Rapid Distance-Based Outlier Detection via Sampling

Mahito Sugiyama[1,2] Karsten Borgwardt[1,2,3]

Machine Learning and Computational Biology Research Group

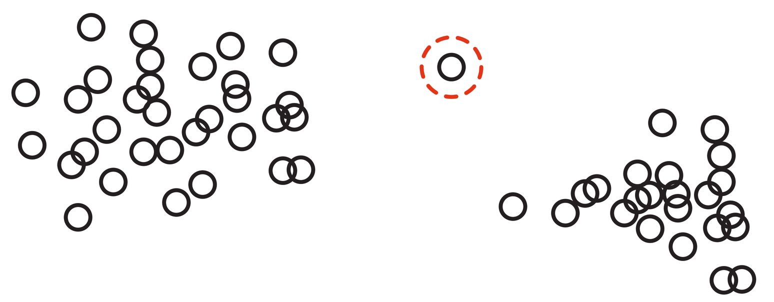
[1]Max Planck Institute for Intelligent Systems, Tübingen, [2]Max Planck Institute for Developmental Biology, Tübingen, [3]ZBIT, Eberhard Karls Universität Tübingen

Unterstützt von / Supported by Alexander von Humboldt Stiftung / Foundation

Background

How can we find outliers efficiently in massive datasets?

- Outliers are objects located far away from the remaining objects
- Outliers appear everywhere:
  - Intrusions in network traffic, credit card fraud, defective products in industry, medical diagnosis from X-ray images, ...
- **Specific task:** Assign an outlierness score to each point using pairwise distances
  - This distance-based approach has been successfully applied in various domains
    - Example: kth nearest neighbor [1,2], LOF [Breunig *et al.* SIGMOD 2000]
  - It does not require to model the underlying probability distribution, which is particularly challenging for high-dimensional data



Problem: Scalability

- Computation of all pairwise distances is needed: $O(n^2)$
- Two state-of-the-art solutions
  1. Partial computation to obtain only top-κ outliers
  2. Indexing of objects
- Unfortunately, both strategies are not sufficient
  1. The number of outliers often increases in direct proportion to the size of the dataset, which deteriorates the efficiency of partial computation
  2. Index structures are often not efficient enough for high-dimensional data

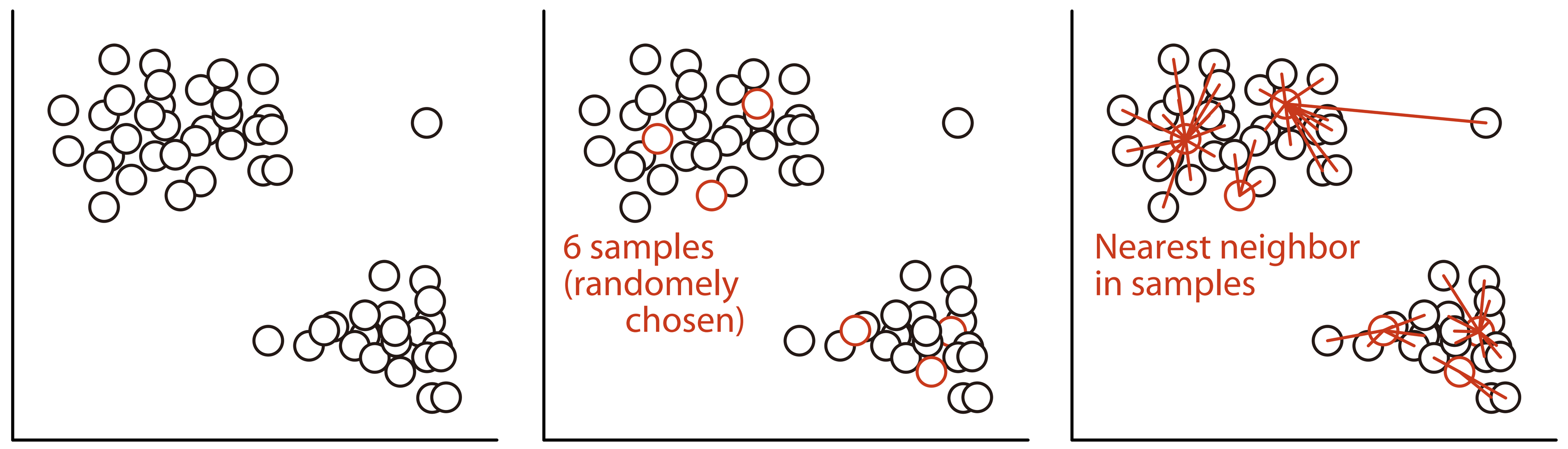
Proposal: Sampling-Based Outlier Detection

