The 20th International Conference Advanced Data Mining and Applications

Anomaly Detection Based on Isolation Mechanisms

Dr Ye Zhu
Senior Lecturer
Deakin University, Australia
ye.zhu@ieee.org

Dr Haolong Xiang
Lecturer
Nanjing University of Information Science and Technology, China
hlxiang@nuist.edu.cn

Mr Xin Han
PhD Student
Deakin University, Australia
xin.han@deakin.edu.au

Mr Yang Cao
PhD Student
Deakin University, Australia
charles.cao@deakin.edu.au

Sydney, Australia 3rd ~ 5th December, 2024

Outline

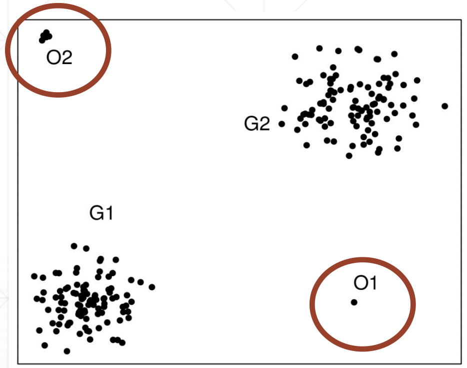
- Backgrounds and Definitions
- Isolation Mechanism
- Data-dependent kernels
    - Isolation kernel
    - Isolation distribution kernel
- Applications
- Parameter Settings and Model Optimisation
- Extensions

Reference

Cao, Yang, Haolong Xiang, Hang Zhang, Ye Zhu, and Kai Ming Ting. "Anomaly Detection Based on Isolation Mechanisms: A Survey." arXiv preprint arXiv:2403.10802 (2024).

What are Anomalies?

- **Anomalies (a.k.a., outliers, novelties): Points that are significantly different from most of the data**
  - ✓ **Rare**
  - ✓ **Irregular**

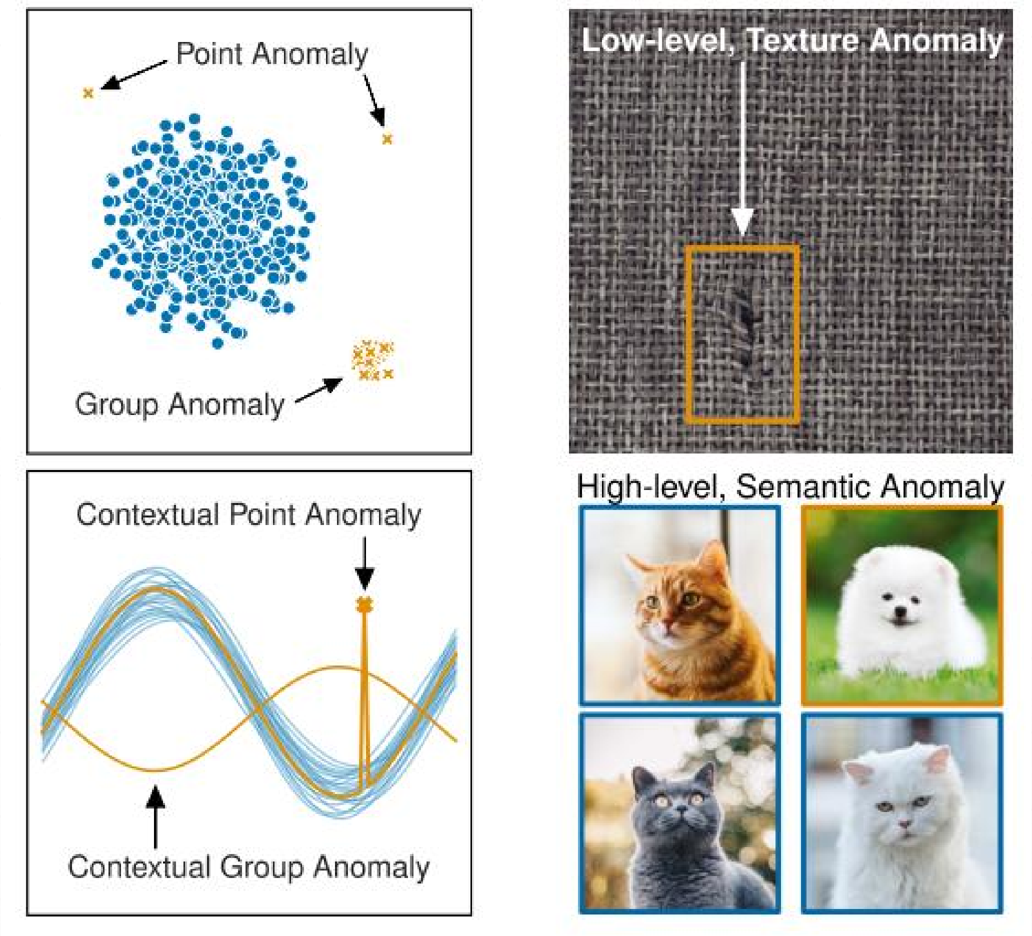


Source: Wikipedia

  - Binary Output versus scoring
    - Binary output generates a yes/no tag
    - **Preferable and more general:** Scoring output generates a real-valued score or rank

Multiple ways to define what makes an anomaly different

Types of Anomalies?



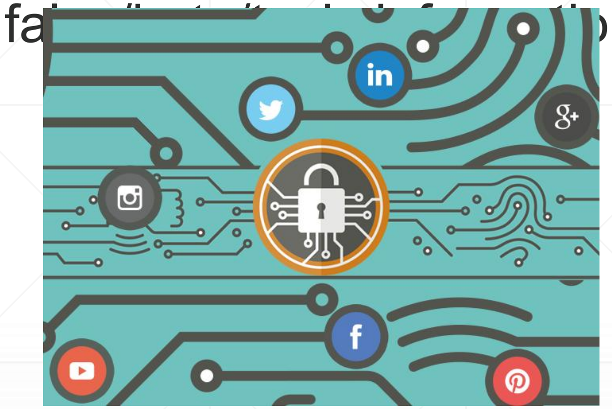
- A **point anomaly** is a single anomalous point.
- A **group anomaly** can be a cluster of anomalies or some series of related points that are anomalous under the joint series distribution.
- A **contextual point anomaly** occurs if a point deviates in its local context, here a spike in an otherwise normal time series.
- A **low-level sensory anomaly** deviates from the low-level features
- A **semantic anomaly** deviates in high-level factors of variation or semantic concepts

Real-World Application Domains

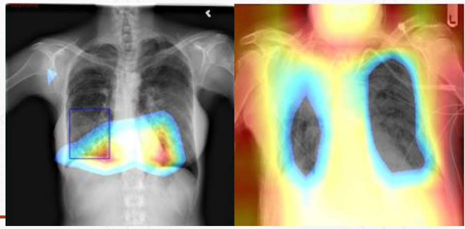
Cybersecurity:
attacks, malware, malicious apps/URLs, biometric spoofing

Finance:
credit card/insurance frauds, market manipulation, money laundering, etc.

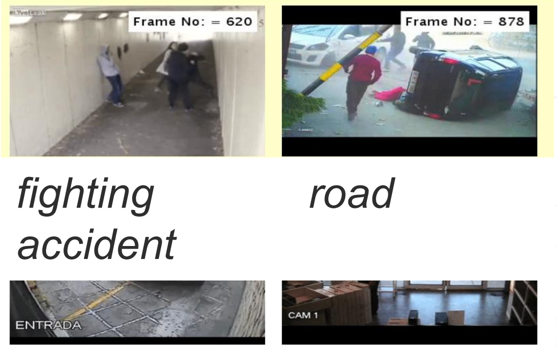
Social Network and Web Security:
false/malicious accounts, fake/hate/toxic information



Healthcare:
lesions, tumours, events in IoT/ICU monitoring, etc.



Video Surveillance:
criminal activities, road accidents, violence, etc.



fighting *road*
accident

shooting *shoplifting*

Industrial Inspection:
Defects, micro-cracks

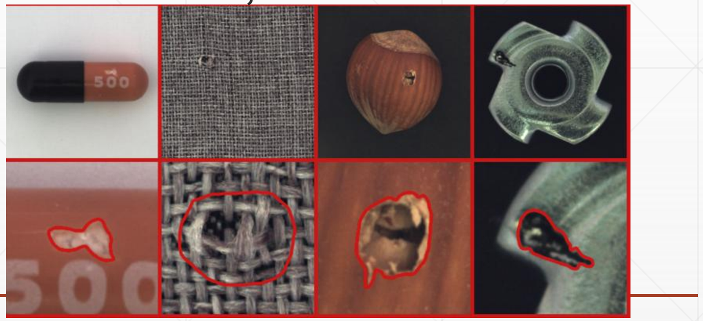


Image source: UCF-Crime data, MVTec AD data, etc.

Origin of Isolation Mechanism

Idea: Anomalies are data points that are few and different.

- As a result, anomalies are susceptible to a mechanism called ***isolation***.

Two key features:

- Data subsampling
- Isolation mechanism to partition the data space

Theoretical Analyses: NIPS2013 [*], MLJ2017 [#]

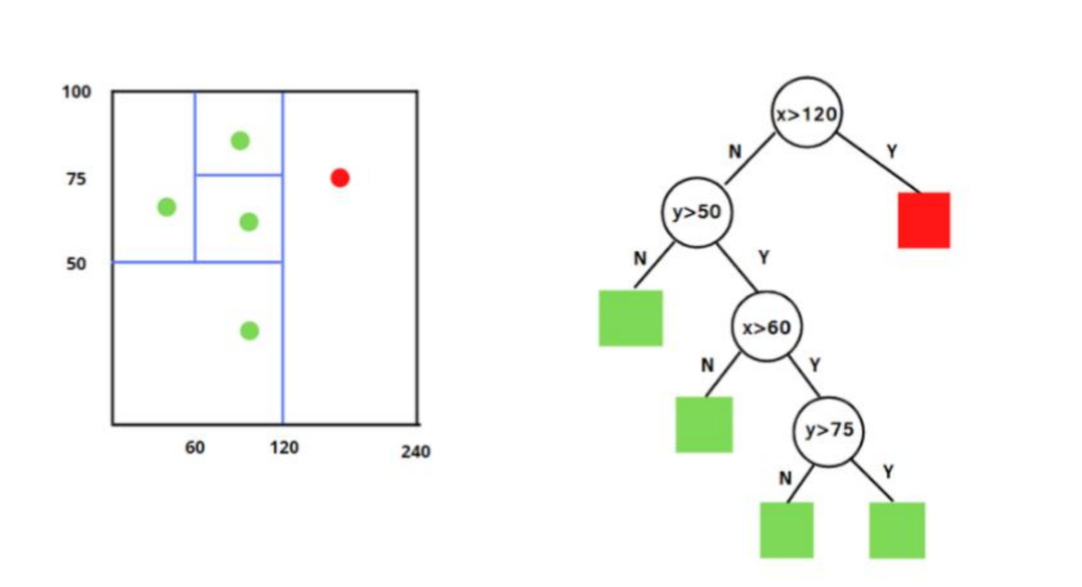
Not just for efficiency as claimed by Aggarwal & Sathe (2017)

