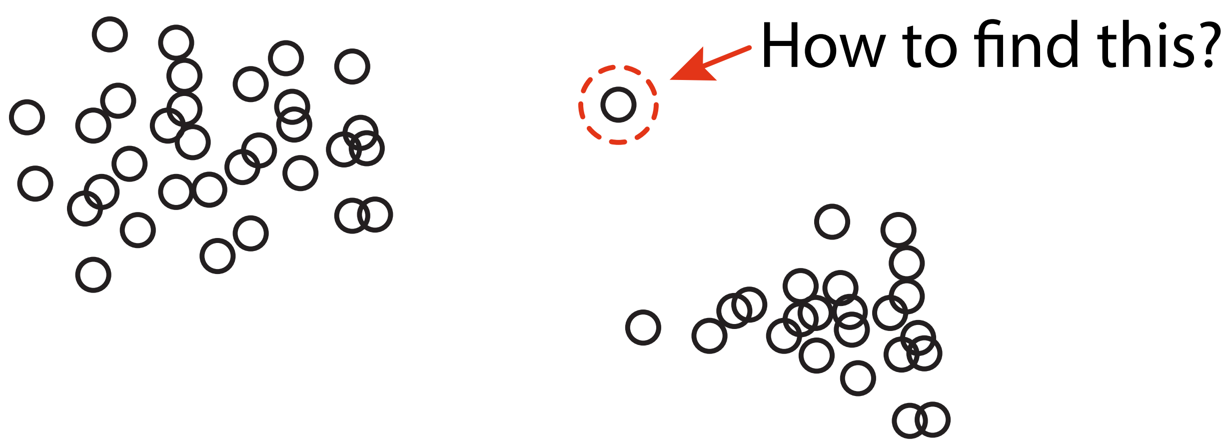
June 16, 2014
Kyoto University
Informatics Seminar

Distance-Based Outlier Detection via Sampling

ISIR, Osaka University
Mahito Sugiyama

Overview

- 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) ?
- I will talk about recent advances in distance-based outlier detection methods



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

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 [Aggrawal, 2013]
  - One of the most popular approaches in outlier detection
    - Distribution-free
    - Easily applicable for various types of data

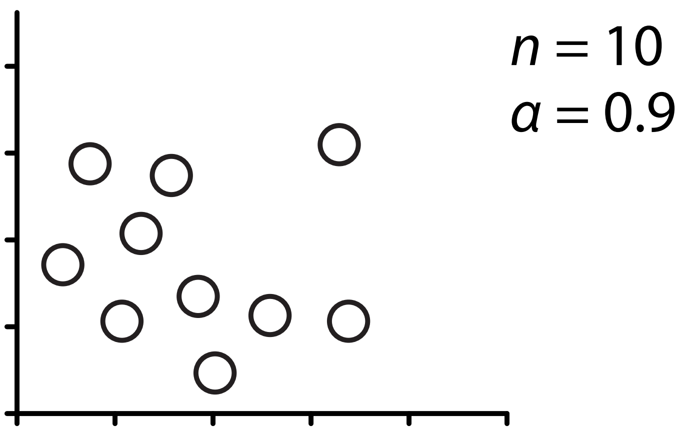
Distance-Based Outlier Detection

- The modern distance-based approach
  - A data point is an outlier, if its locality is sparsely populated [Aggrawal, 2013]
  - 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]

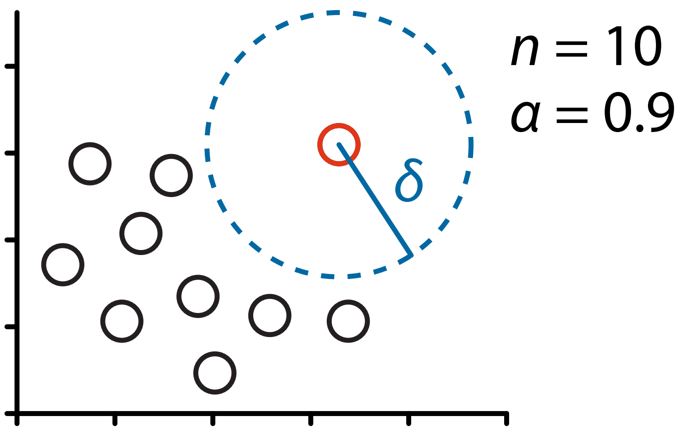
The First Distance-Based Method

- 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
- 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



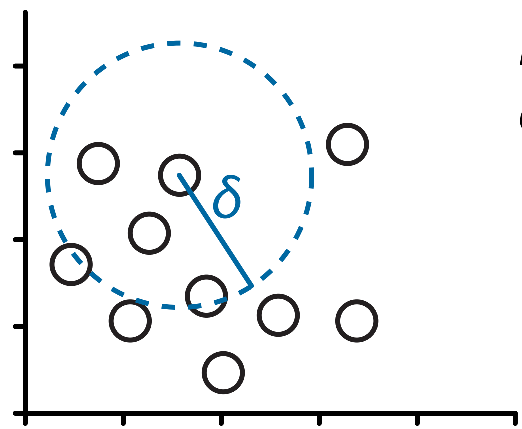
The First Distance-Based Method

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



$n = 10$
$\alpha = 0.9$

From Classification to Ranking

- Two drawbacks of DB(α, δ)-outliers
  1. Setting the distance threshold δ is difficult in practice
     - Setting α is not so difficult since it is always close to 1
  2. 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
  - The most basic distance-based approach to date

From Classification to Ranking

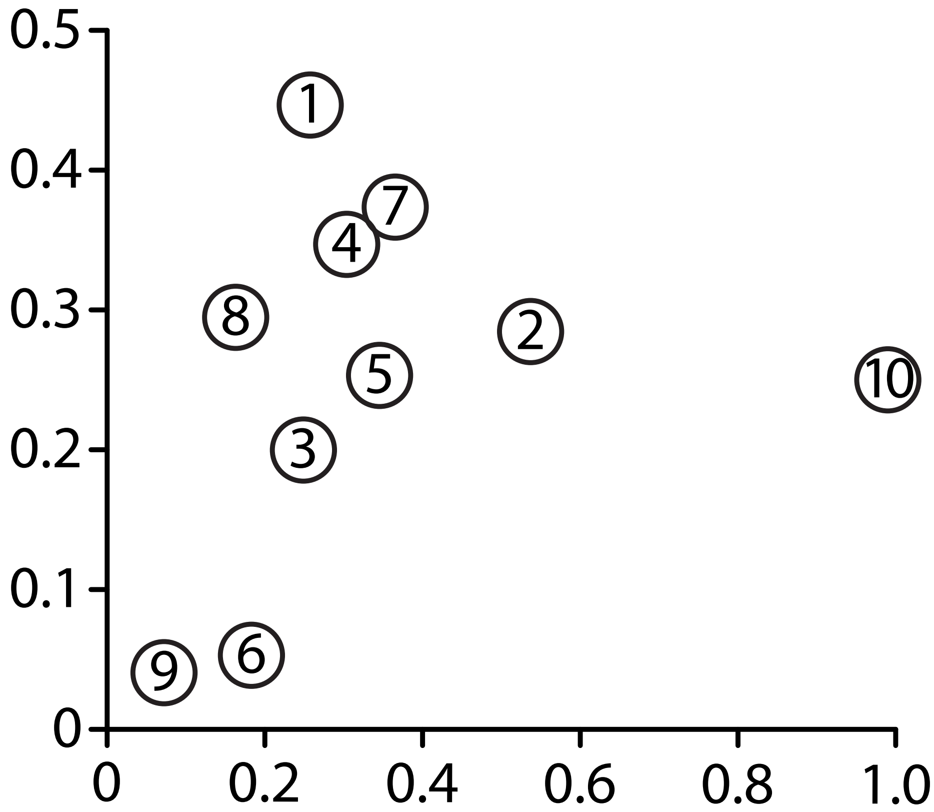
- Two drawbacks of DB(α, δ)-outliers
  1. Setting the distance threshold δ is difficult in practice
     - Setting α is not so difficult since it is always close to 1
  2. The lack of a ranking of outliers
- Ramaswamy *et al.* proposed to measure the outlierness by the *k*th-nearest neighbor (*k*th-NN) distance
  - Ramaswamy, S., Rastogi, R., Shim, K., “Efficient algorithms for mining outliers from large data sets”, SIGMOD 2000
  - The most basic distance-based approach to date
- ***From this study, the task of DB outlier detection becomes a ranking problem***
  - ***do not perform binary classification***

The *k*th-Nearest Neighbor Distance

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

The *k*th-Nearest Neighbor Distance

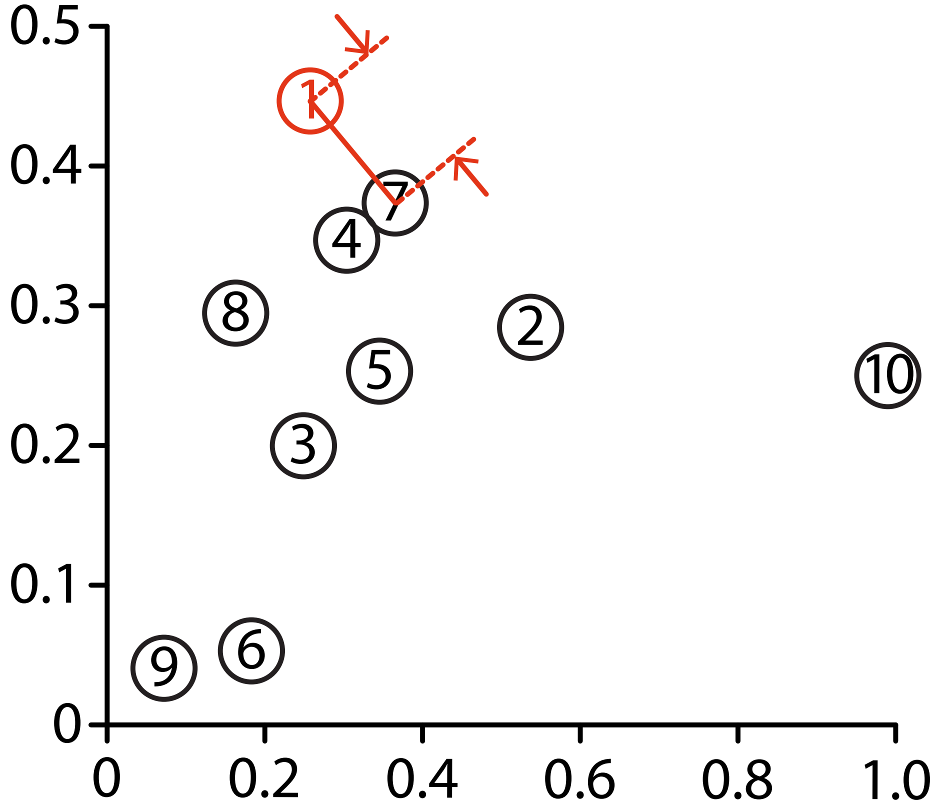
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X



