beta} \frac{\log \alpha}{\log \alpha + \log \beta}$$

   - This value tends to be small



Evaluation criteria

- Precision v.s. Recall (Sensitivity)
  - Recall = TP / (TP + FN)
  - Precision = TP / (TP + FP)
- Effectiveness is usually measured by AUPRC (area under the precision-recall curve)
  - Equivalent to the average precision over all possible cut-offs on the ranking of outlierness
- cf. ROC curve: False Positive Rate (FPR) v.s. Sensitivity
  - FPR = FP / (FP + TN) = 1 – Specificity
  - Sensitivity = TP / (TP + FN)

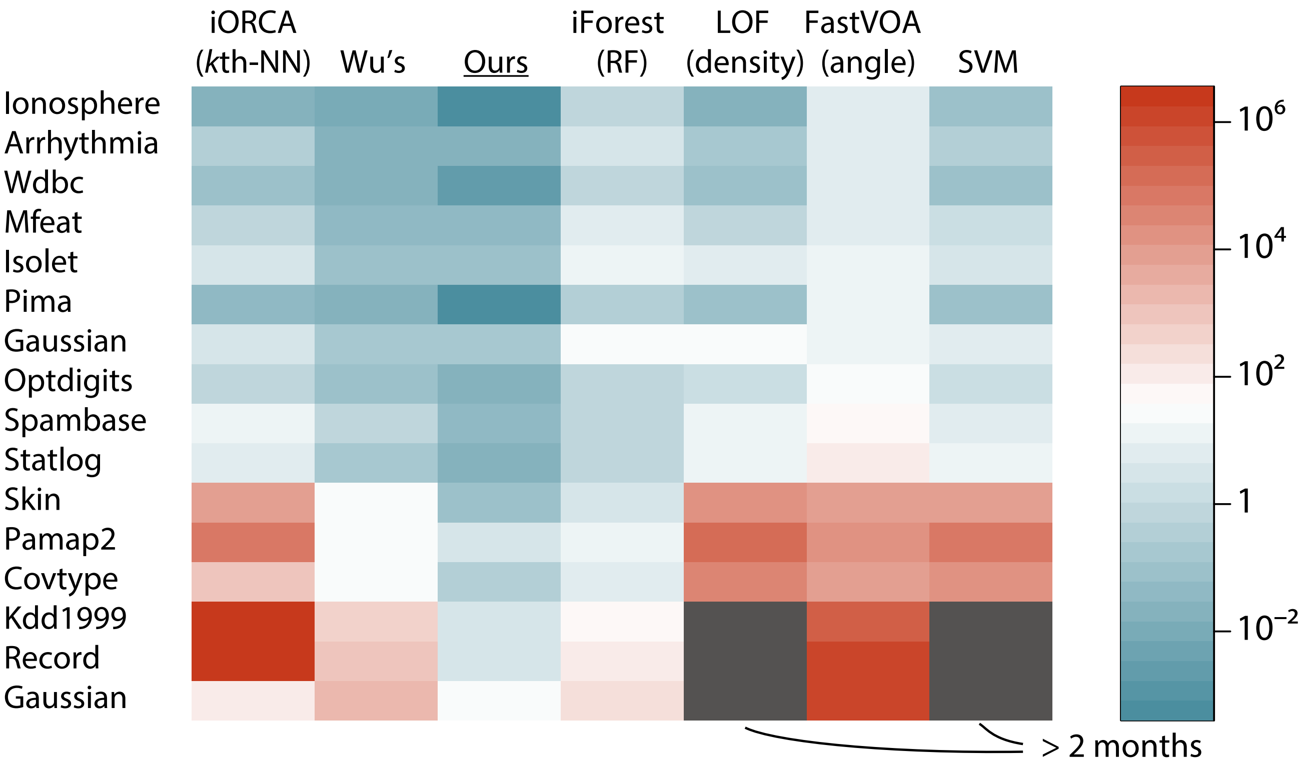
Relationship

	Ground truth		
	Condition Positive	Condition Negative	
Test Outcome Positive	True Positive	False Positive (Type I Error)	Precision TP / (TP + FP)
Test Outcome Negative	False Negative (Type II Error)	True Negative	
	Sensitivity (Recall) TP / (TP + FN)	Specificity TN / (FP + TN) = 1 – FPR False Positive Rate (FPR) FP / (FP + TN)	