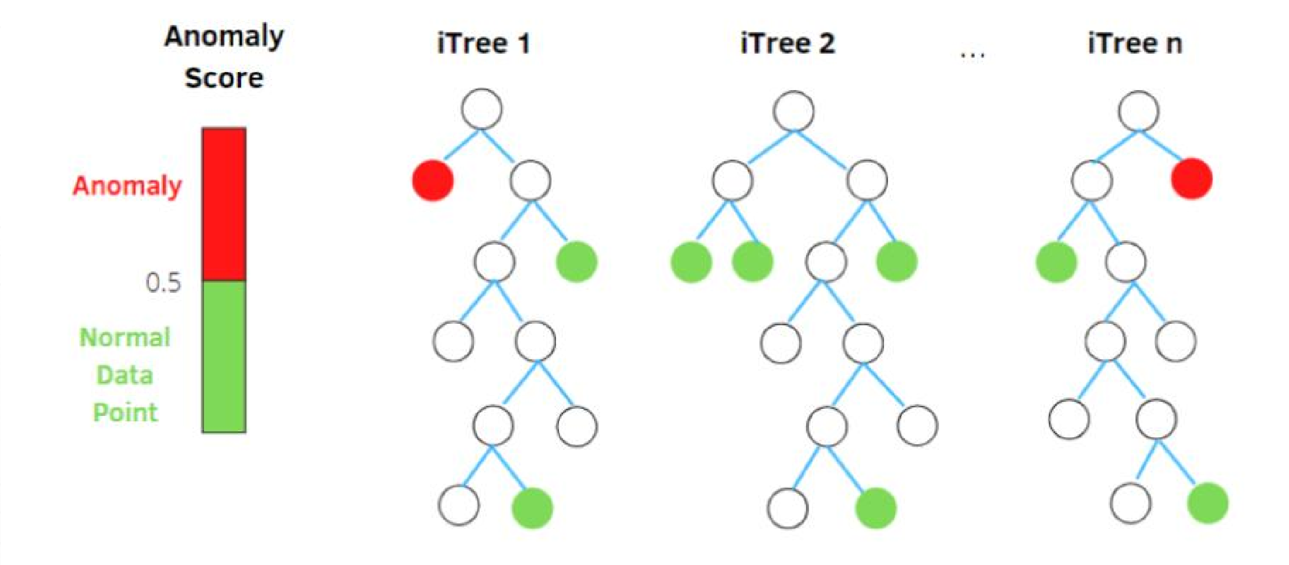
[*] Sugiyama M., Borgwardt K. (2013) Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems* 26, 467-475

[#] Ting, K. M., Washio, T., Wells, J. R., Aryal, S. (2017). Defying the Gravity of Learning Curve: A Characteristic of Nearest Neighbour Anomaly Detectors. *Machine Learning*. Vol 106, Issue 1, 55-91.

Isolation Forest

The Isolation Forest ‘isolates’ observations (subsample) by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

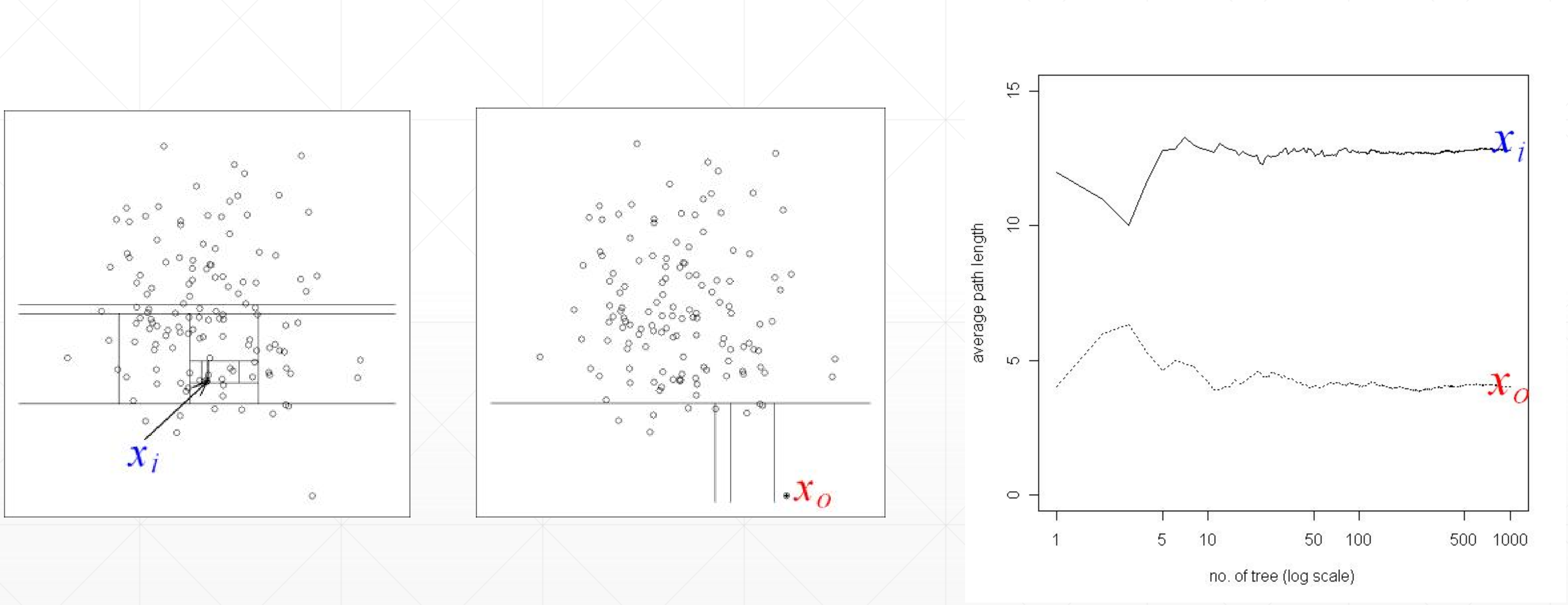




$Score(\boldsymbol{x}) = \frac{1}{t}\sum_{i=1}^{t} \ell_i(\boldsymbol{x})$where $\ell_i(\boldsymbol{x})$ is the path length of test point $\boldsymbol{x}$ traversed in tree i.

Liu F.T., Ting K.M. and Zhou Z.H. (2008) Isolation Forest, In *Proceedings of IEEE ICDM*, p:413–422.
Liu F.T., Ting K.M., and Zhou Z.H. (2012) Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery,* 39(3), p:1-39.

Isolation Forest (cont.)

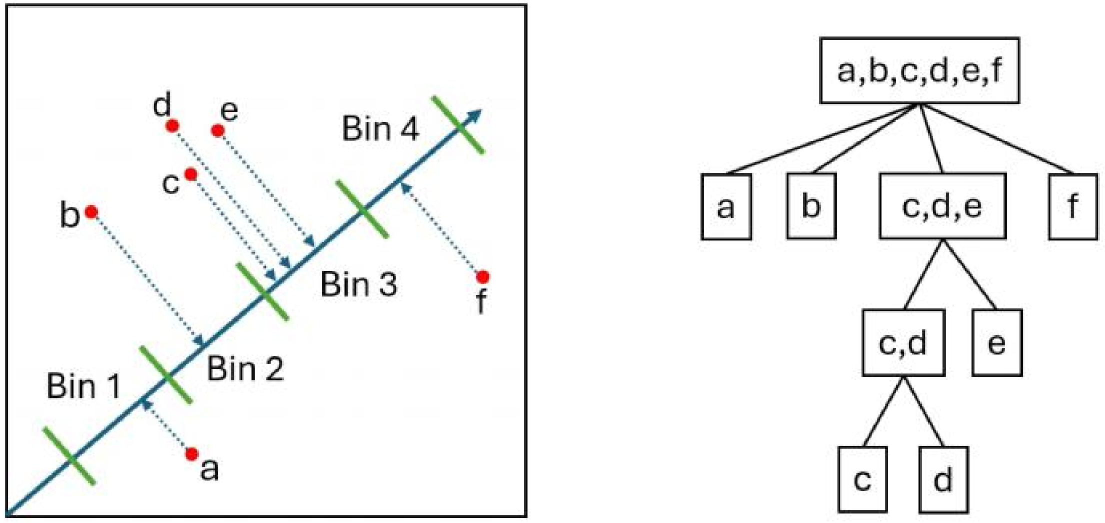


Source: Liu et al. 2008

Locality-sensitive hashing iForest

Projecting each point on a random vector and dividing the projected values into equal sized sequential intervals, then each bin becomes a branch of an LSHiTree.

Each level of the tree may have a different number of branches, thus, reducing the height of the tree for higher efficiency in anomaly score calculation.

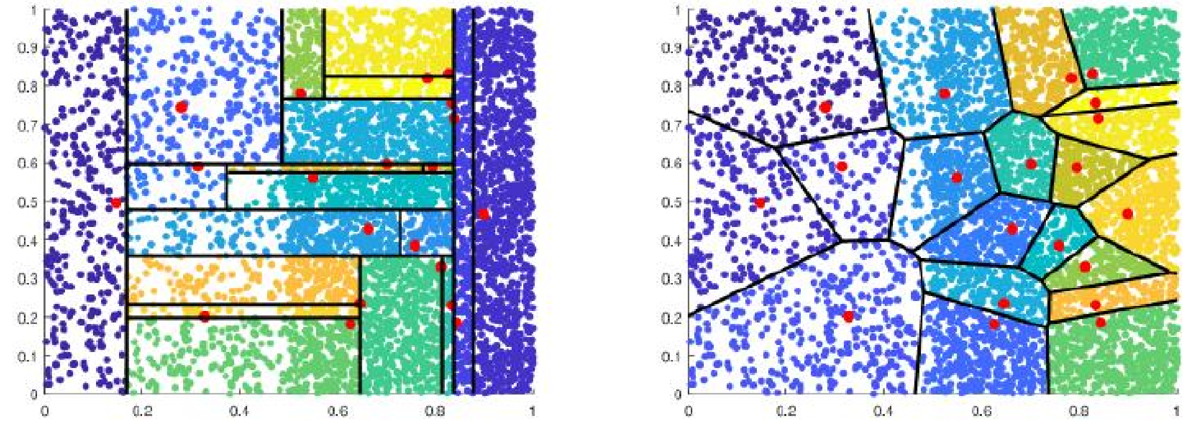


(a) Data projection and hash

(b) One LSHiTree

Isolating partitions

- Large in sparse regions and small in dense regions
- Adapt to local data distribution
- This characteristic is important not only for point anomaly detection, but also for deriving data dependent kernels (to be described later).



(a) Axis-parallel splitting

(b) NN partitioning

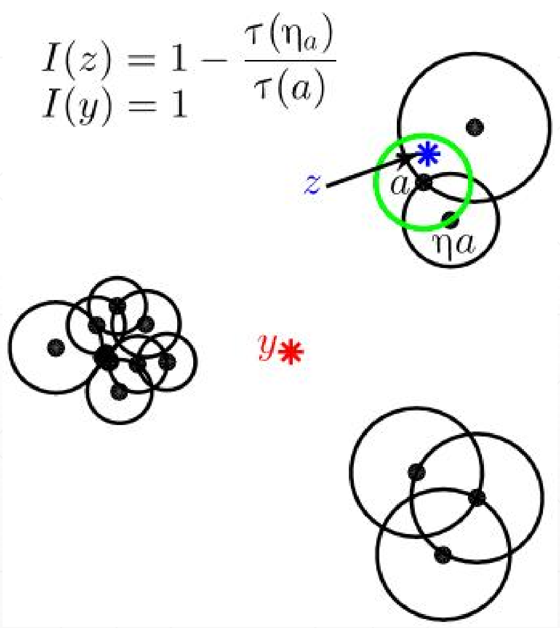
Source: Qin et al 2019

Qin, X., Ting, K. M., Zhu, Y., & Lee, V. C. (2019, July). Nearest-neighbour-induced isolation similarity and its impact on density-based clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 4755-4762).