id	score

The *k*th-Nearest Neighbor Distance

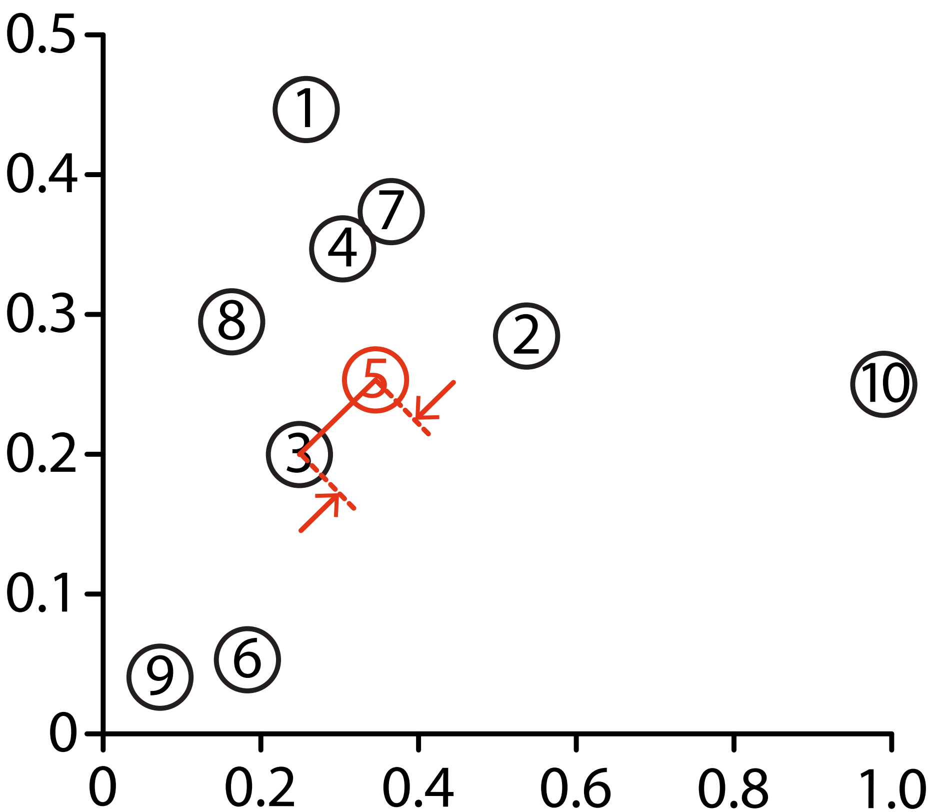
- The kth-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X



id	score
1	0.109

The *k*th-Nearest Neighbor Distance

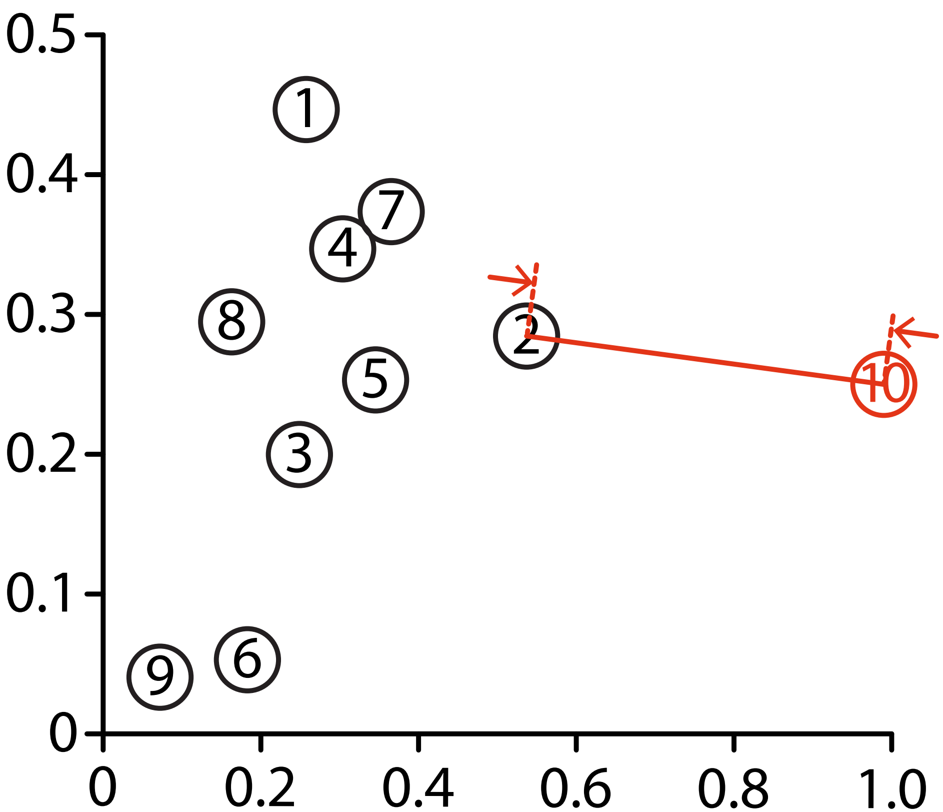
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X



id	score
1	0.109
5	0.103

The *k*th-Nearest Neighbor Distance

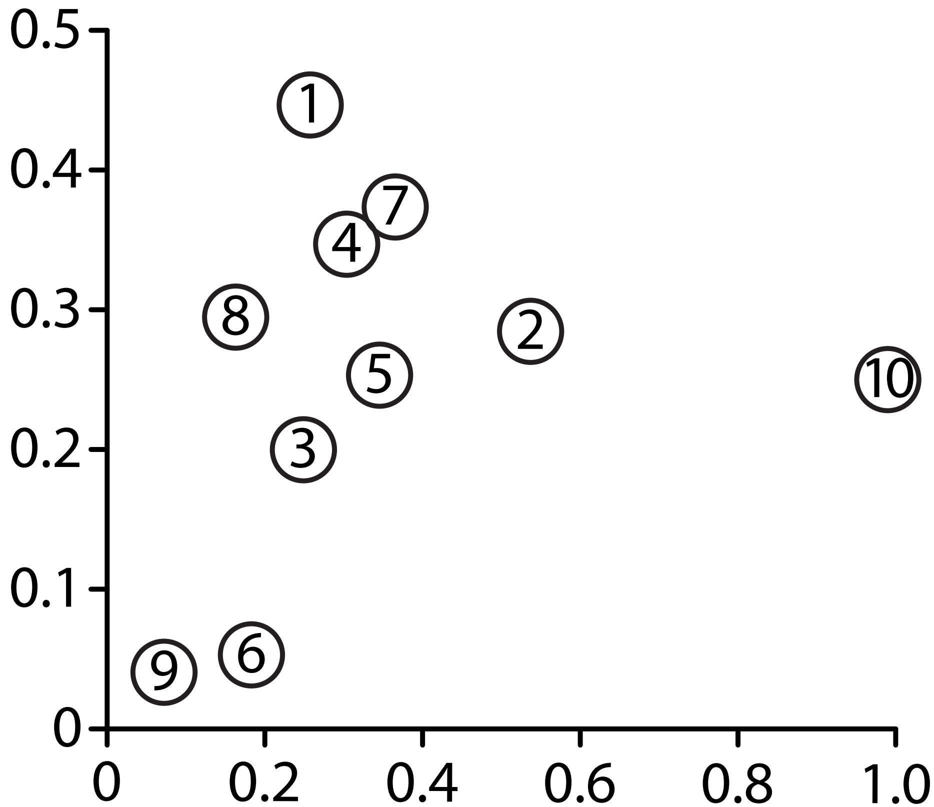
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X



id	score
10	0.454
1	0.109
5	0.103

The *k*th-Nearest Neighbor Distance

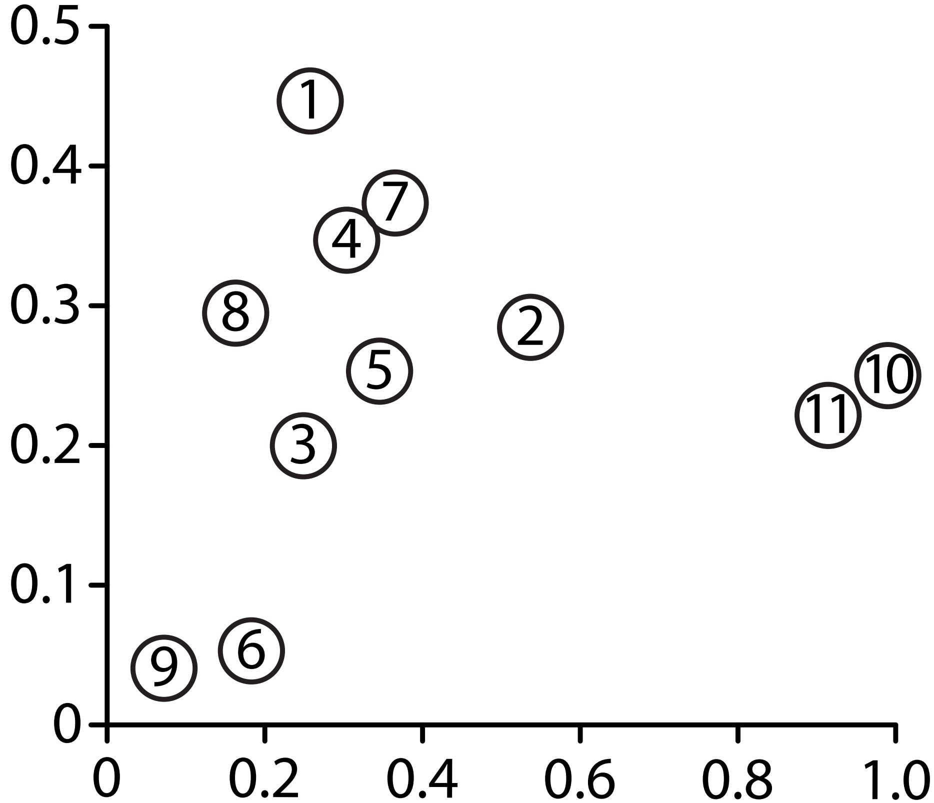
- The *k*th-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X