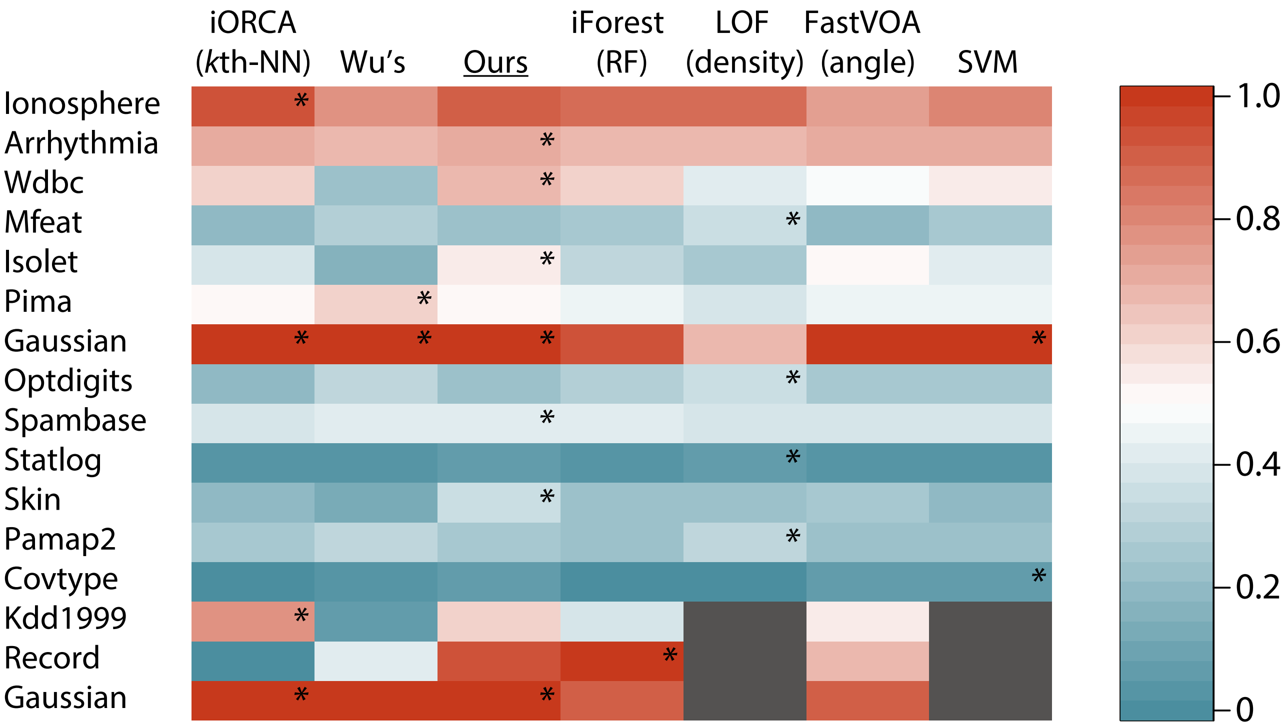
Datasets for Experiments (* are synthetic data)

	# of objects	# of outliers	# of dims
Ionosphere	351	126	34
Arrhythmia	452	207	274
Wdbc	569	212	30
Mfeat	600	200	649
Isolet	960	240	617
Pima	768	268	8
Gaussian*	1000	30	1000
Optdigits	1688	554	64
Spambase	4601	1813	57
Statlog	6435	626	36
Skin	245057	50859	3
Pamap2	373161	125953	51
Covtype	286048	2747	10
Kdd1999	4898431	703067	6
Record	5734488	20887	7
Gaussian*	10000000	30	20