Isolation mechanism comparison

- Isolation mechanism
  - hyper-rectangles (iForest) vs hyper-spheres (iNNE)
  - t sets of hyper-rectangles/spheres
  - Each set is constructed from a small sample ($\psi \ll n$)
- Measure: path length vs radius of hyper-sphere
- Anomaly Score: average measure over t sets

iNNE

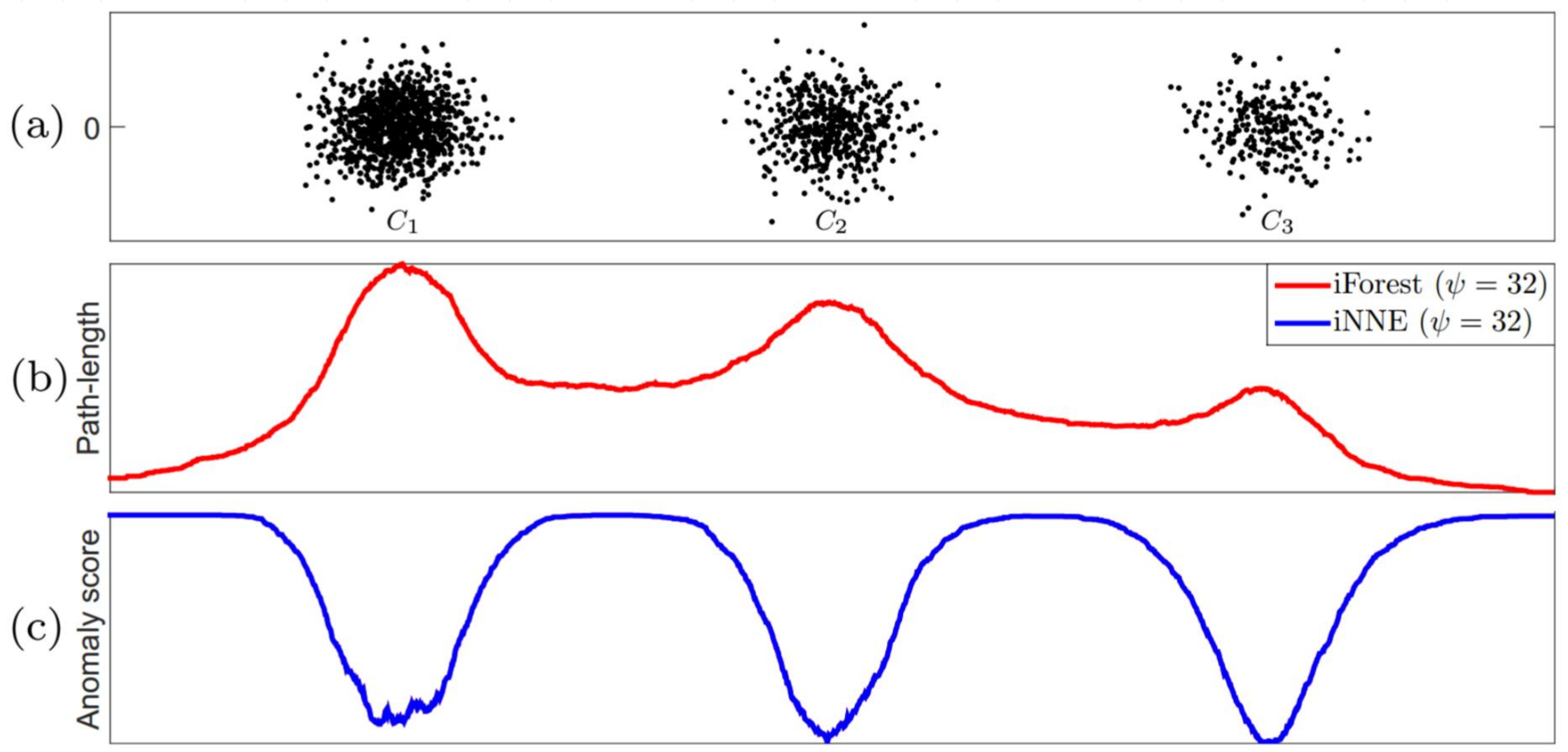


$$I(z) = 1 - \frac{\tau(\eta_a)}{\tau(a)}$$
$$I(y) = 1$$

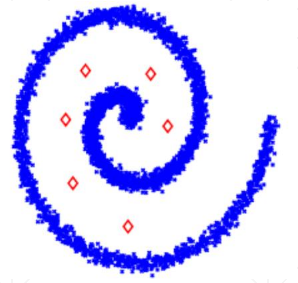
Source: Tharindu et al 2018

Bandaragoda, T.R., Ting, K.M., Albrecht, D., Liu, F.T., Zhu, Y. and Wells, J.R., 2018. Isolation‐based anomaly detection using nearest‐neighbor ensembles. *Computational Intelligence*, *34*(4), pp.968-998.