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

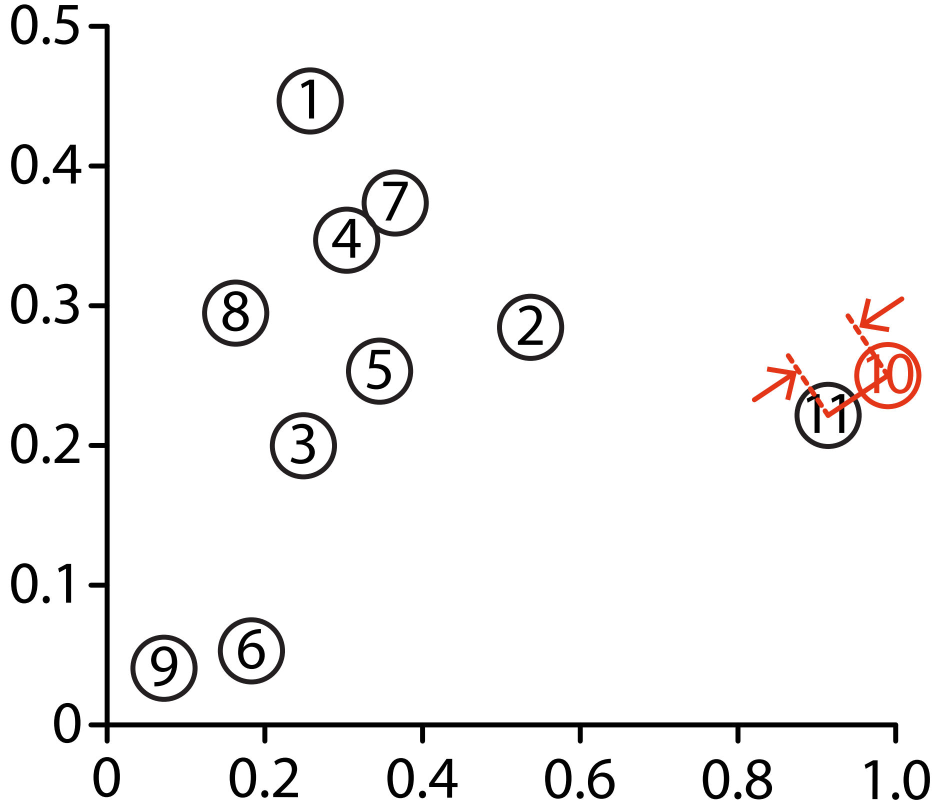
- The kth-NN score $q_{k\text{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X



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

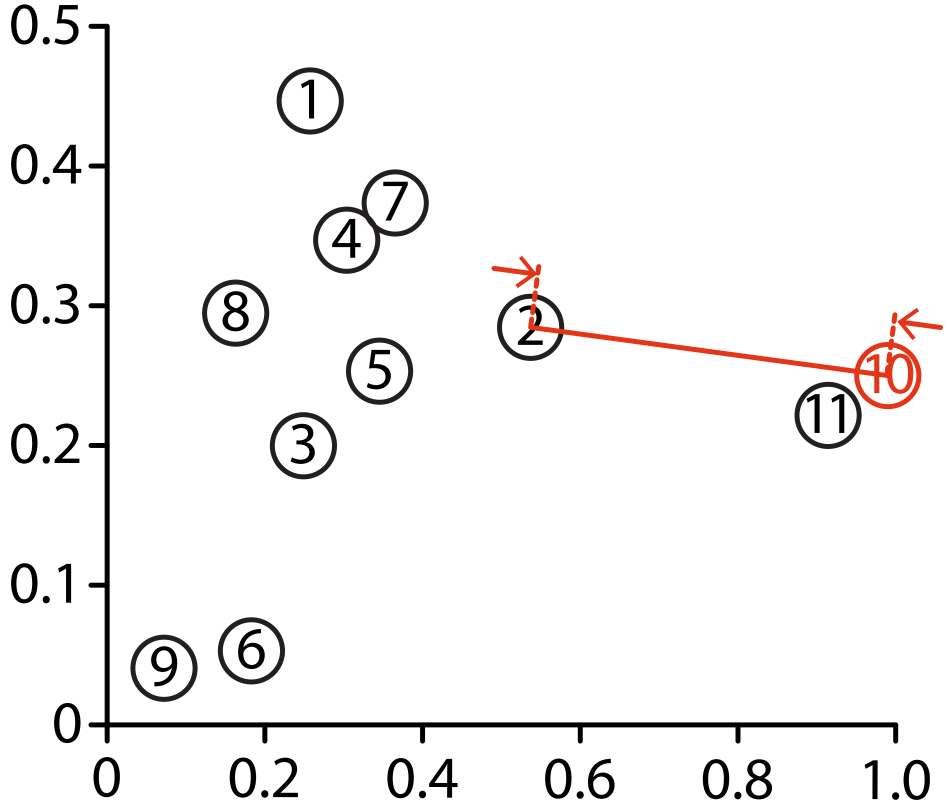
- The *k*th-NN score $q_{\mathrm{kthNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its *k*th-NN in X



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 kth-Nearest Neighbor Distance

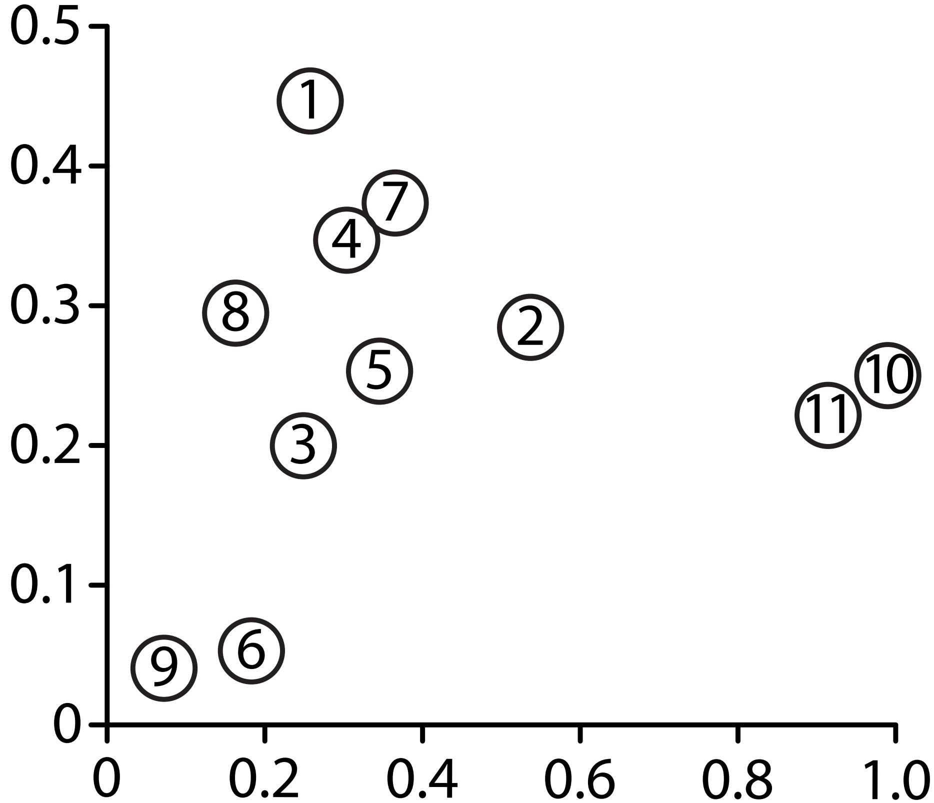
- The kth-NN score $q_{\mathrm{kthNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X



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 kth-Nearest Neighbor Distance

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

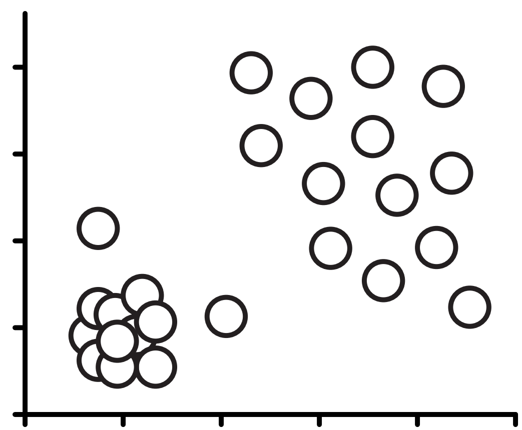
- The kth-NN score $q_{k\mathrm{thNN}}(x) := d^k(x; X)$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X
- Let $\alpha = (n - k)/n$
- For any threshold δ,
  the set of DB(α, δ)-outliers $= \{x \in X \mid q_{k\mathrm{thNN}}(x) \geq \delta\}$

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, SIGKDD 2003]
     - iORCA [Bhaduri et al., SIGKDD 2011]