Solution: Sampling

- 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 $\Theta(ms)$



- **Related work:** Wu and Jermaine [3] proposed a sampling-based method:
  - Our method: one-time sampling, their method: iterative sampling for each point

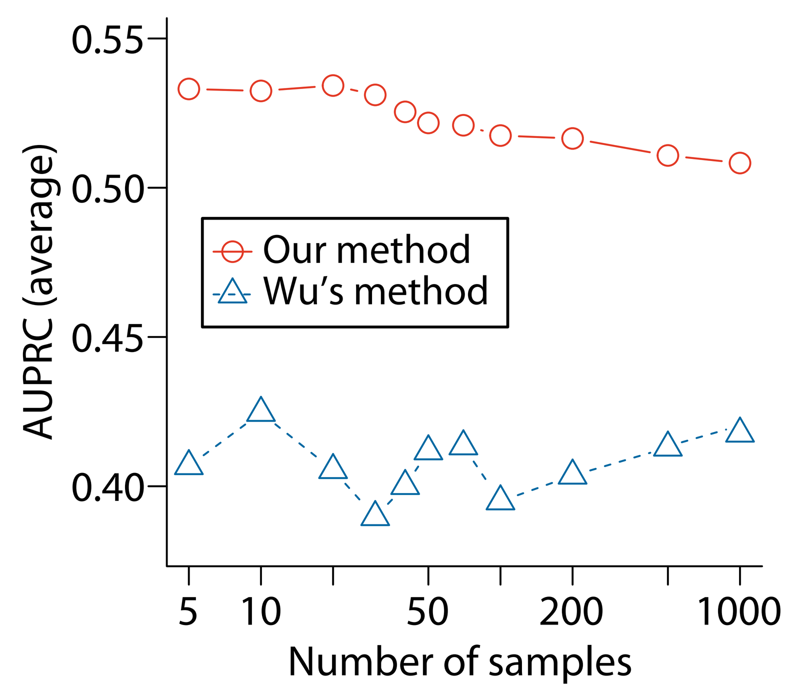
Experimental Results

- Comparison partners:
  $k$thNN (the latest technique iORCA [2] is used), Wu and Jermaine's method [3] (iterative sampling), iForest (random forest-like method) [Liu *et al.* 2012], LOF, FastVOA (angle-based method) [Pham and Pagh, 2012], One-class SVM [Schölkopf *et al.* 2001]
  - Parameters were set to be the same in the original papers or popular values