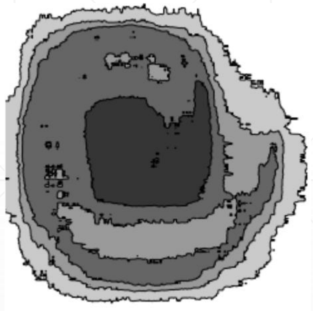
iForest versus iNNE



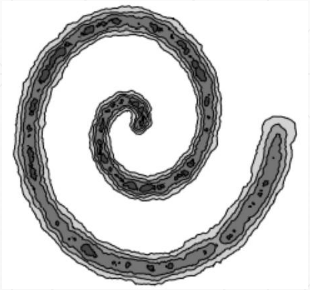
Source: Tharindu et al 2018



6 red anomalies



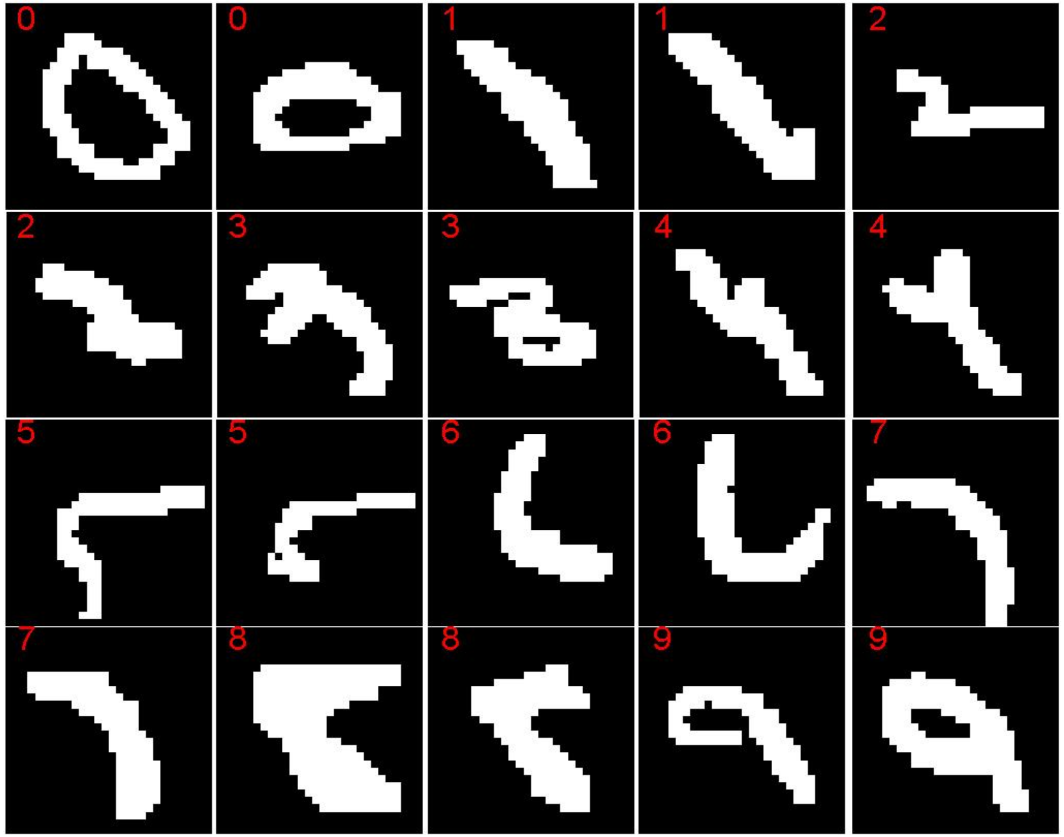
iForest contour



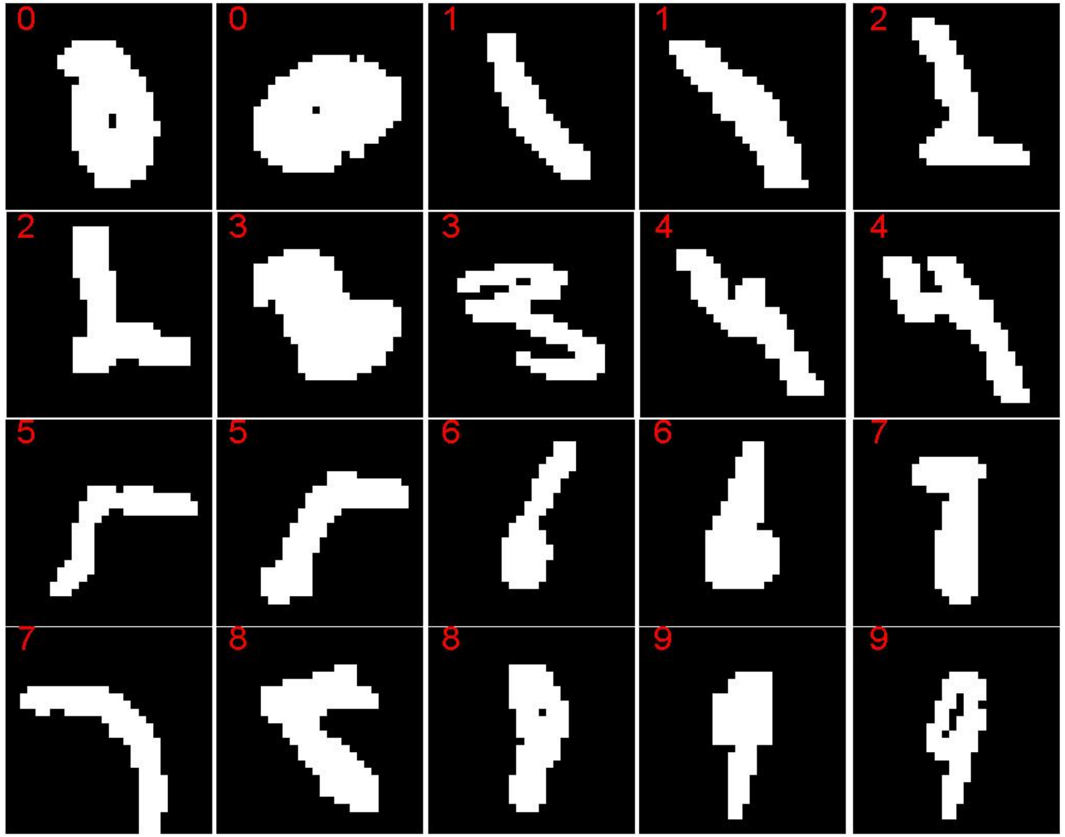
iNNE contour

Example handwritten digits: MNIST top 2 anomalies per digit

iForest

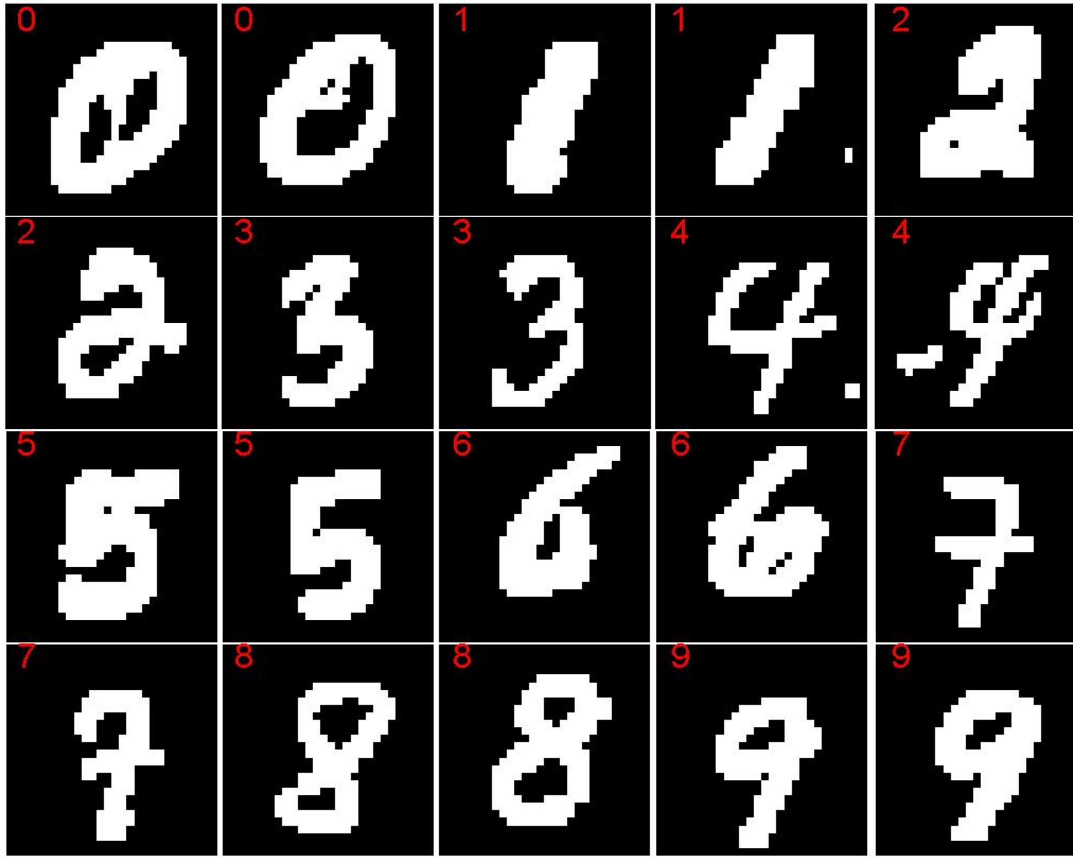


iNNE

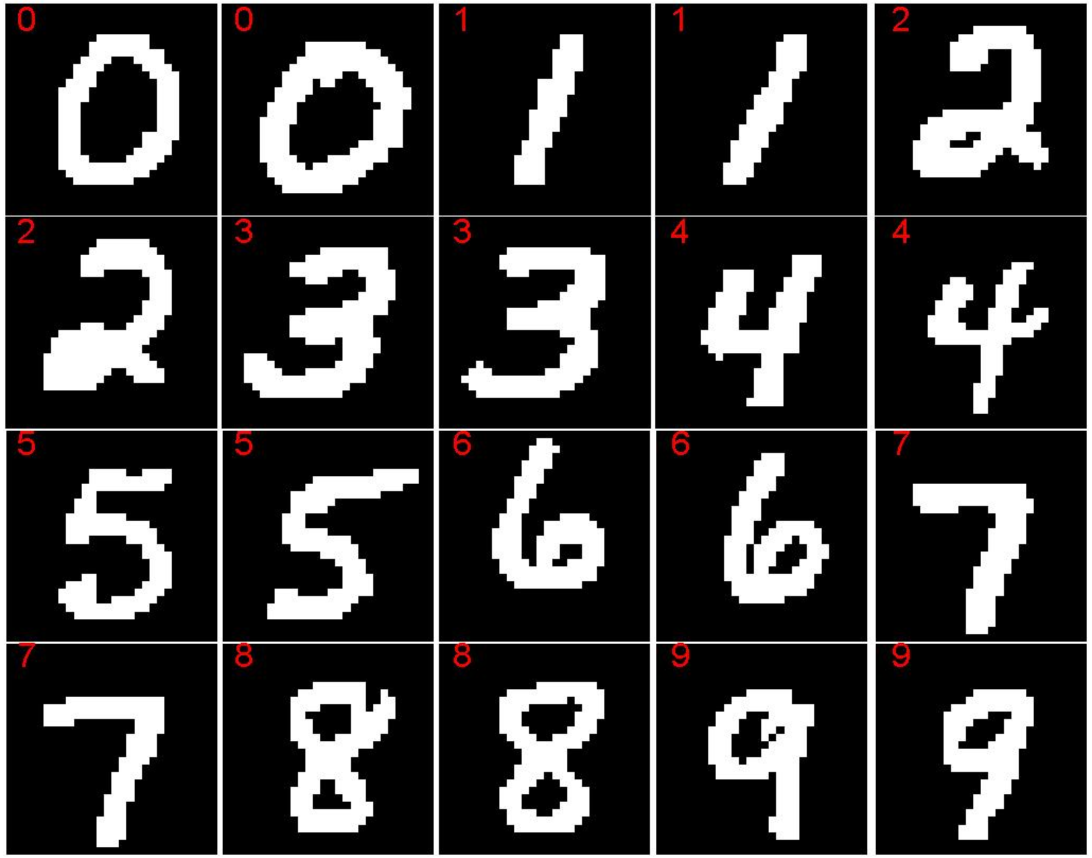


Example handwritten digits: MNIST
bottom 2 anomalies (most typical example) per digit

iForest



iNNE



Data-dependent kernels

Since the idea of Isolation was conceived, it was never confine to point anomaly detection only.

Two notable recent developments:

- **Isolation Kernel** (IK): A data dependent kernel
  - derived from a dataset based on the isolation mechanism,
  - unlike Gaussian kernel, IK has no closed form expression,
  - requires no learning.
- **Isolation Distributional Kernel** (IDK) measures the similarity of two distributions, based on the framework of kernel mean embedding.

Isolation Kernel

The key idea of the Isolation kernel [AAAI19] is using a space partitioning strategy to split the whole data space into ψ non-overlapping partitions based on a random sample of ψ points from a given dataset. The similarity between any two points is the expectation that these two points are found in the same partition.

$$K_{\psi}(\mathbf{x},\ \mathbf{y}|D) = \mathbb{E}_{\mathcal{H}_{\psi}(D)}[\mathbb{I}(\mathbf{x},\ \mathbf{y} \in \theta|\theta \in H)]$$

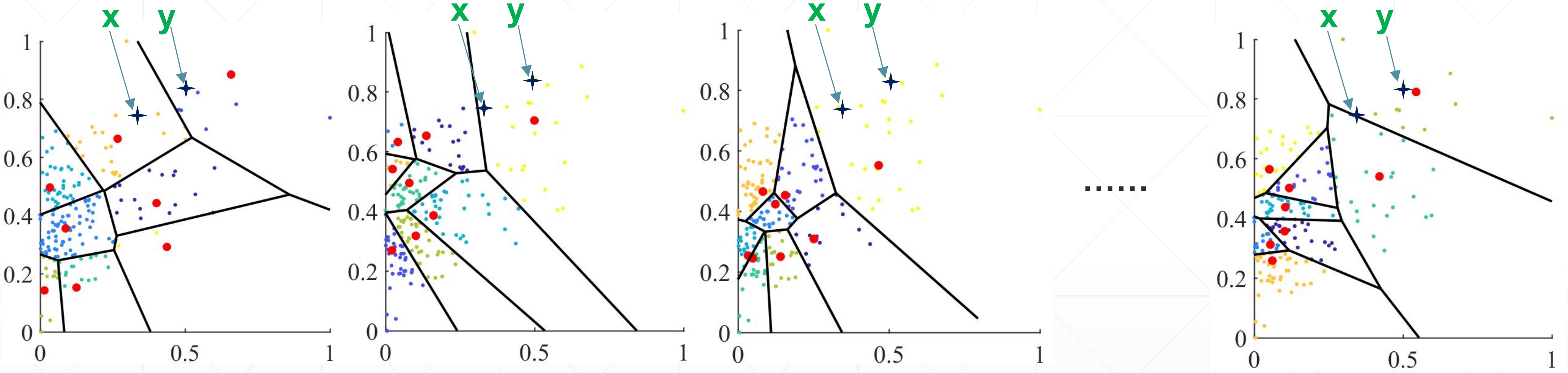
$$\cong \frac{1}{t}\sum_{i=1}^{t} \mathbb{I}(\mathbf{x},\ \mathbf{y} \in \theta|\theta \in H_{i}),$$

where $H \in \mathcal{H}_{\psi}(D)$ is one partitioning based on a subsample with size

and $\mathbb{I}$ is an indicator function.

[AAAI19] Qin, X., Ting, K. M., Zhu, Y., & Lee, V. C. (2019, July). Nearest-neighbour-induced isolation similarity and its impact on density-based clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 4755-4762).

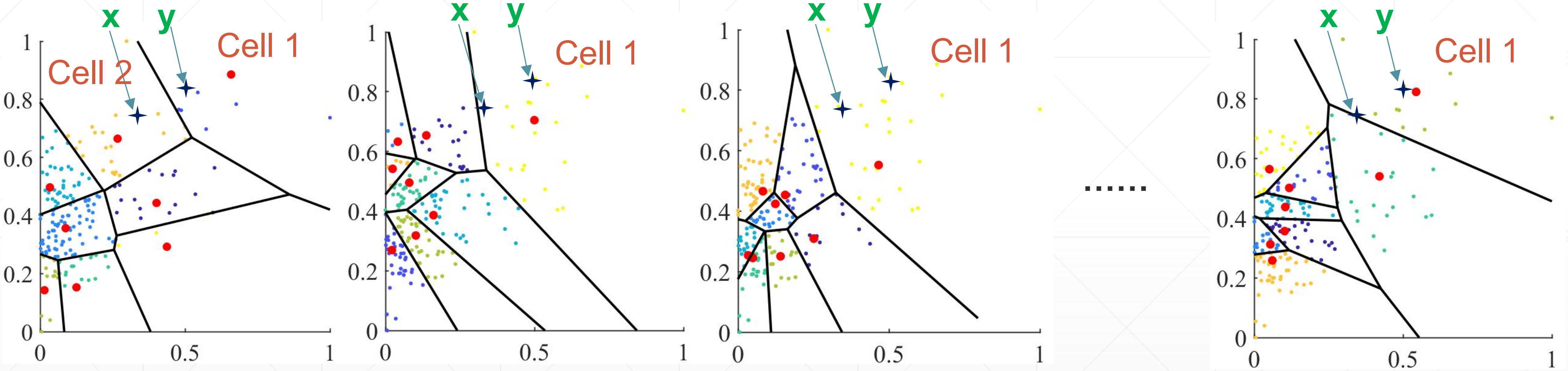
Isolation Kernel Calculation (NN-Voronoi Diagram)



We can use a nearest neighbour method to split a data space into 8 non-overlapping partitions, and independently conduct this partitioning strategy for t=100 trials. If two points $\mathbf{x}$ and $\mathbf{y}$ are located in the same partition (sharing the same nearest subsample point) in 25 out of 100 trials, then the similarity between $\mathbf{x}$ and $\mathbf{y}$ is estimated as 0.25, i.e., $K_8\left(\mathbf{x},\ \mathbf{y}|D\right) = 0.25$.

Isolation Kernel Feature Map

$\Phi(\mathbf{x})$ is a binary vector that represents the partitions in all the partitionings, where $\mathbf{x}$ falls in to only one of ψ cells in each partitioning.