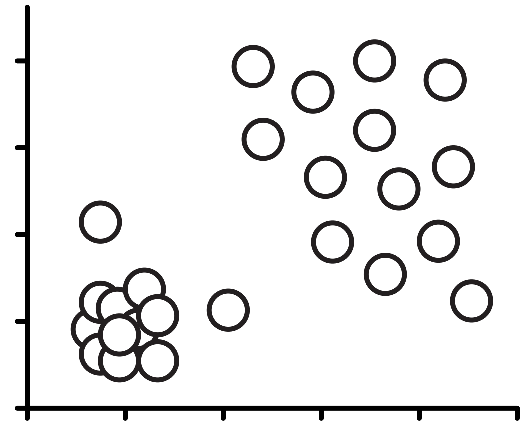
Two Drawbacks of the *k*th-NN Approach

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



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, SIGKDD 2003]
     - iORCA [Bhaduri et al., SIGKDD 2011]
2. Detection ability
   - **Solution:** Introduce other definitions of the outlierness
     - Density-based (LOF) [Breunig et al. SIGKDD 2000]
     - Angle-based (ABOD) [Kriegel et al. SIGKDD 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_kthNN(x) = ∞    (k = 1 for simplicity)
for each x' ∈ X \ {x}
  if d(x, x') < q_kthNN(x)
    q_kthNN(x) = d(x, x')
  end if
end for
```

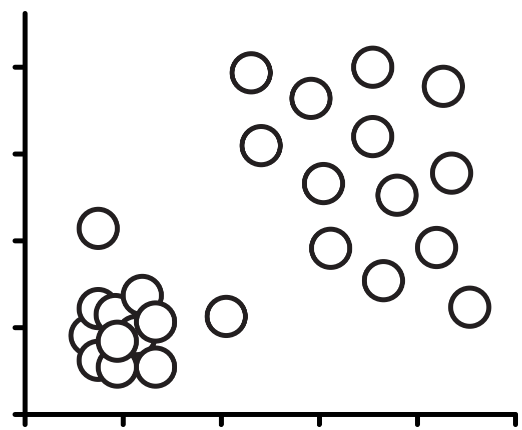
    the current value $q_{k\text{thNN}}(x)$ is monotonically decreasing
- In the `for` loop, if $q_{k\text{thNN}}(x)$ becomes smaller than the mth largest score so far, x never becomes an outlier
  - The `for` loop can be terminated earlier

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
  - This r is a 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

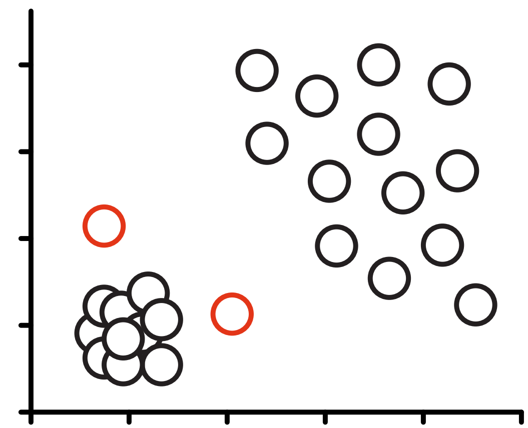
1. Scalability; $O(n^2)$
   - **Solution:** Partial computation of the pairwise distances to compute scores only for the top-*t* outliers
     - ORCA [Bay & Schwabacher, SIGKDD 2003]
     - iORCA [Bhaduri et al., SIGKDD 2011]
2. Detection ability
   - **Solution:** Introduce other definitions of the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



LOF (Local Outlier Factor)

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

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

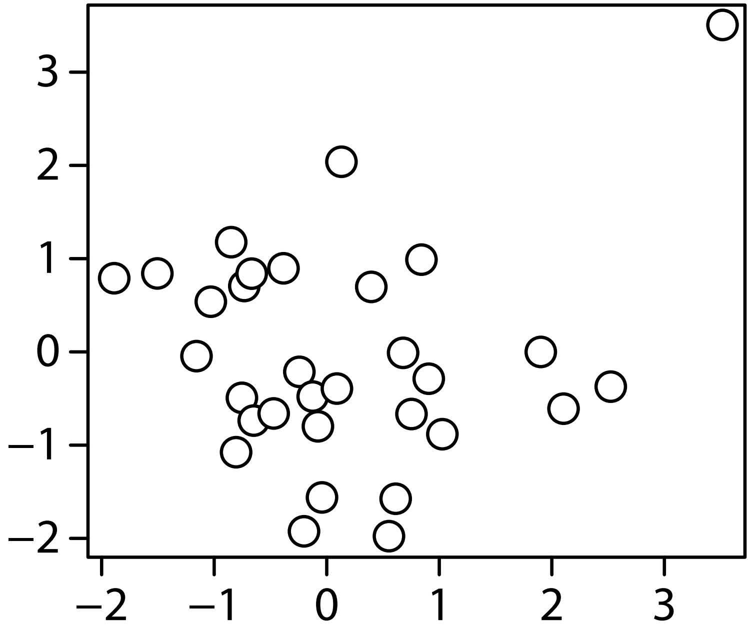
- The ratio of the local reachability density of x and the average of the local reachability densities of its kNNs

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
- For example, a ML library Jubatus (`http://jubat.us/en/`) supports LOF as an outlier detection technique