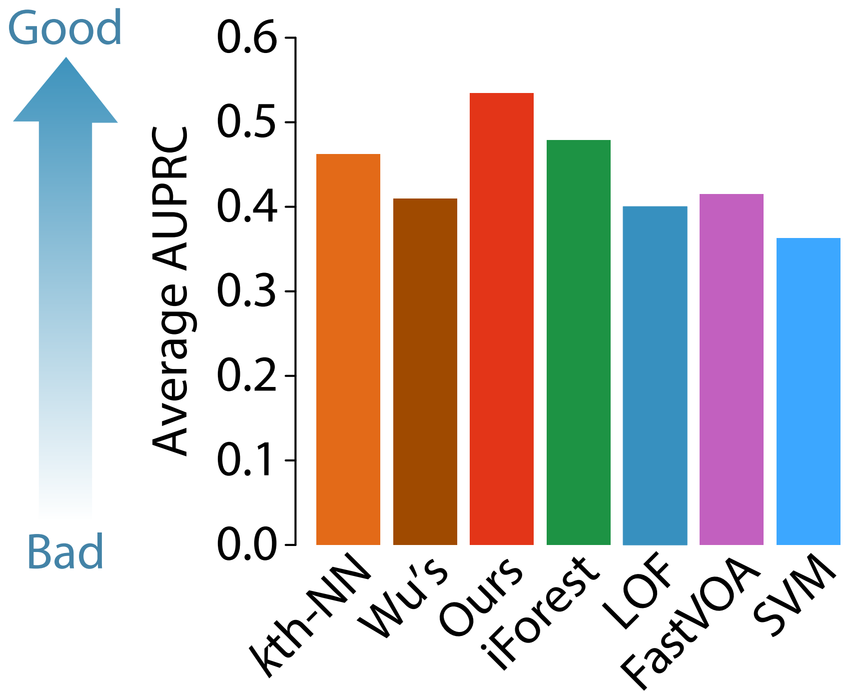
Runtime (seconds)



AUPRC (* shows the best score)



Average of AUPRC over all datasets

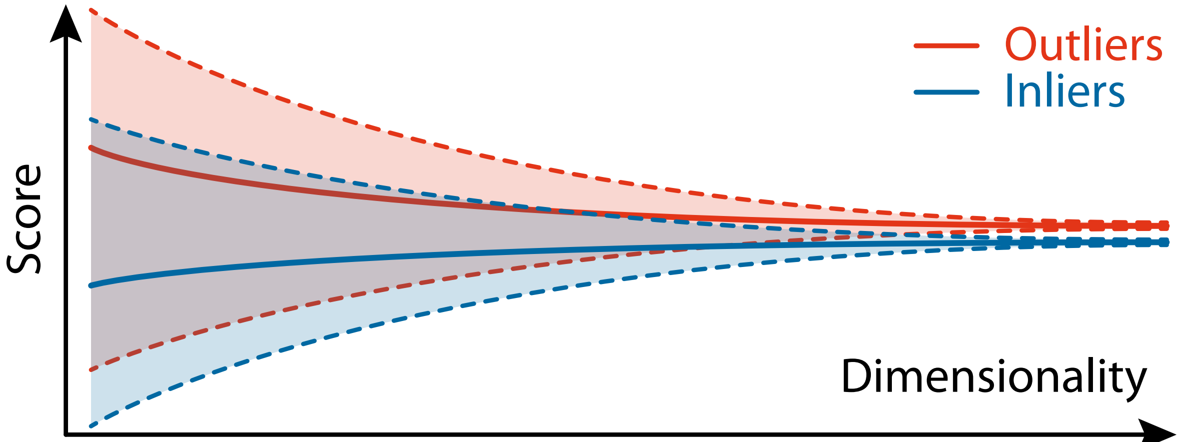


How about High-dimensional Data ?

- So-called “the curse of dimensionality”
- There is an interesting paper that studies outlier detection in high-dimensional data
  - Zimek, A., Schubert, E., Kriegel, H.-P., “A survey on unsupervised outlier detection in high-dimensional numerical data”, Statistical Analysis and Data Mining (2012)

Fact about High-Dimensional Data

- High-dimensionality is not always the problem
  - If all attributes are relevant, detecting outliers becomes easier and easier as attributes (dimensions) increases
  - Of course, it is not the case if irrelevant attributes exist



When Data Is Supervised

- First choice: Optimize parameters by cross validation
  - Sample size in Sugiyama-Borgwardt method
  - Determine the threshold for outliers from rankings
- Classification methods can be used, but it is generally difficult as positive and negative data are unbalanced

Summary

- *k*th-NN method is the standard
- If there are different density regions, LOF is recommended
- The most advanced (yet simple) method is the sampling-based method
  - sampling is a powerful tool in outlier detection