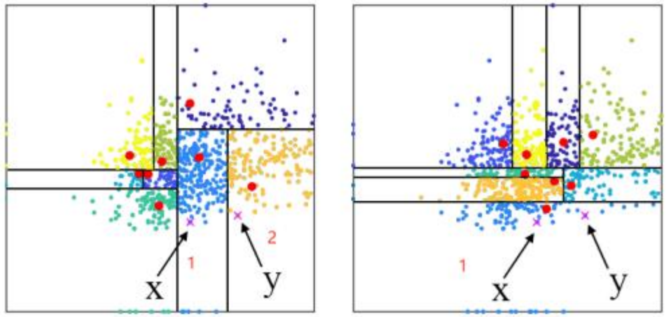


$\Phi(\mathbf{x})$ - > [0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ...... 1 0 0 0
0 0 0 0]
$\Phi(\mathbf{y})$ - > [1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ...... 1 0 0 0
0 0 0 0]

$$K_{\psi}(\mathbf{x},\ \mathbf{y}|D) = \frac{1}{t} < \Phi(\mathbf{x}), \Phi(\mathbf{y})) >$$

Other partitioning strategies

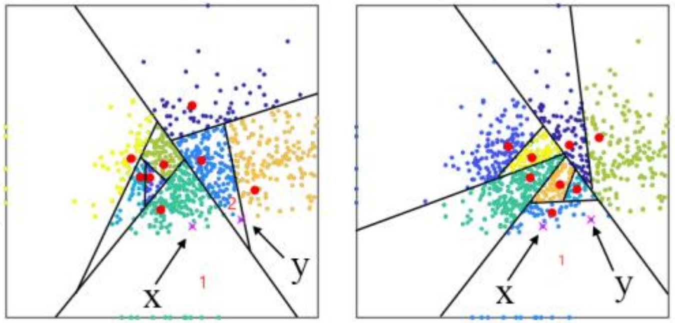
Axis parallel



$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,1\,0\,0\,0\,0\,0\,0]$

$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [1\,0\,0\,0\,0\,0\,0\,0]$

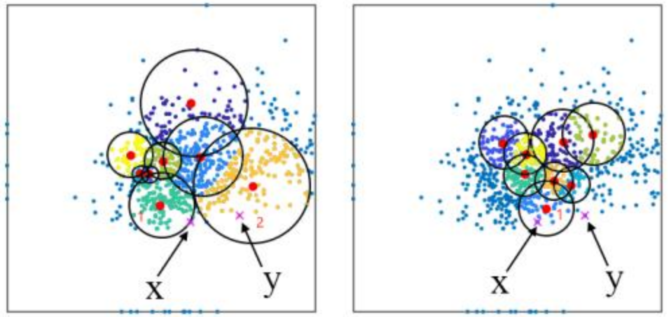
Random hyperplane



$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,1\,0\,0\,0\,0\,0\,0]$

$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [1\,0\,0\,0\,0\,0\,0\,0]$

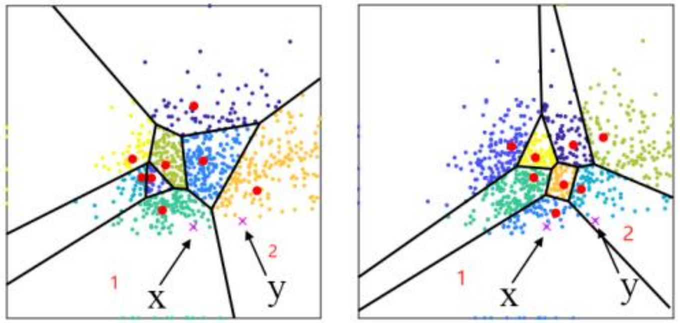
Hypersphere



$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,1\,0\,0\,0\,0\,0\,0]$

$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,0\,0\,0\,0\,0\,0\,0]$

Vonoroi diagram

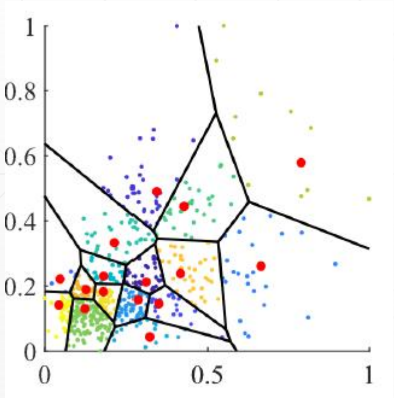


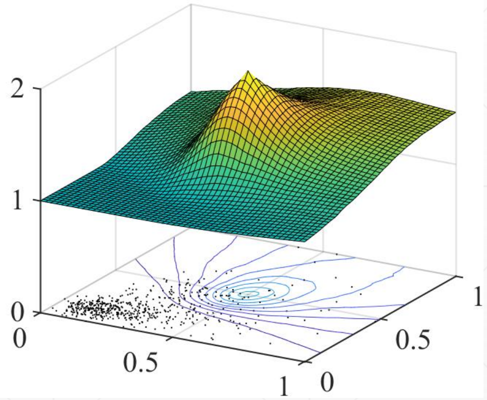
$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,1\,0\,0\,0\,0\,0\,0]$

$\Phi(x) = [1\,0\,0\,0\,0\,0\,0\,0]$
$\Phi(y) = [0\,1\,0\,0\,0\,0\,0\,0]$

Isolation Kernel Properties

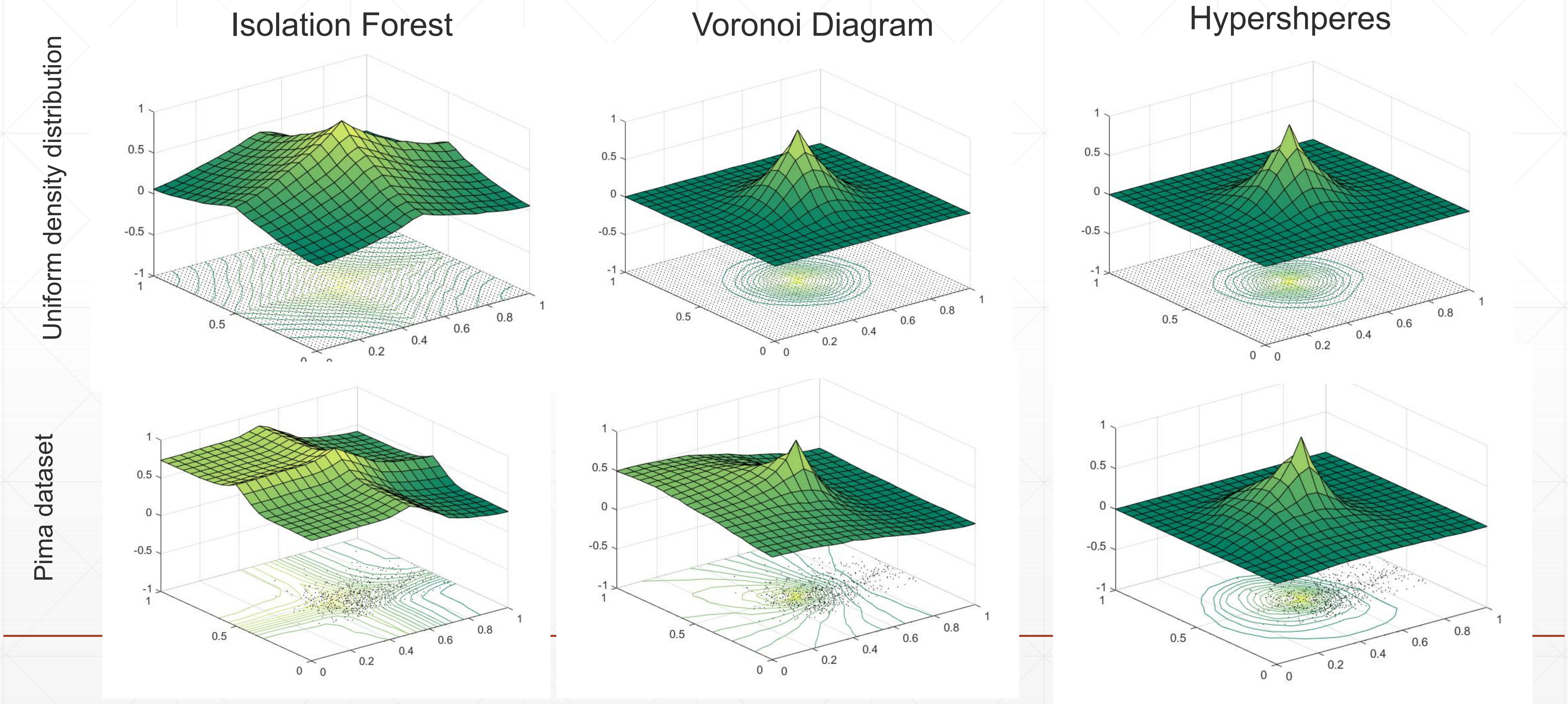
- Isolation kernel adapts to local density distribution. The isolation mechanism of IK produces large partitions in sparse regions and small partitions in dense regions, based on the random subsamples.
- The probability of two points from the dense cluster falling into the same isolating partition is lower than two points of equal inter-point distance from the sparse cluster, i.e., two points in a sparse region are more similar than two points of equal inter-point distance in a dense region.





Contours with reference to point (0.5, 0.5).

Example kernel distributions of Isolation Kernel of three different implementations



Isolation Forest

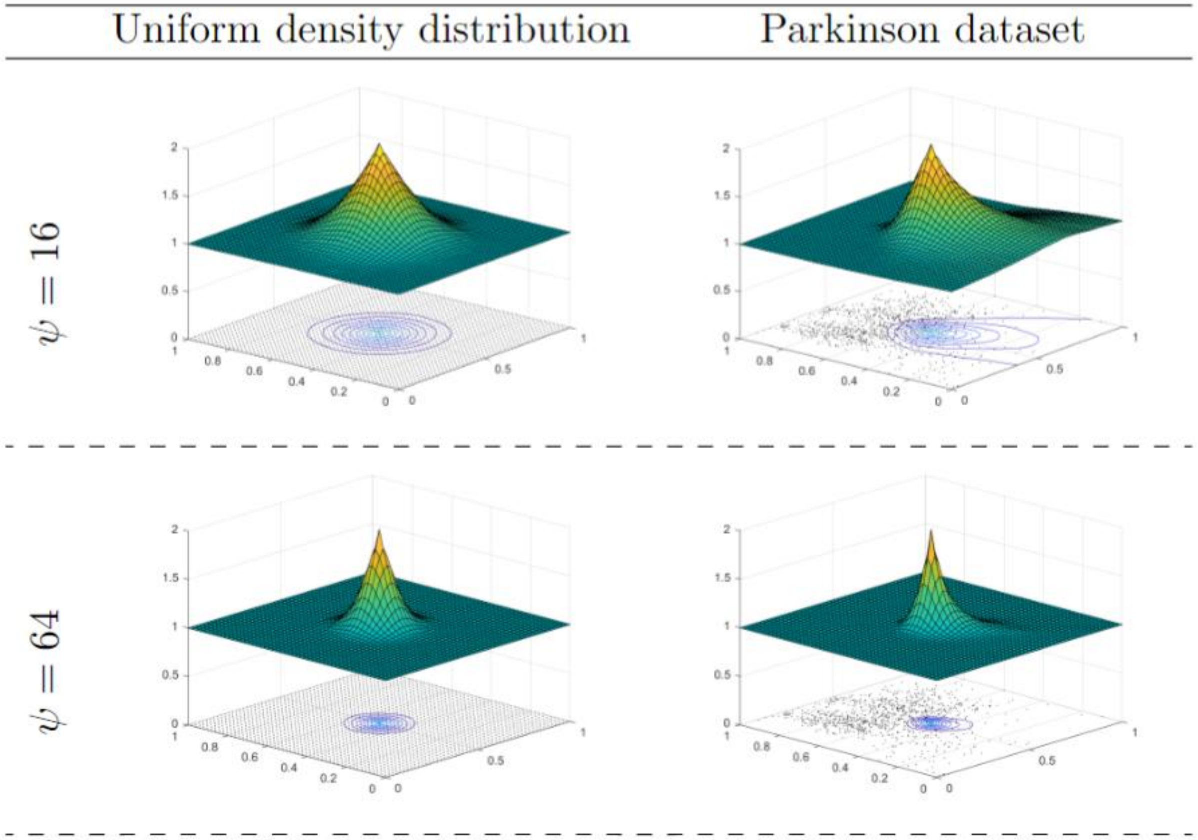
Voronoi Diagram

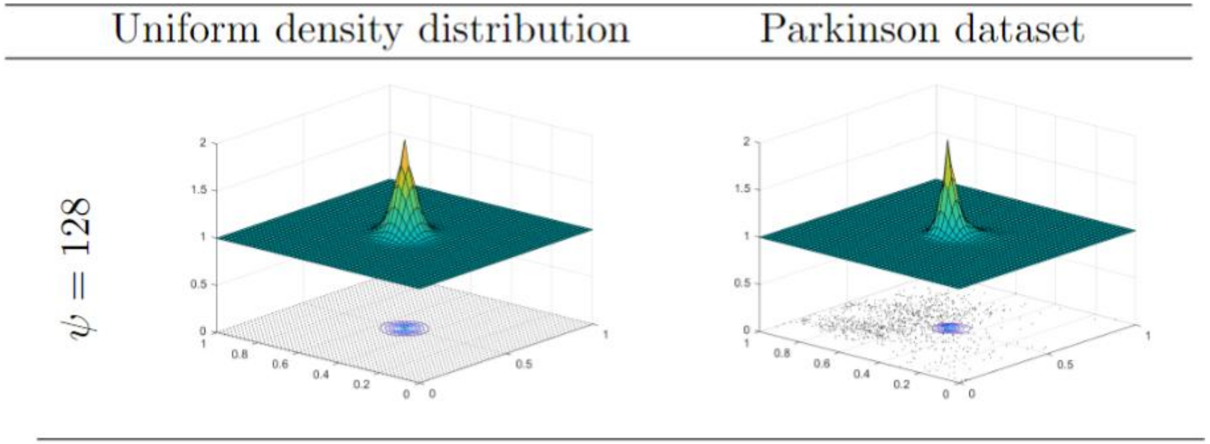
Hypershperes

Uniform density distribution

Pima dataset

Subsample Effects





Contours of the Isolation kernel based on Voronoi Diagram with reference to point (0.5, 0.5)