- The main drawback: scalability
  - $O(n^2)$ is needed for neighbor search
  - Same as kth-NN

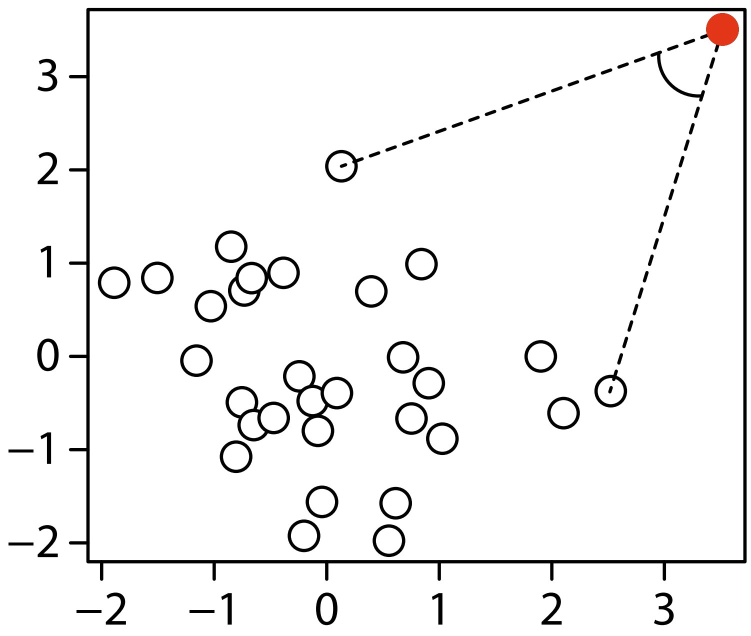
ABOD (Angle-Based Outlier Detection)

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



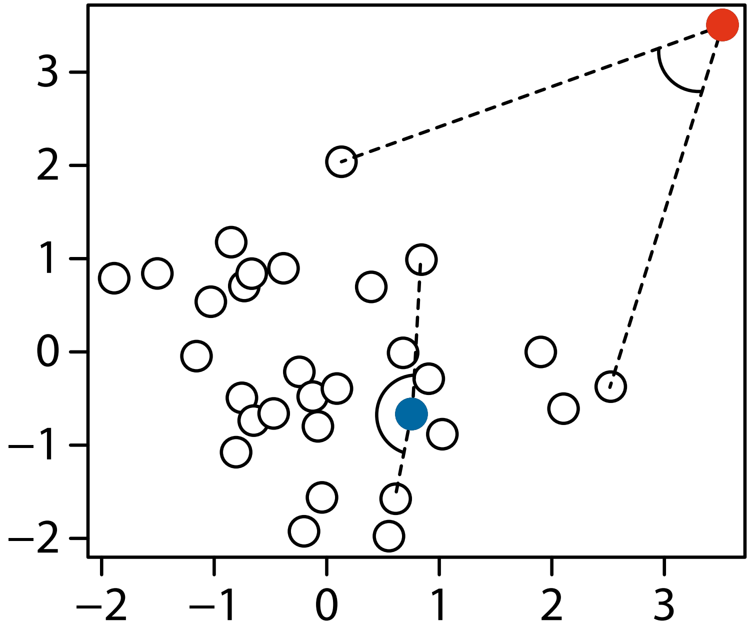
ABOD (Angle-Based Outlier Detection)

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



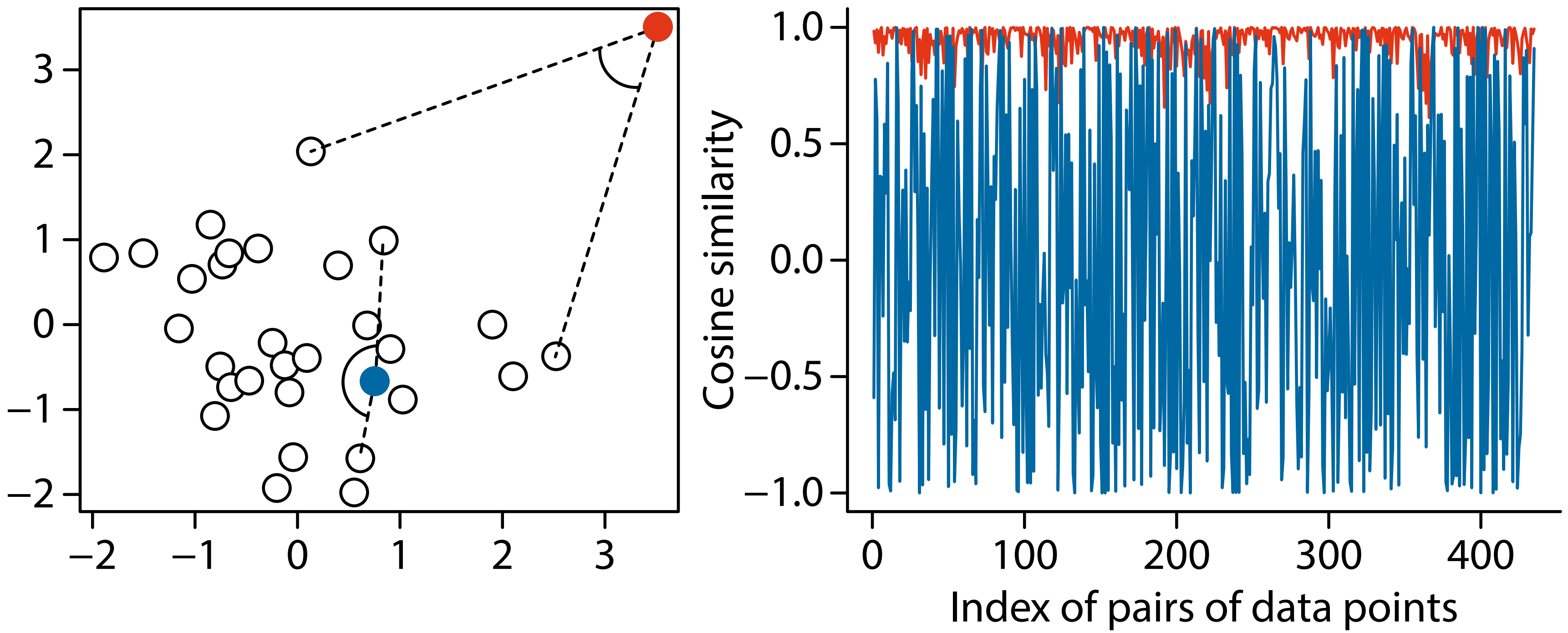
ABOD (Angle-Based Outlier Detection)

- 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 $\mathrm{ABOF}(x) := \mathrm{Var}_{y,z \in X}\, s(y - x, z - x)$
  - $s(x, y)$ is the similarity between vectors x and y, for example, 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 speeded-up 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 (Best Paper Runner-Up)
  - 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:
  1. 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
  2. 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

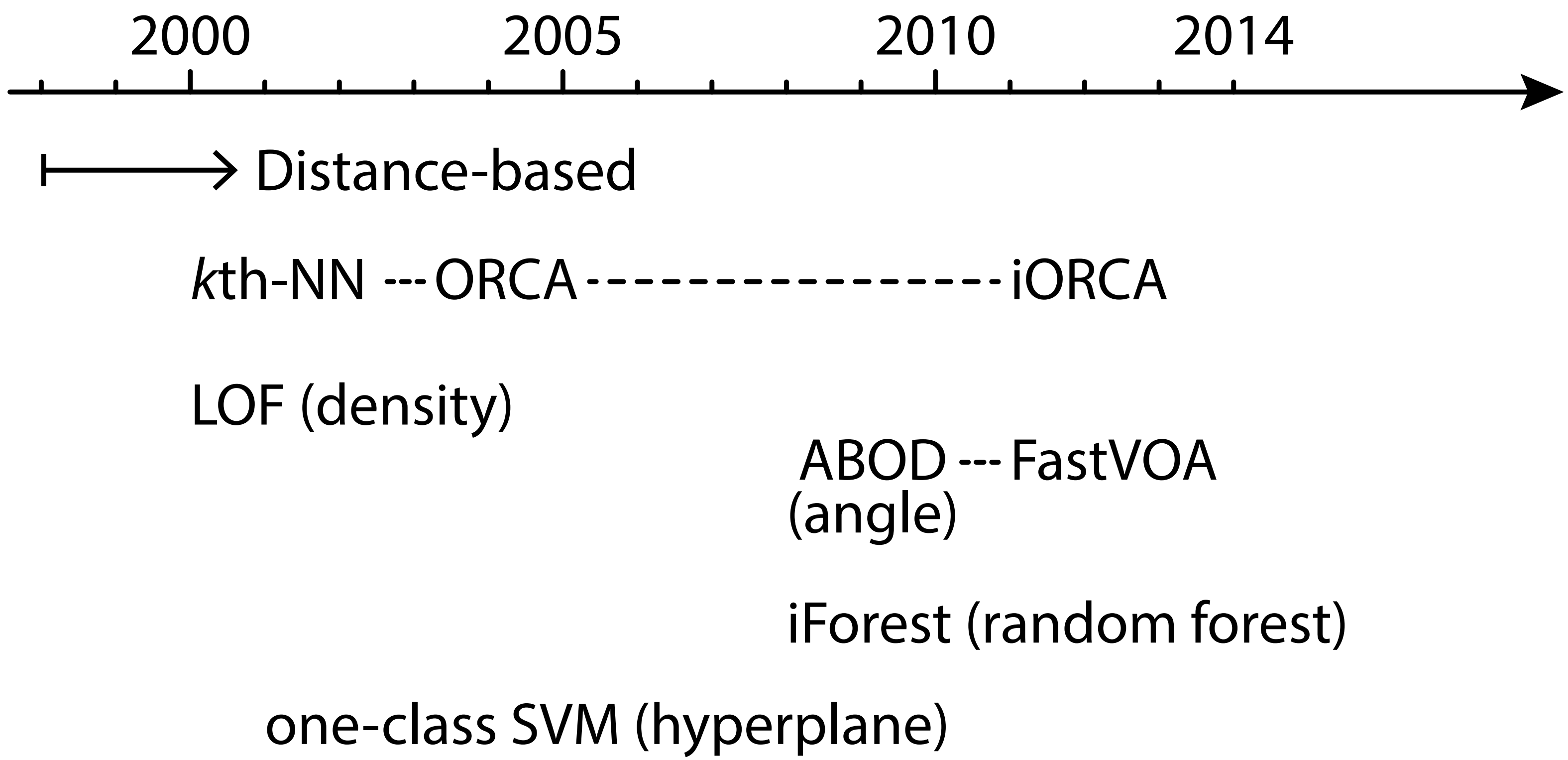
- A technique via hyperplanes by Schölkopf *et al.*
- The score of a vector **x** is $\rho - (w \cdot \Phi(\mathbf{x}))$
  - Φ: a feature map
  - w and ρ are the solution of the following quadratic program:

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

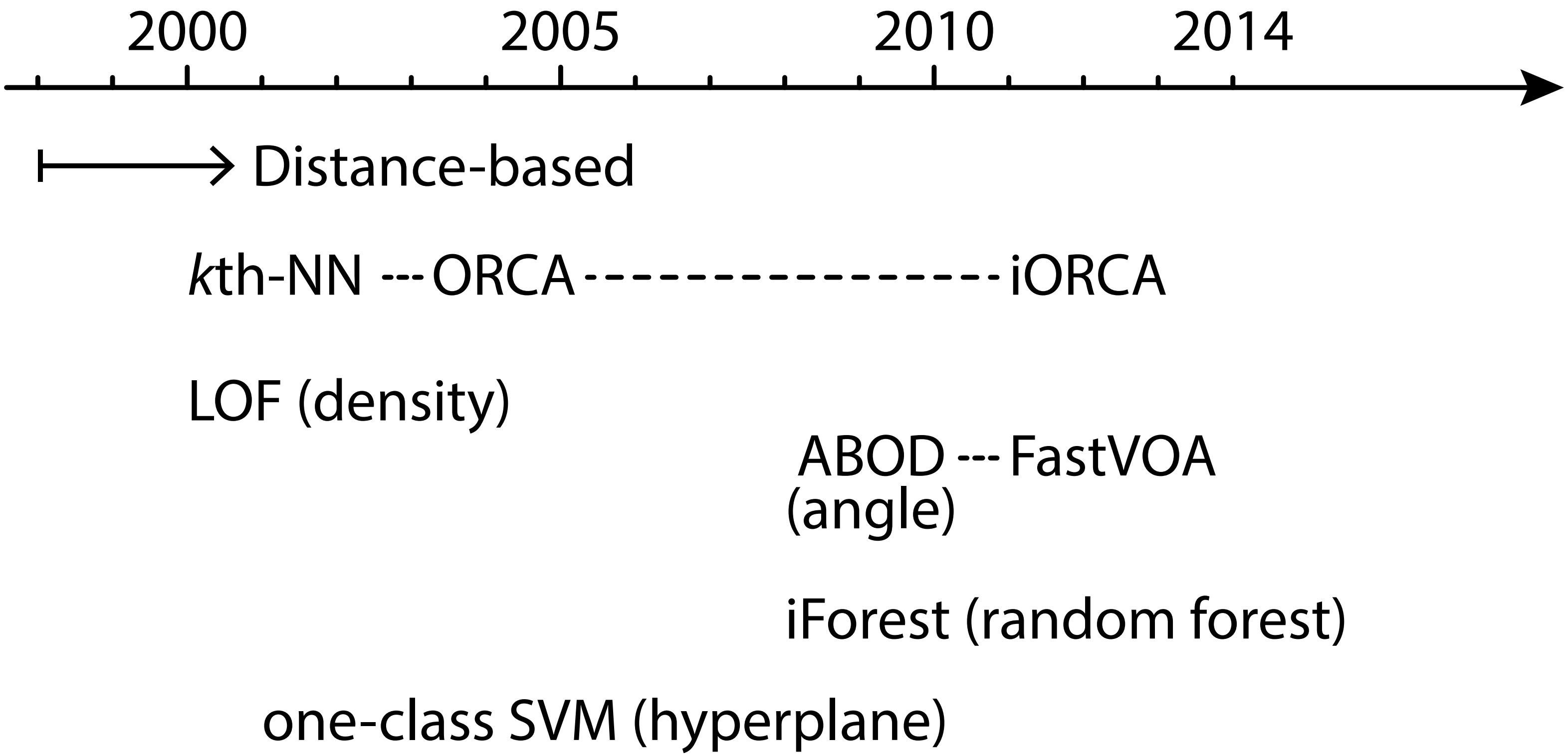
    subject to $(w \cdot \Phi(x_i)) \geq \rho - \xi_i,\ \ \xi_i \geq 0$

  - 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)

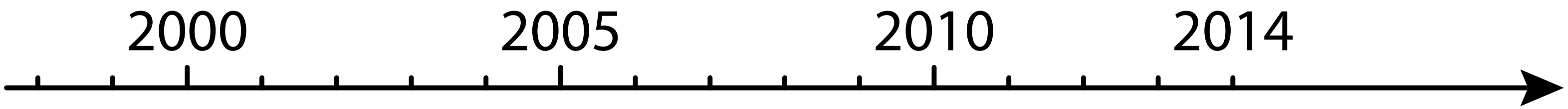
Timeline



The common drawback: Scalability

(except for iForest)

Timeline



Distance-based

*k*th-NN --- ORCA ------------------ iORCA

LOF (density)

ABOD --- FastVOA
(angle)

iForest (random forest)

one-class SVM (hyperplane)

Sampling
(Sugiyama-Borgwardt)

Outlier Detection via Sampling

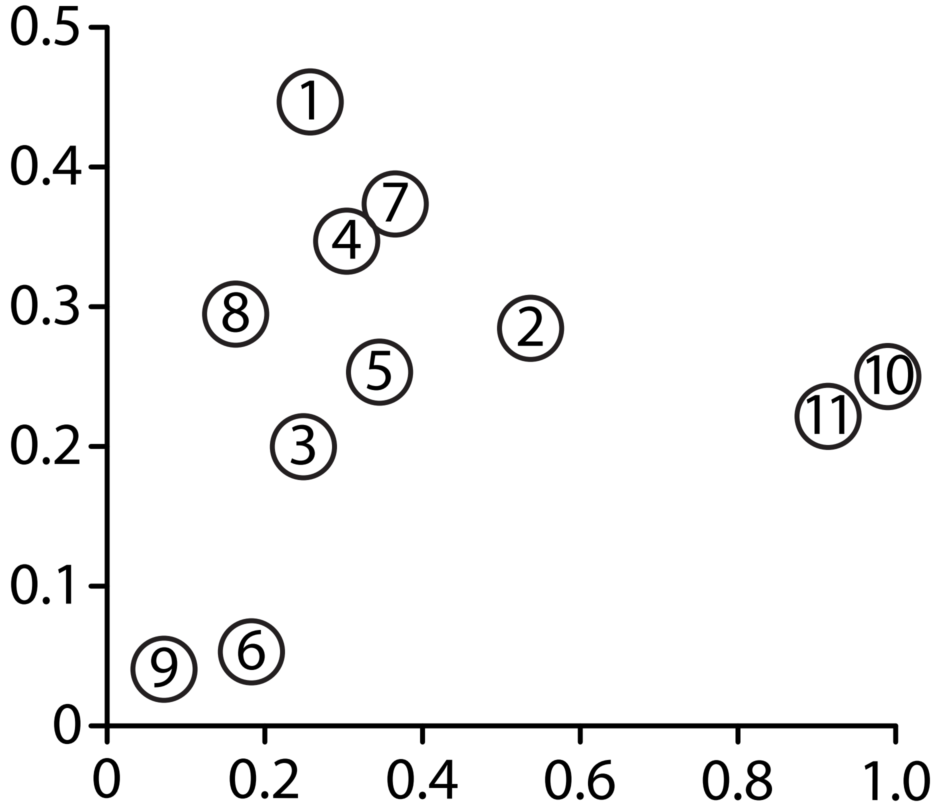
- (Sub-)Sampling was largely ignored in outlier detection