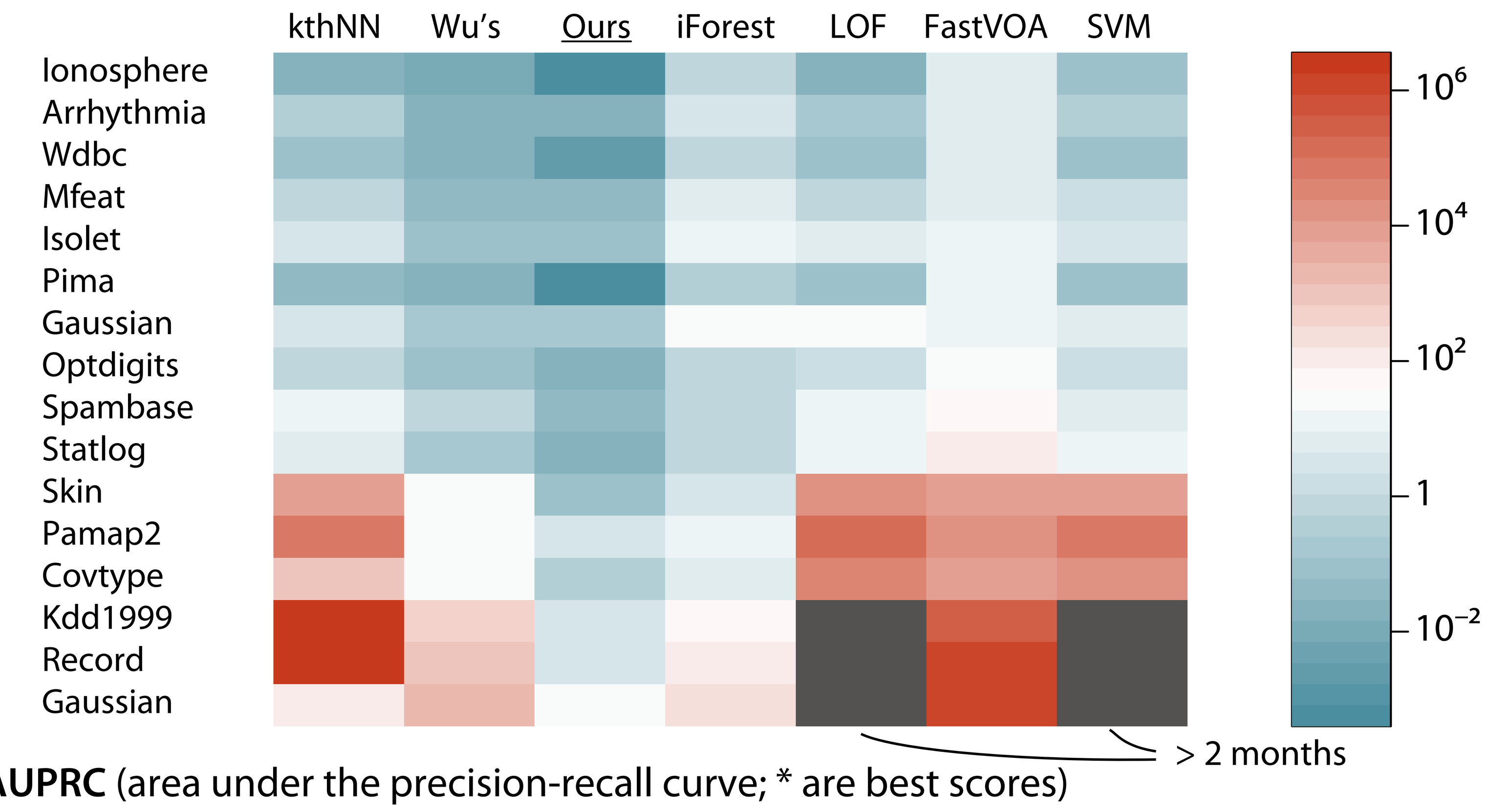
Datasets (* are synthetic)