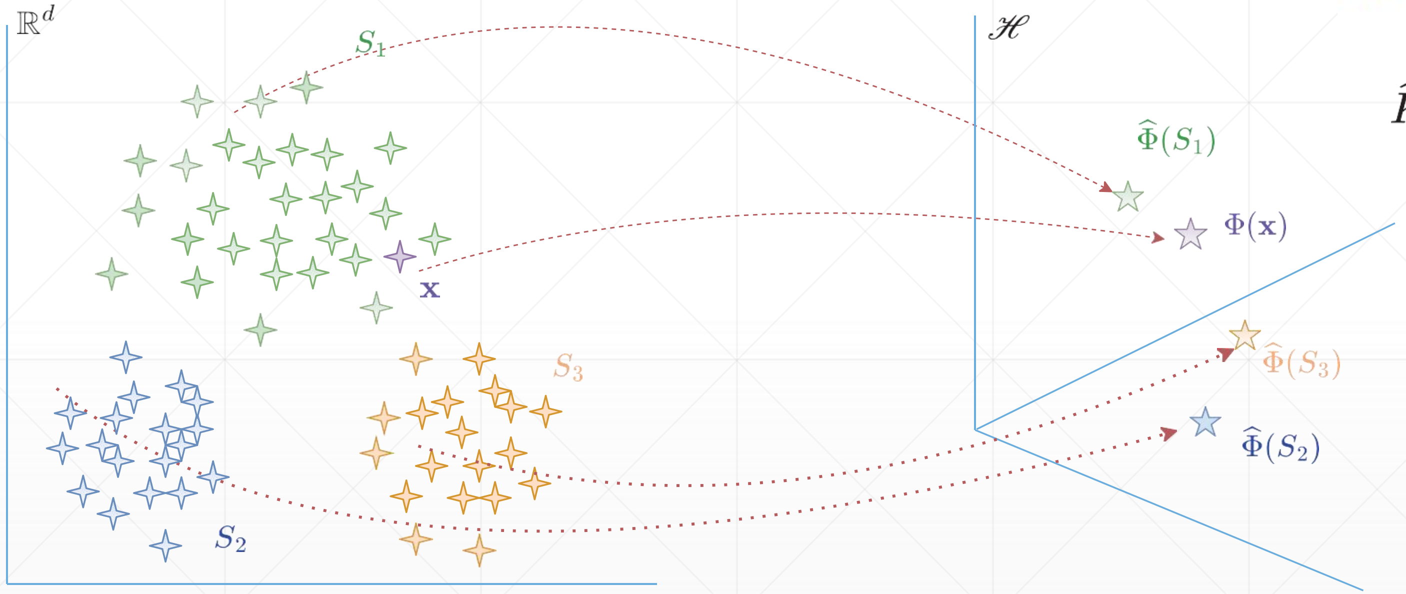
Point-Set Kernel

Given a point $\mathbf{x}$ and a set $A = \{\mathbf{y}_i\}_{i=1}^{p}$, and $\mathbf{x},\ \mathbf{y}_\mathbf{i} \in R^d$, the point-set similarity between $\mathbf{x}$ and A is the average pairwise similarity between $\mathbf{x}$ and every point in A, defined as follows:

$$\widehat{K}_\psi(\mathbf{x},\ \mathrm{A}|D) = \frac{1}{|\mathrm{A}|}\sum_{\mathbf{y}\in A} K_\psi(\mathbf{x},\ \mathbf{y}|D) = \frac{1}{t} < \Phi(\mathbf{x}), \widehat{\Phi}(A) >$$

Where $\widehat{\Phi}(A) = \frac{1}{|A|}\sum_{\mathbf{y}} \Phi(\mathbf{y})$ is the kernel mean map of K_ψ.

Point-Set Kernel (cont.)



we normalise it to $[0,1]$ as

$$\widehat{K}_\psi(\mathbf{x}, A|D) = \frac{\langle \Phi(\mathbf{x}), \widehat{\Phi}(A) \rangle}{\sqrt{\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle} \sqrt{\langle \widehat{\Phi}(A), \widehat{\Phi}(A) \rangle}}$$

$$\widehat{\Phi}(A) = \frac{1}{|A|} \sum_{\mathbf{y} \in A} \Phi(\mathbf{y})$$

Because $\widehat{\Phi}(A)$ can be pre-calculated, estimating the similarity between a point and a set points costs constant time $O(1)$.

Isolation Distributional Kernel (IDK)

$$\widehat{K}(P_S,\ P_T) = \frac{1}{|S||T|}\sum_{x\in S}\sum_{y\in T}\kappa(x, y)$$

1. As κ (Isolation Kernel) is a characteristic kernel, then its kernel mean map is injective, i.e.,

   $\| \widehat{\varphi}(P_S) - \widehat{\varphi}(P_T)\|_H\ = 0$ if and only if $P_S = P_T$.

2. Data dependent property: Two distributions, as measured by IDK derived in sparse region, are more similar than the same two distributions, as measured by IDK derived in dense region.

- Key in improving task-specific performance

3. It has finite-dimensional feature map: $\widehat{K}(P_S,\ P_T) = \langle\widehat{\Phi}(P_S), \widehat{\Phi}(P_T)\rangle$

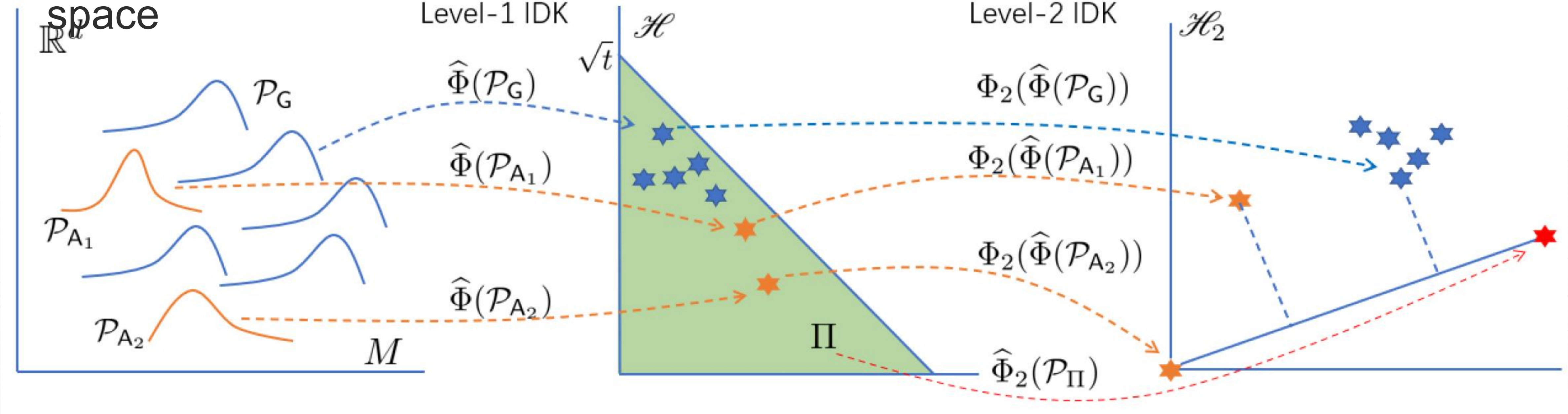
- Key in low time complexity

Application: Group Anomaly Detection

IDK^2 : Using two levels of IDK to detect group anomalies [KDD20, TKDE22]

Level-1 maps each group to a point in Level-1 Hilbert space

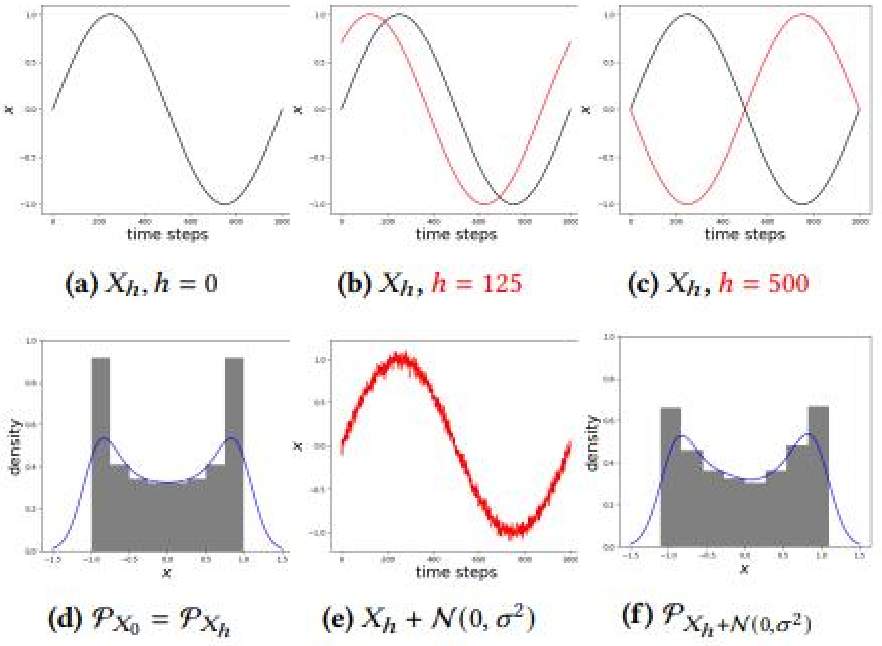
Level-2 maps level-1 pts and the set of level-1 pts to pts in Level-2 Hilbert space


$\mathbb{R}^d$
Level-1 IDK
$\mathcal{H}$
Level-2 IDK
$\mathcal{H}_2$
$\sqrt{t}$
$\mathcal{P}_G$
$\widehat{\Phi}(\mathcal{P}_G)$
$\Phi_2(\widehat{\Phi}(\mathcal{P}_G))$
$\widehat{\Phi}(\mathcal{P}_{A_1})$
$\Phi_2(\widehat{\Phi}(\mathcal{P}_{A_1}))$
$\mathcal{P}_{A_1}$
$\Phi_2(\widehat{\Phi}(\mathcal{P}_{A_2}))$
$\widehat{\Phi}(\mathcal{P}_{A_2})$
$\mathcal{P}_{A_2}$
M
Π
$\widehat{\Phi}_2(\mathcal{P}_\Pi)$

[TKDE22] K.M. Ting, B.-C. Xu, T. Washio, Z.-H. Zhou. Isolation Distributional Kernel: A New Tool for Point and Group Anomaly Detections. IEEE Transactions on Knowledge and Data Engineering (2022).