  - Find outliers from samples seems hopeless
- We proposed to use samples as a reference set
  - Sugiyama, M., Borgwardt, K.M., "Rapid Distance-Based Outlier Detection via Sampling", NIPS 2013
  - Sample size is surprisingly small, which is sometimes 0.0001% of the total number of data points
  - Accuracy is competitive with state-of-the-art methods

Outlier Detection via Sampling

- (Sub-)Sampling was largely ignored in outlier detection
  - Find outliers from samples seems hopeless
- We proposed to 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

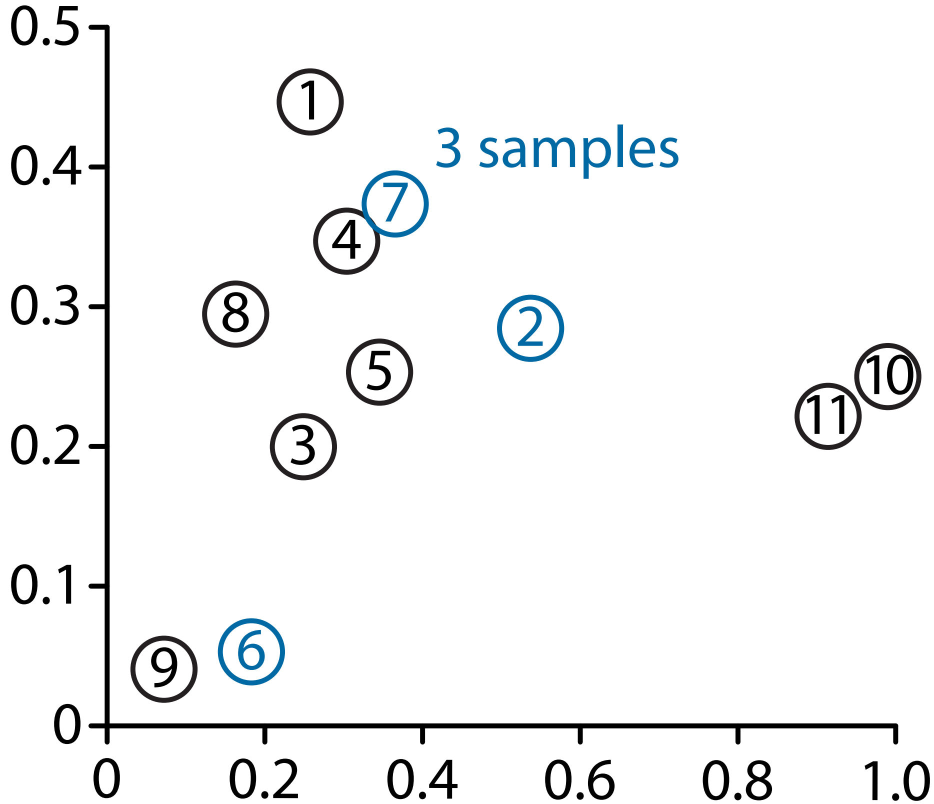
- Ensemble method with subsampling was also proposed:
  - Zimek, A. et al., “Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles”, SIGKDD 2013

Sugiyama-Borgwardt Method



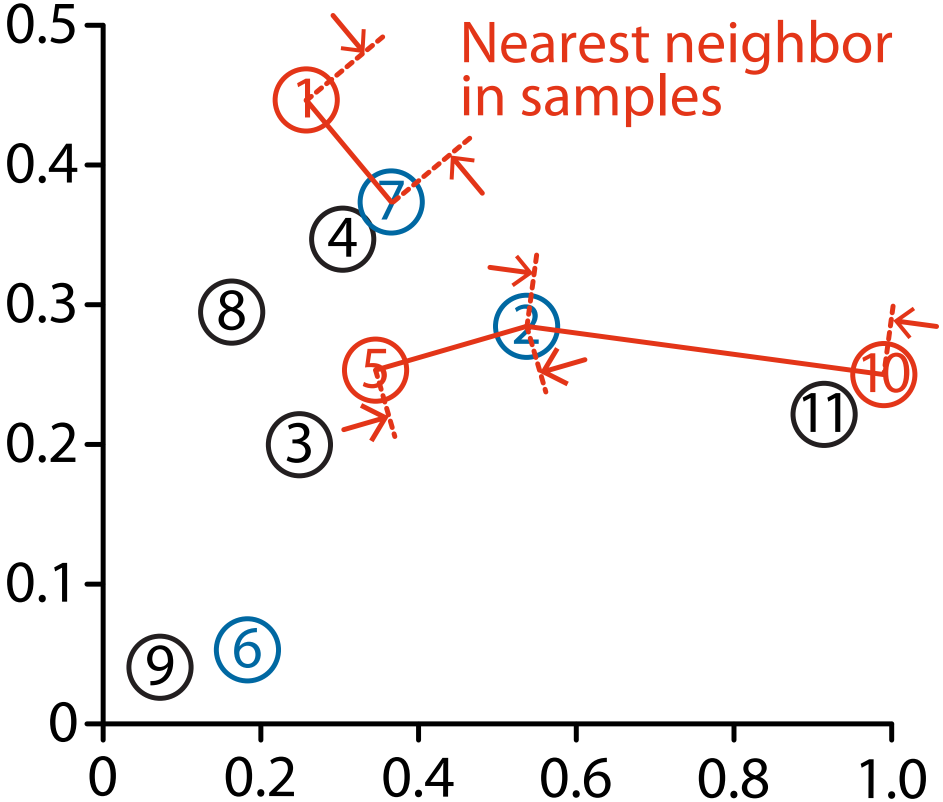
id	score

Sugiyama-Borgwardt Method



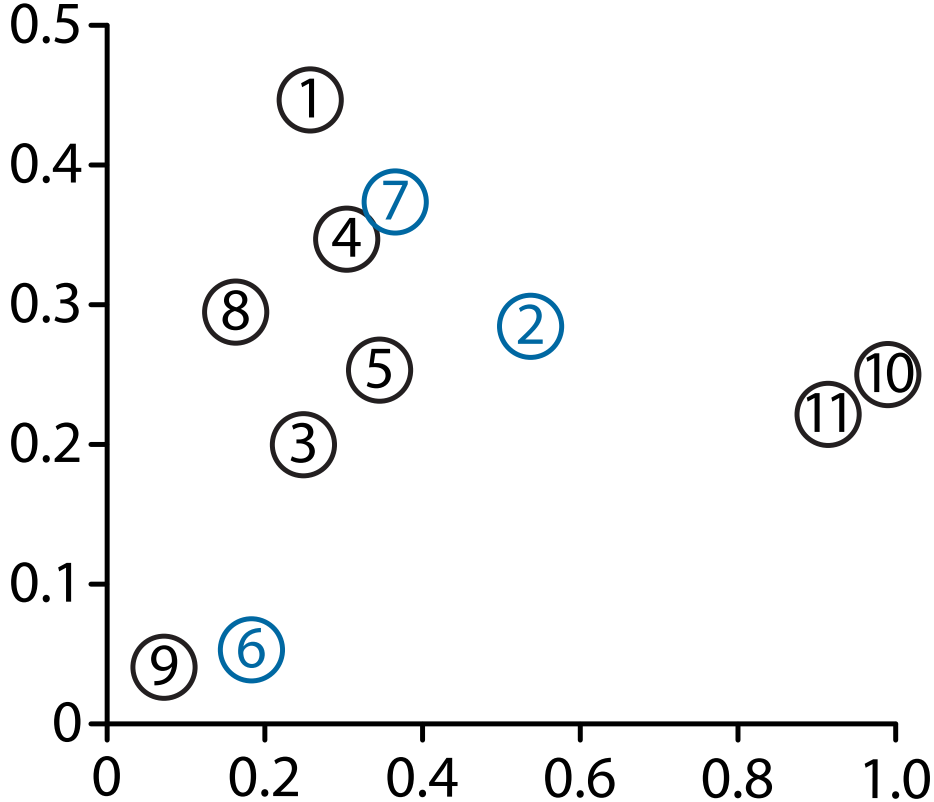
id	score

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

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

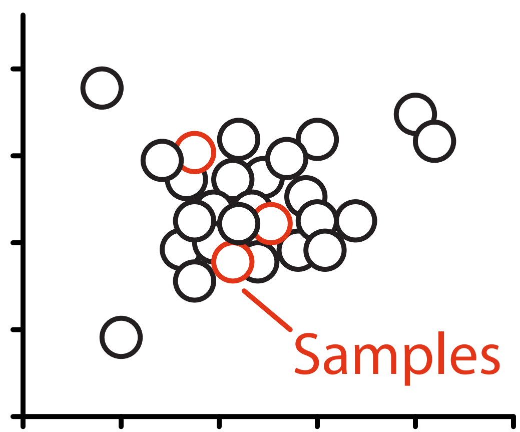
  - Input parameter: the number of samples $s = |S(X)|$
  - The time complexity is $\Theta(nms)$ and
    the space complexity is $\Theta(ms)$

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

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, which is similar to an outlier, slip into the sample set



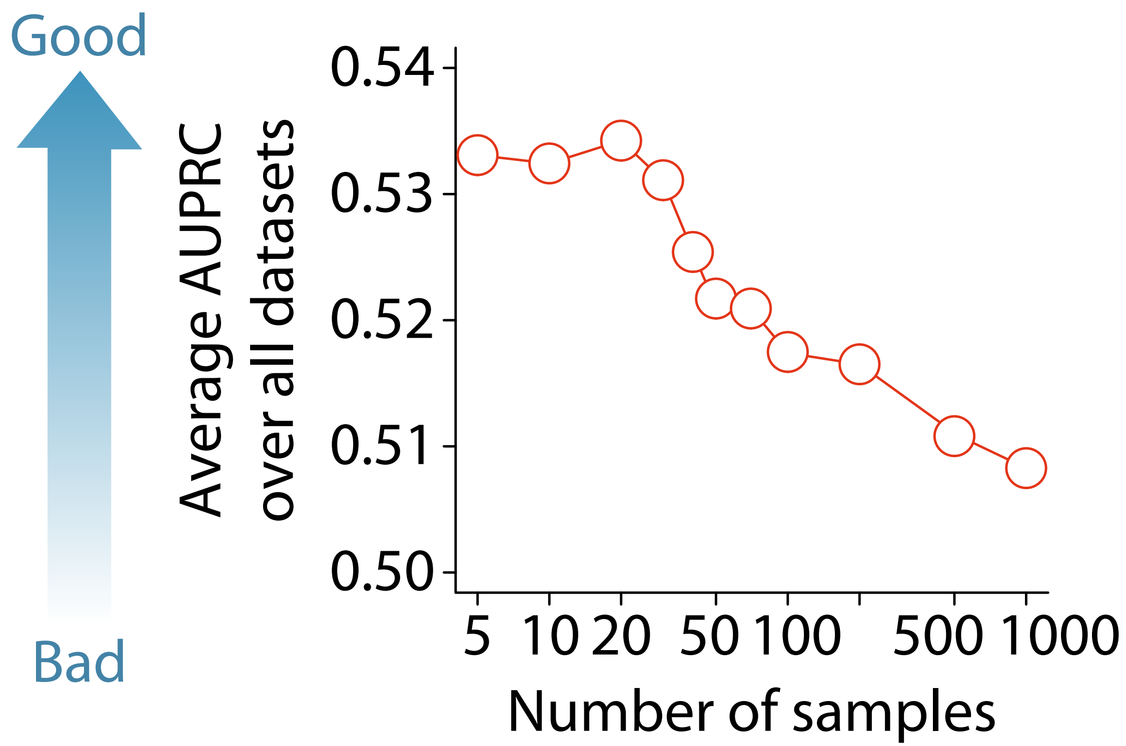
Experiments

- Examine state-of-the-art methods using synthetic and real-world datasets
  - Real data were collected from UCI repository
  - Points in the smallest class was assumed to be outliers
- Comparison partners:
  - *k*th-NN (iORCA), LOF, ABOD (FastVOA), iForest, one-class SVM, Wu and Jermaine's method
- Effectiveness was measured by AUPRC (area under the precision-recall curve)