	# 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

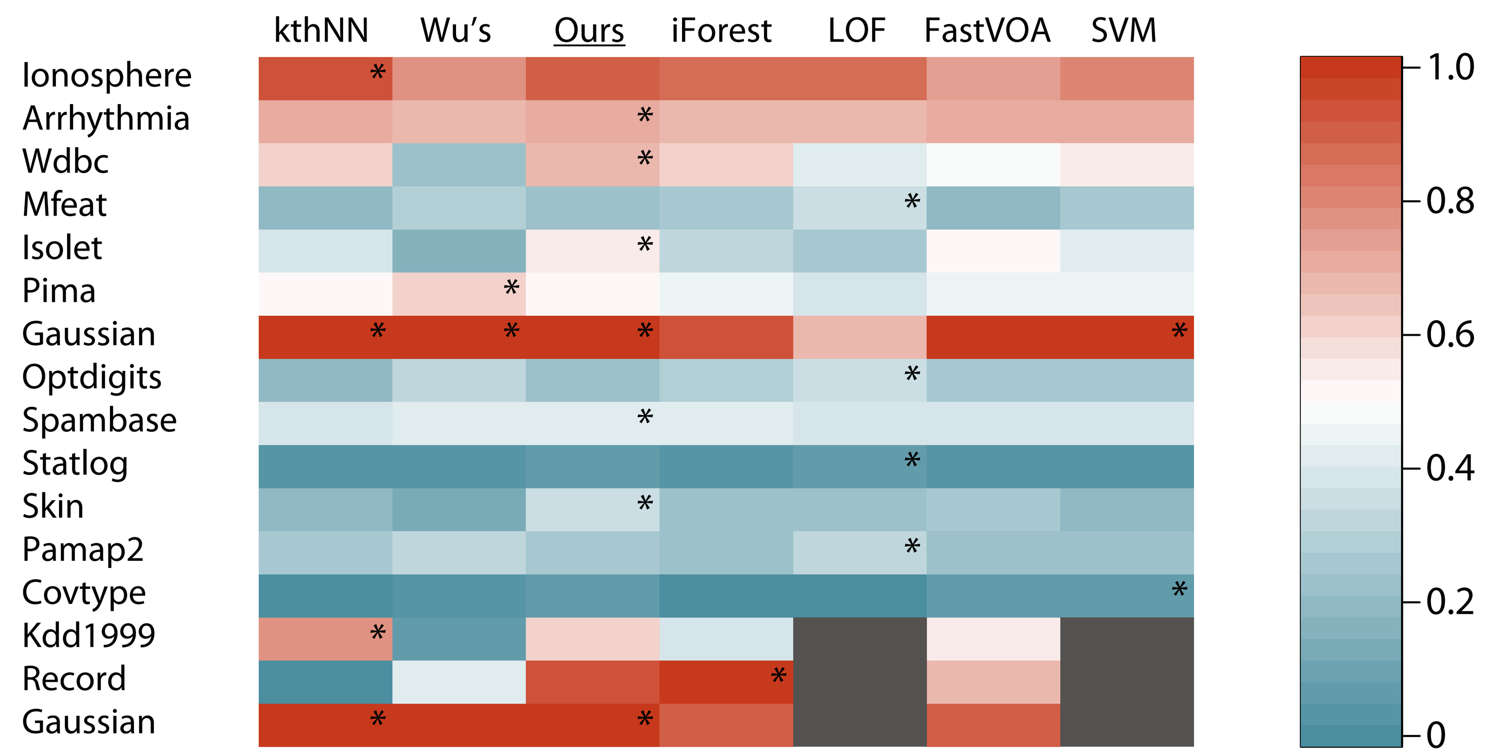
Sensitivity in sample sizes



Running time (in seconds)



AUPRC (area under the precision-recall curve; * are best scores)



	kthNN	Wu's	Ours	iForest	LOF	FastVOA	SVM
Average	0.453	0.410	**0.534**	0.479	0.400	0.468	0.421
Avg.rank	3.750	3.875	**2.188**	3.875	4.538	4.563	4.000
RMSD	0.259	0.274	**0.068**	0.133	0.152	0.140	0.094