Application: Time Series Anomaly Detection

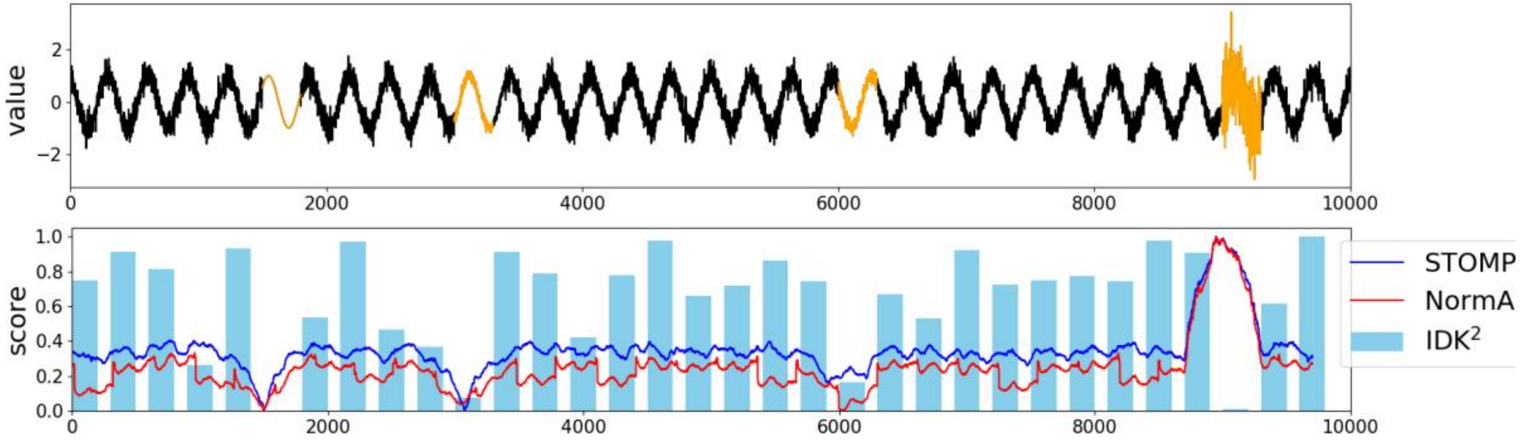
- A new treatment for timeseries. This is a paradigm shift from the time domain and frequency domain approaches that have been around for more than 100 years [VLDB22].



(a) $X_h, h = 0$ (b) X_h, $h = 125$ (c) X_h, $h = 500$

(d) $\mathcal{P}_{X_0} = \mathcal{P}_{X_h}$ (e) $X_h + \mathcal{N}(0, \sigma^2)$ (f) $\mathcal{P}_{X_h+\mathcal{N}(0,\sigma^2)}$

Figure 1: Example sine waves (with m = 1000) and their pdfs

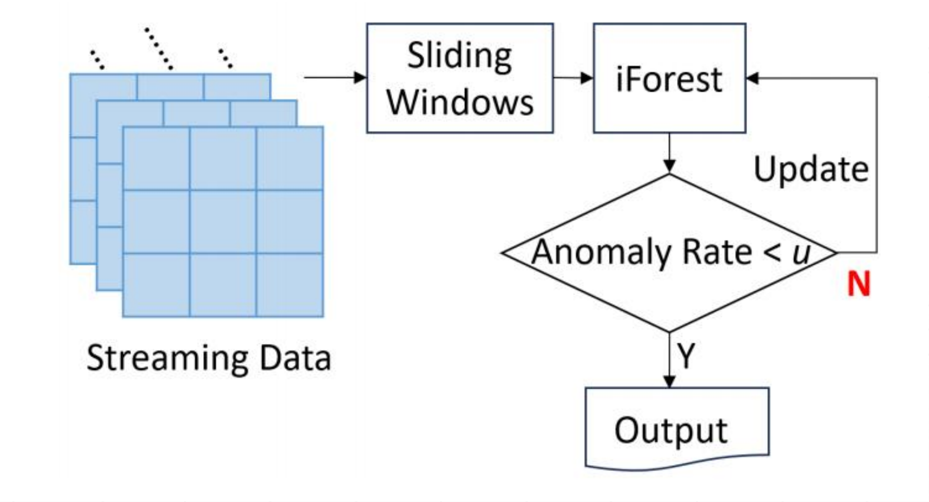


(a) Noisy_sine

Kai Ming Ting, Zhongyou Liu, Hang Zhang, Ye Zhu (2022) A New Distributional Treatment for Time Series and an anomaly detection investigation. To appear in VLDB22

Application: Streaming Anomlay Detection

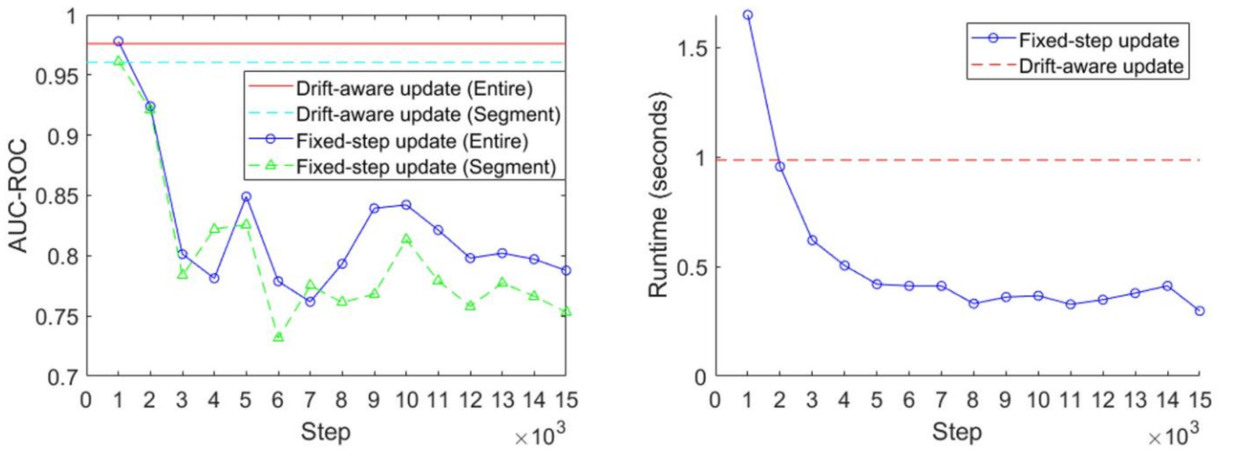
Conditional update - iForestASD



Ding Z, Fei M (2013) An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proc Vol 46(20):12–17

Reconstruct strategy - iForests, iNNEs, IDKs

Keep updating the models using a sliding window, i.e., given a point, build the anomaly detection models based on the preceding and succeeding windows, and then get the anomaly scores for both windows. The final anomaly score is the smaller one between them.



(a) AUC-ROC

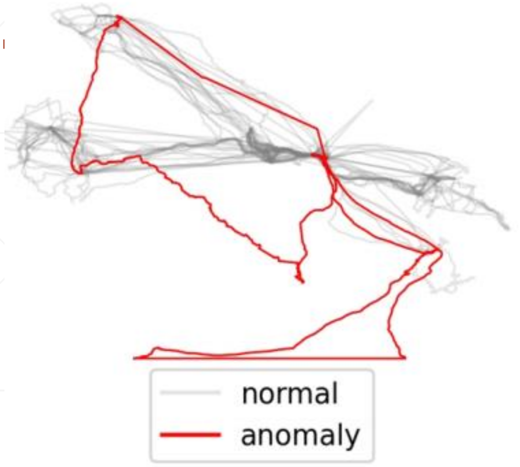
(b) Runtime

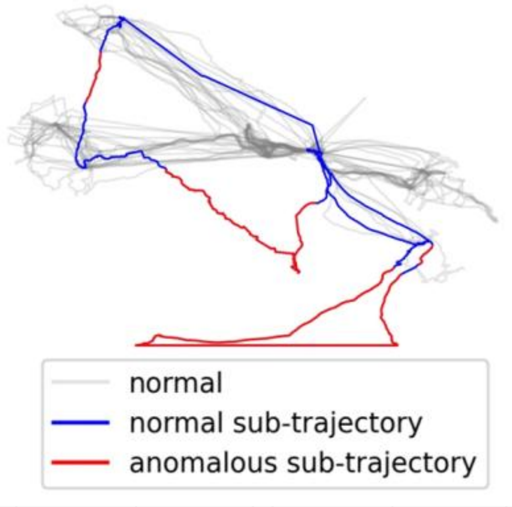
Cao, Yang, Yixiao Ma, Ye Zhu, and Kai Ming Ting. "Revisiting streaming anomaly detection: benchmark and evaluation." Artificial Intelligence Review 58, no. 1 (2025): 1-24.

Application: Trajectory Anomaly Detec

Trajectory data provide rich information on how people move around a city at various spatial and temporal resolutions, emerging as a critical source of insight into **network traffic dynamics and traveler behaviors**.

- Existing **trajectory distance measures** have high time complexity because their core computations rely on a point-to-point distance measure.
- **Low fidelity of existing distance measures**. Most existing distance measures are sensitive to sampling rate, outliers, spatial/temporal lengthening and contraction
- Existing trajectory clustering algorithms have **either effectiveness issues or high time complexity.**





Wang, Yufan, Zijing Wang, Kai Ming Ting, and Yuanyi Shang. "A Principled Distributional Approach to Trajectory Similarity Measurement and its Application to Anomaly Detection." Journal of Artificial Intelligence Research 79 (2024): 865-893.

Parameter Settings and Model Optimisation

- Isolation-based anomaly detection algorithms have two key parameters:

  the sub-sampling size $\psi$and the number of ensembles t.

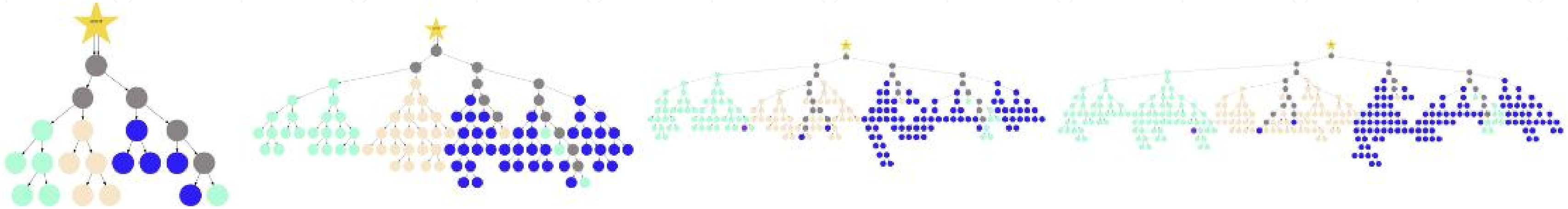
- The original paper of iForest recommends to use ψ= 256 and t = 100 as the default. However, the parameters should be tuned based on a given dataset, so we consider the range of $\psi \in$ $[2^1, 2^2, 2^3 ... 2^{10}$].
- We can fix t to 100 or 200, as a larger value of t tends to produce more stable results but also longer running time.
- Since anomalies are assumed to be rare, we may search for the best parameter ψ in IK-based anomaly detection that leads to maximum stability of anomaly score distribution, with a measure of instability, e.g., variance, entropy or Gini coefficient.