  - Equivalent to the average precision over all possible cut-offs on the ranking of outlierness

Datasets (* are synthetic)

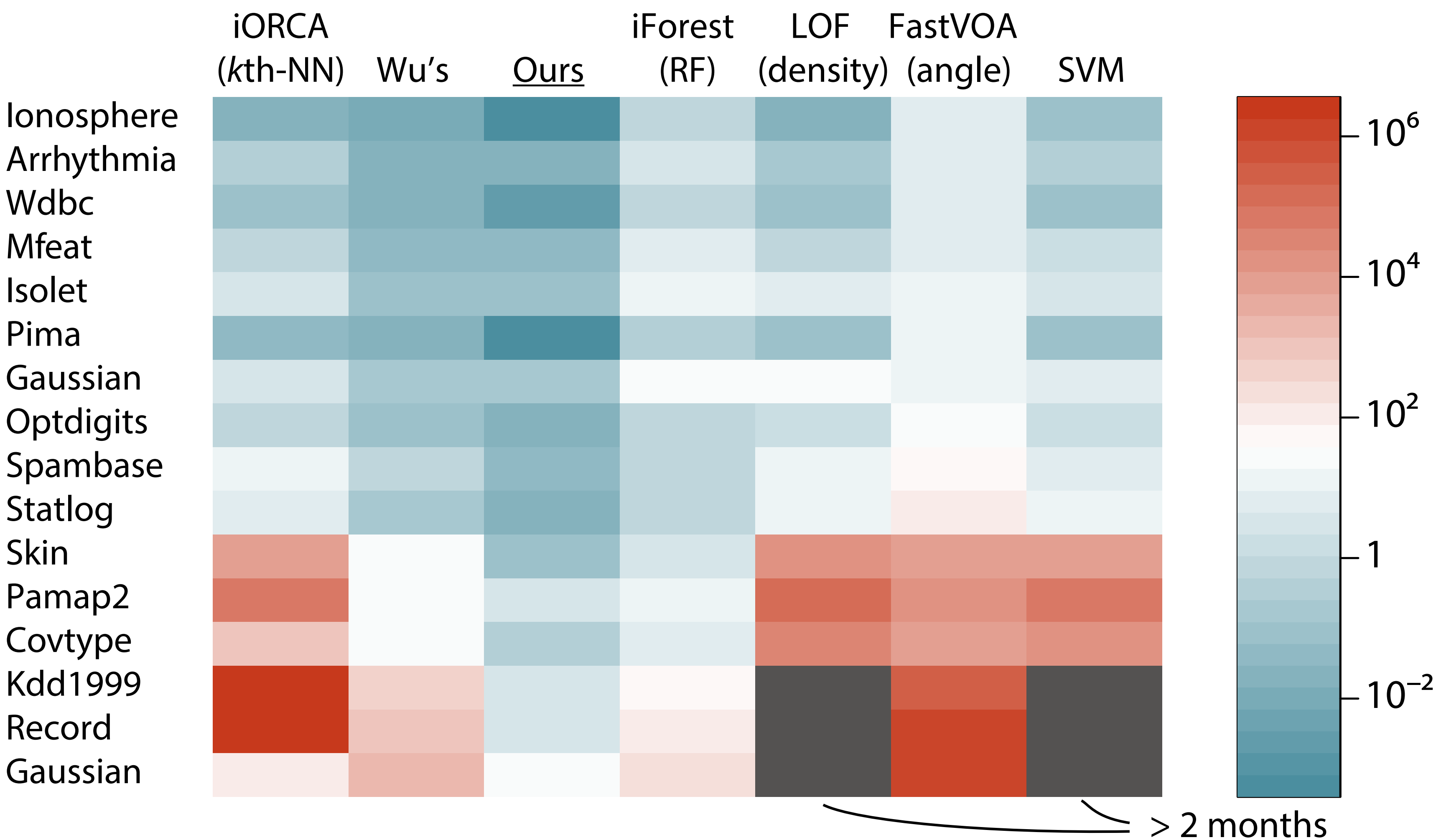
	# of objects	# of outliers	# of dims
Ionosphere	351	126	34
Arrhythmia	452	207	274
Wdbc	569	212	30
Mfeat	600	200	649
Isolet	960	240	617
Pima	768	268	8
Gaussian*	1000	30	1000
Optdigits	1688	554	64
Spambase	4601	1813	57
Statlog	6435	626	36
Skin	245057	50859	3
Pamap2	373161	125953	51
Covtype	286048	2747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

Sensitivity in sample sizes

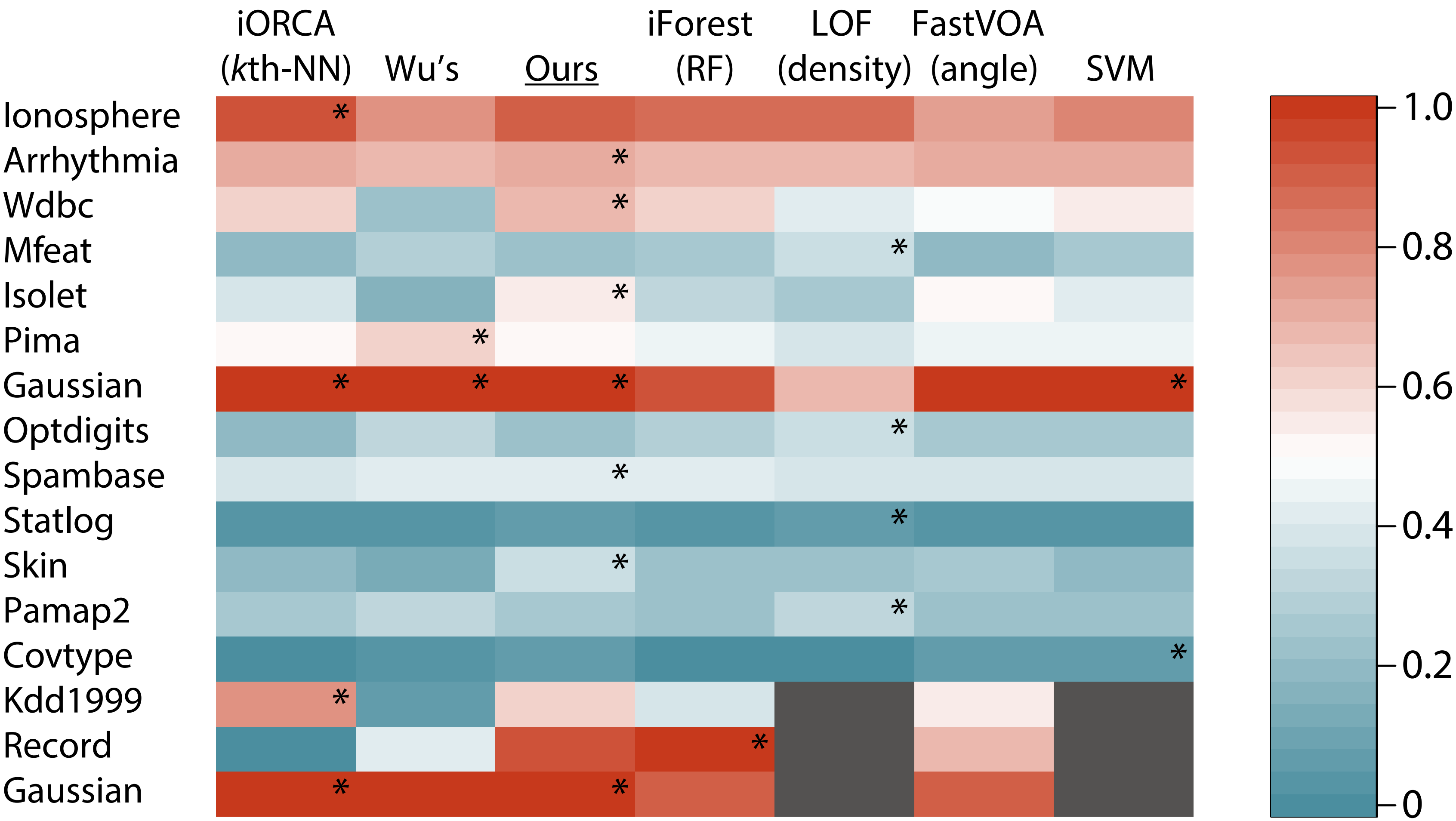


- Interestingly, the effectiveness was maximized at a rather small sample size, 20
  - Monotonically decreased as the sample size increased further

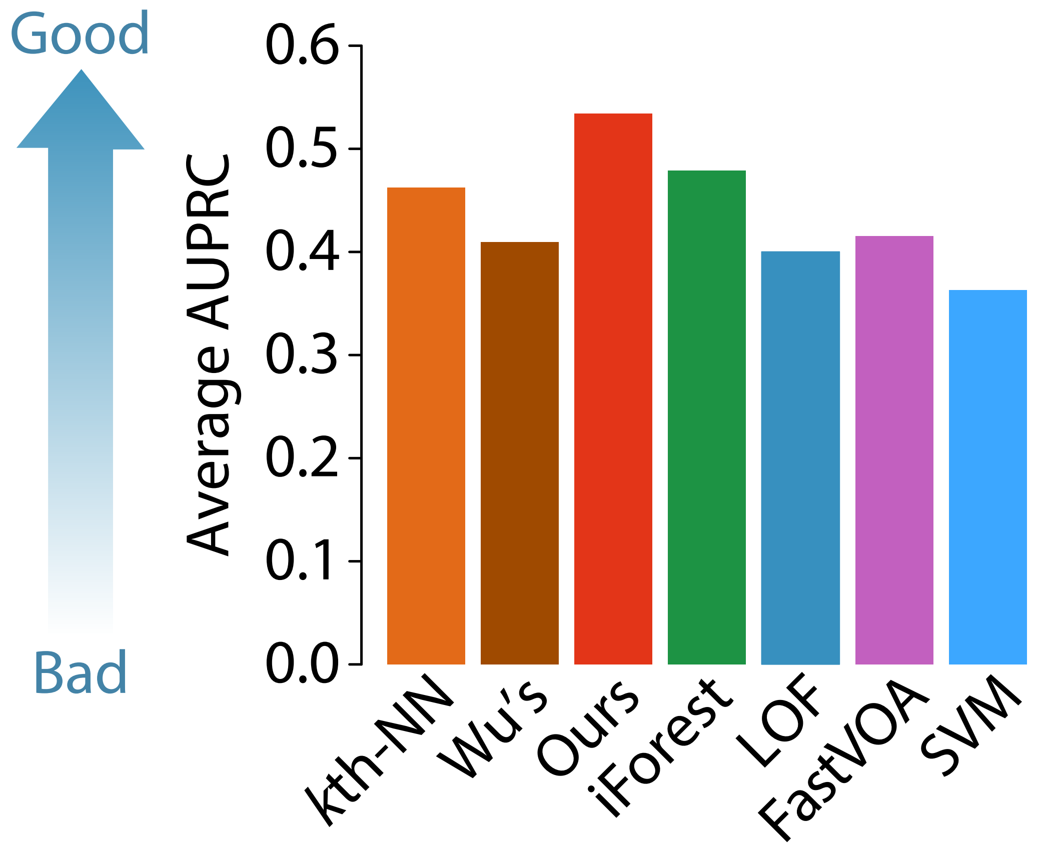
Running Time (seconds)



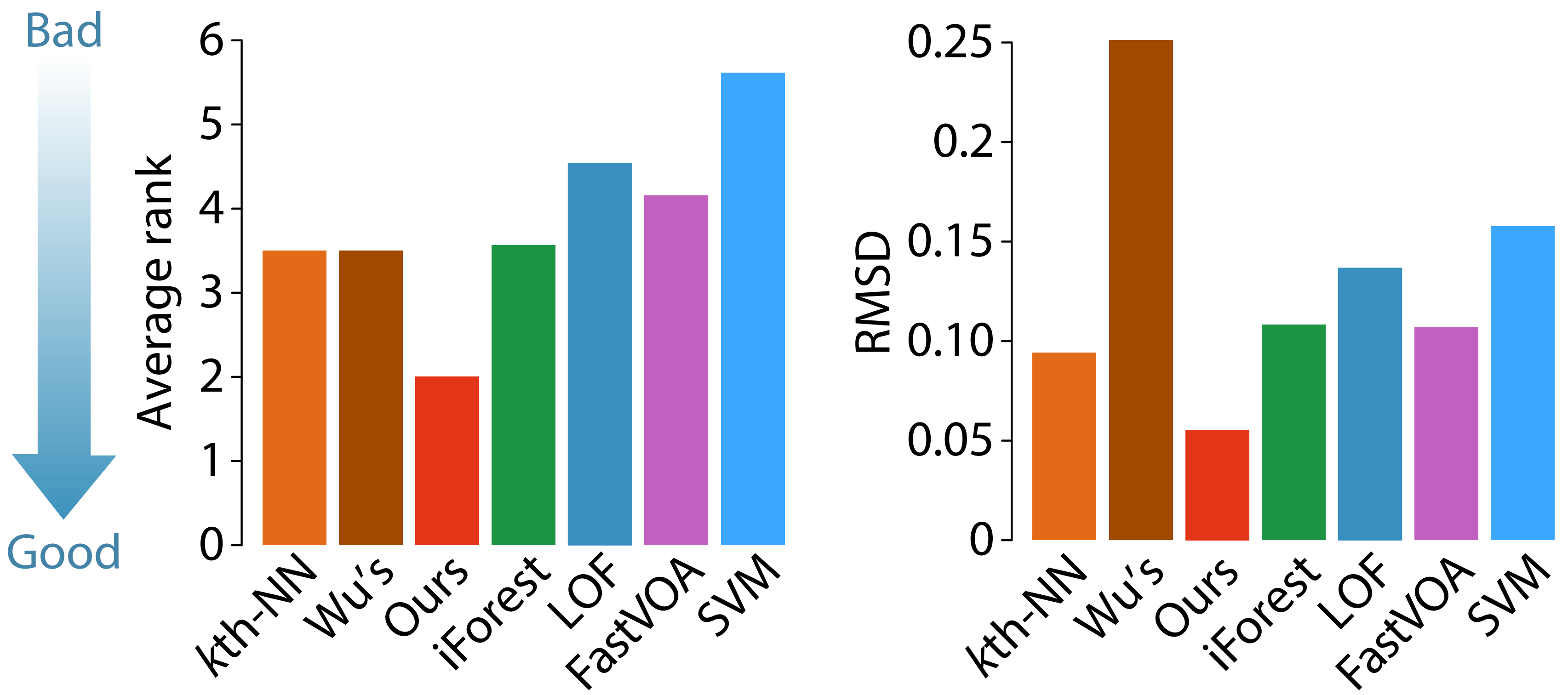
AUPRC (* are best scores)



Average of AUPRC over all datasets



Other statistics



- RMSD: the root-mean-square deviation to the best scores, rewarding methods that are always close to the best result

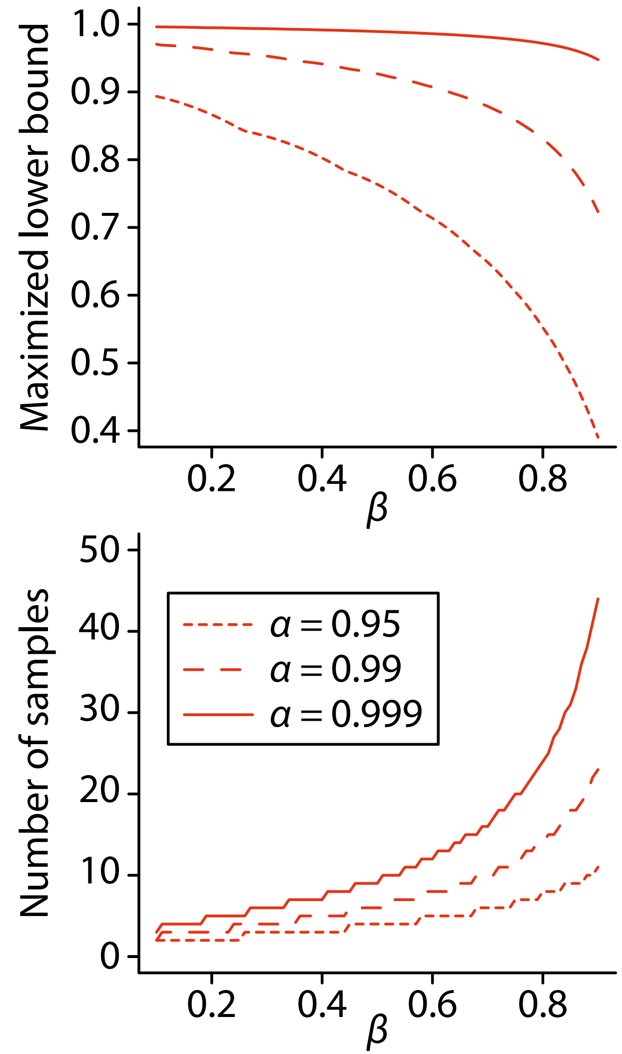
Notations

- $X(\alpha;\delta)$: 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 $x' \in \overline{X}(\alpha;\delta)$,
   $\Pr\big(q_{\mathrm{Sp}}(x) > q_{\mathrm{Sp}}(x')\big) \geq \alpha^s(1-\beta^s)$
   (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

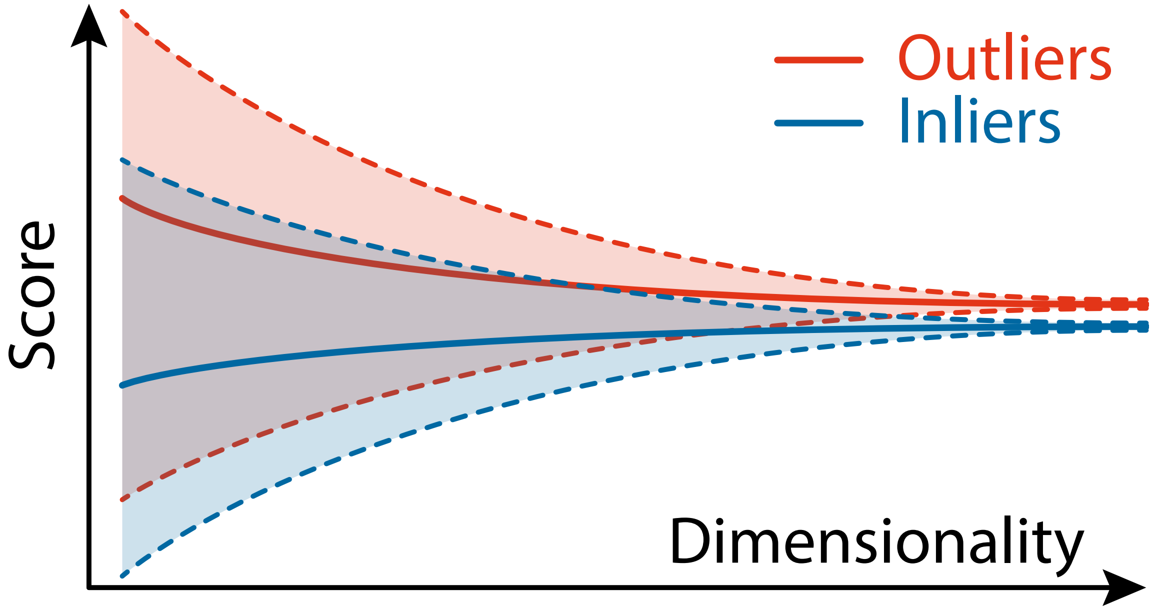


How about High-dimensional Data ?

- 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



Conclusion

- Sampling is a powerful tool in outlier detection
- Sugiyama-Borgwardt method is
  - much (2 to 6 orders of magnitude) faster than exhaustive methods
  - the most effective on average
- Future work:
  - On-line outlier detection with updating samples
  - Apply to other data types
- Thanks to:

MAX-PLANCK-GESELLSCHAFT

MLCB

Unterstützt von / Supported by

Alexander von Humboldt Stiftung/Foundation