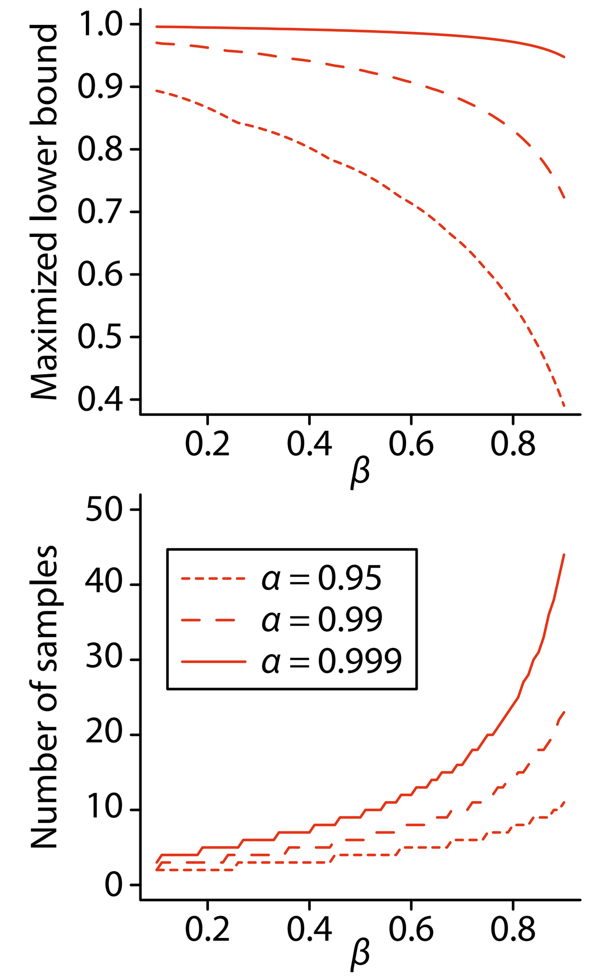
Theoretical Analysis

Main results

- $X(\alpha;\delta)$: the set of Knorr and Ng's DB(α,δ)-outliers:

$$x \in X(\alpha;\delta) \quad \text{if} \quad \left|\{x' \in X \mid d(x,x') > \delta\}\right| \geq \alpha n$$

  - $\delta \in \mathbb{R}$ is a distance threshold
  - $\alpha \in \mathbb{R}$ $(0 \leq \alpha \leq 1)$ is the fraction of objects which locate far away from x; this should be close to 1 by definition of outliers
  - NOTE: These parameters are not needed in practice



- $\overline{X}(\alpha;\delta) = X \setminus X(\alpha;\delta)$, the set of inliers
- Define β $(0 \leq \beta \leq \alpha)$ as the minimum value s.t.

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

- **Result 1:** For $x \in X(\alpha;\delta)$ and $x' \in \overline{X}(\alpha;\delta)$,

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

  (s is the number of samples)
  - This lower bound tends to be high in a typical setting (α is large, β is moderate)
- **Result 2:** This bound is maximized at

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

  - This value tends to be small

More detailed results

- 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 $C \in \mathcal{P}_\delta$,

$$\Pr\bigl(\forall x' \in C,\ q_{\mathrm{Sp}}(x) > q_{\mathrm{Sp}}(x')\bigr) \geq \alpha^s(1-\beta^s) \text{ with } \beta = (n-|C|)/n$$

- 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 $\varphi(s) = \sum_{\forall i;\, s_i \geq 0} f(s_1,\dots,s_l;\mu,p_1,\dots,p_l)$, where f is the probability mass function of the multinomial distribution, and $\gamma = |I(\alpha;\delta)|/n$. Then

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

Conclusion

- Our method is much (2 to 6 orders of magnitude) faster than exhaustive methods
- Our method is the most effective on average

References

1. Original paper on distance-based outliers: Korr, E. M., Ng, R. T., and Tucakov, V. **Distance-based outliers: algorithms and applications.** *The VLDB Journal*, 8(3):237–253, 2000.
2. State-of-the-art of $k$thNN detection: Bhaduri, K., Matthews, B. L., and Giannella, C. R. **Algorithms for speeding up distance-based outlier detection.** KDD 2011.
3. Related sampling-based method: Wu, M. and Jermaine, C. **Outlier detection by sampling with accuracy guarantees.** KDD 2006.