Isolation Mechanisms Applied to Clustering

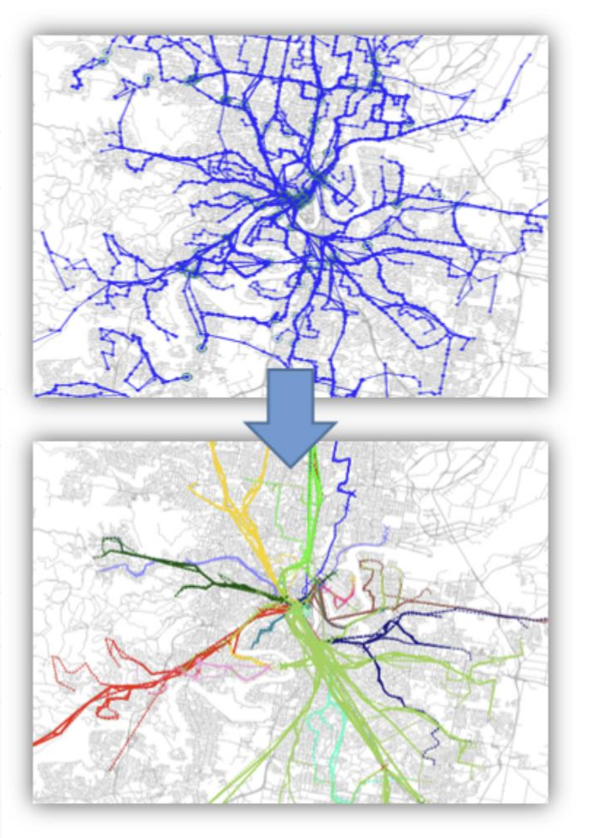
- Using IK similarity to significantly improve existing density-based clustering on datasets with varied densities. [AAAI19]
- A new class of clustering algorithm called psKC (point-set Kernel Clustering). [TKDE22]
    - Up to early 2022, psKC is the only clustering algorithm which is both effective and efficient---a quality which is all but nonexistent in current clustering algorithms. It is also the only kernel-based clustering which has linear time complexity.
- A new class of online Hierarchical Clustering (StreaKHC) that makes use of the idea of point-set kernel. [KDD22]
    - Up to Sep 2022, StreaKHC is the fastest online hierarchical clustering algorithm without proximation or sampling.
- Utilising Distributional kernel for massive trajectory clustering. [ICDM2023]

Clustering Demonstration

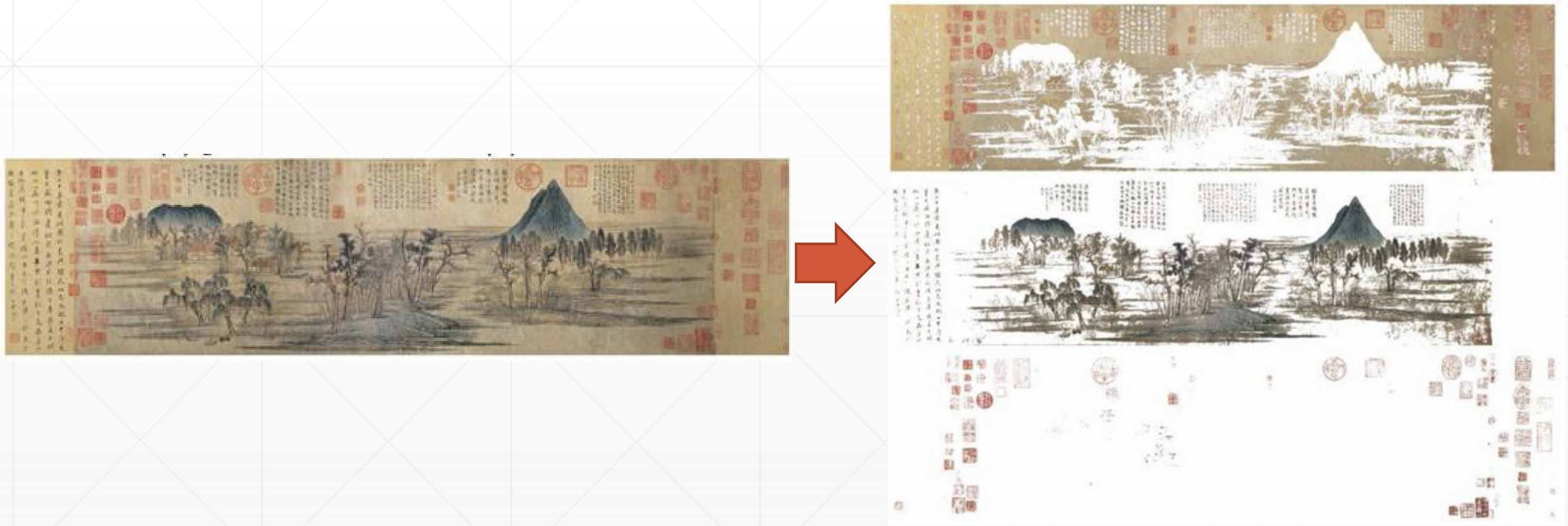
Streaming Hierarchical Clustering



Trajectory clustering

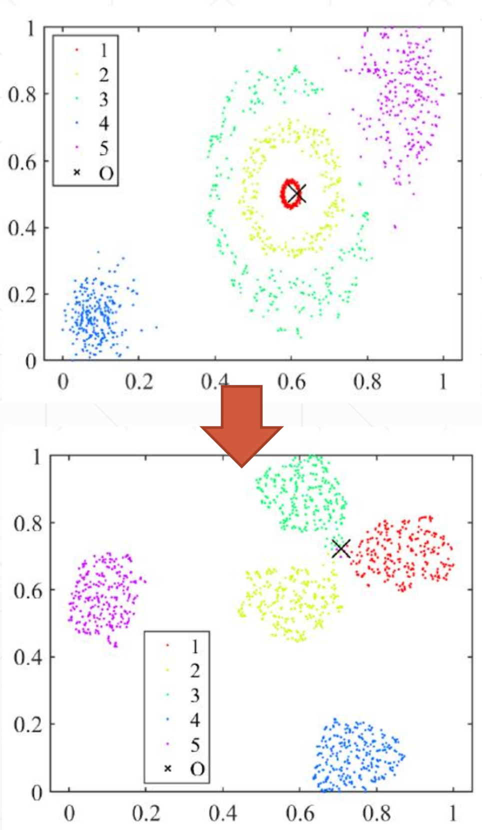


Point-Set Kernel Clustering

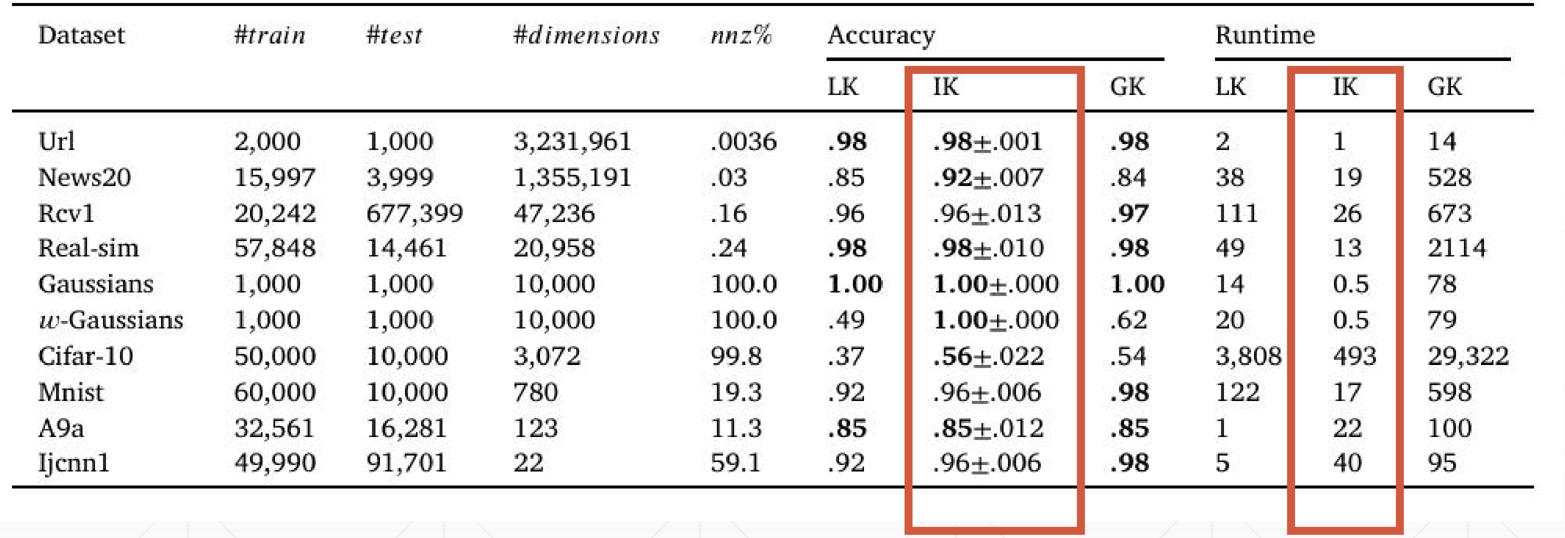


Isolation Mechanisms Applied to High-dimensional Data

Improving the visualisation using t-SNE on subspace clusters [JAIR2021,IJCAI2022]



Improving SVM classification accuracy & runtime [AIJ2024]



Dataset	*#train*	*#test*	*#dimensions*	*nnz%*	Accuracy			Runtime		
					LK	IK	GK	LK	IK	GK
Url	2,000	1,000	3,231,961	.0036	**.98**	**.98±.001**	**.98**	2	1	14
News20	15,997	3,999	1,355,191	.03	.85	**.92±.007**	.84	38	19	528
Rcv1	20,242	677,399	47,236	.16	.96	.96±.013	**.97**	111	26	673
Real-sim	57,848	14,461	20,958	.24	**.98**	**.98±.010**	**.98**	49	13	2114
Gaussians	1,000	1,000	10,000	100.0	**1.00**	**1.00±.000**	**1.00**	14	0.5	78
w-Gaussians	1,000	1,000	10,000	100.0	.49	**1.00±.000**	.62	20	0.5	79
Cifar-10	50,000	10,000	3,072	99.8	.37	**.56±.022**	.54	3,808	493	29,322
Mnist	60,000	10,000	780	19.3	.92	.96±.006	**.98**	122	17	598
A9a	32,561	16,281	123	11.3	**.85**	**.85±.012**	**.85**	1	22	100
Ijcnn1	49,990	91,701	22	59.1	.92	.96±.006	**.98**	5	40	95

Ting, Kai Ming, Takashi Washio, Ye Zhu, Yang Xu, and Kaifeng Zhang. "Is it possible to find the single nearest neighbor of a query in high dimensions?." Artificial Intelligence 336 (2024): 104206.

Zhu, Ye, and Kai Ming Ting. "Improving the effectiveness and efficiency of stochastic neighbour embedding with isolation kernel." Journal of Artificial Intelligence Research 71 (2021): 667-695.

Other works

- Multi-instance learning [KDD19]
- Graph classification via Isolation Graph Kernel [AAAI21]
- IDK can be interpreted as a kernel density estimator called Isolation Kernel Density Estimator [ICDM21, KAIS22]
- More are coming…

[KDD19] Bi-Cun Xu, Kai Ming Ting, Zhi-Hua Zhou (2019) Isolation Set-Kernel and Its Application to Multi-Instance Learning. Proceedings of The ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 941-949.
[AAAI21] Bi-Cun Xu, Kai Ming Ting, Yuan Jiang (2021) Isolation Graph Kernel. Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence. 10487–10495.
[KAIS22] Kai Ming Ting, Takashi Washio, Jonathan R. Wells, Hang Zhang, Ye Zhu (2022) Isolation Kernel Density Estimation. To be appear in Knowledge and Information Systems.

Conclusion

- Isolation-based methods refer to methods that employ an isolation mechanism to construct isolating partitions in the input space.
- Isolation Kernel and Isolation Distributional Kernel are efficient and effective data dependent kernels. They are derived from data directly; and they have no closed form expression and does not require learning.
- Isolation-based methods have been shown to be the key in achieving large scale clustering, visualization, anomaly detection, classification, online kernel learning, etc.

Isolation-based anomaly detection code can be obtained from:
https://shorturl.at/1BLEB

Other Isolation-based methods can be obtained from:
https://github.com/IsolationKernel/Codes