Appendix

Evaluation criteria

- Precision v.s. Recall (Sensitivity)
  - Recall = TP / (TP + FN)
  - Precision = TP / (TP + FP)
- 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	Ground truth: 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)	

Wu and Jermaine's method

- Define the score of x as $d^k(x; S_x(X))$
  - $d^k(x; X)$ is the distance between x and its kth-NN in X
  - $S_x(X)$ is a subset of X, which is randomly and iteratively sampled for each object x
- Closely related to our method when $k = 1$
  - our method performs sampling only once
  - Wu's method performs sampling per each object
- Wu, M., Jermaine, C., "Outlier detection by sampling with accuracy guarantees", SIGKDD 2006

More Detailed Analysis

- A δ-*partition* $\mathcal{P}_\delta$ of $\overline{X}(\alpha;\delta)$:
  $\forall C \in \mathcal{P}_\delta$, $\max_{x,y\in C} d(x,y) < \delta$ and $\bigcup_{C\in\mathcal{P}_\delta} C = \overline{X}(\alpha;\delta)$
- For an outlier $x \in X(\alpha;\delta)$ and a cluster $C \in \mathcal{P}_\delta$,
  $\Pr\big(\, \forall x' \in C,\ q_{\mathrm{Sp}}(x) > q_{\mathrm{Sp}}(x')\,\big) \geq \alpha^s(1-\beta^s)$ with $\beta = (n-|C|)/n$
- Let $I(\alpha;\delta) \subset \overline{X}(\alpha;\delta)$ s.t. $\forall x \in X(\alpha;\delta)$, $\min_{x'\in I(\alpha;\delta)} d(x,x') > \delta$, $\mathcal{P}_\delta = \{C_1,\dots,C_l\}$ be a δ-partition of $I(\alpha;\delta)$, and $p_i = |C_i|/|I(\alpha;\delta)|$ for each $i \in \{1,\dots,l\}$
- Let $\varphi(s) = \sum_{\forall i;\, s_i \gneq 0} f(s_1,\dots,s_l;\mu,p_1,\dots,p_l)$, where f is the probability mass function of the multinomial distribution, and $\gamma = |I(\alpha;\delta)|/n$. Then

$$\Pr\big(\, \forall x \in X(\alpha;\delta), \forall x' \in \overline{X}(\alpha;\delta),\ q_{\mathrm{Sp}}(x) > q_{\mathrm{Sp}}(x')\,\big) \geq \gamma^s \max_{\mathcal{P}_\delta} \varphi(s)$$