Which Anomaly Detector should I use?

Kai Ming Ting & Sunil Aryal
Federation University Australia

Takashi Washio
Osaka University

Tutorial on 20 November 2018 at IEEE International Conference on Data Mining, Singapore.

Contents

Introduction: Anomaly Detection

- Definition of anomaly:
  - “Something that deviates from what is standard, normal, or expected” Oxford Dictionary
- Example Applications
  - Event detection in sensor networks
  - Disturbance detection in ecosystems
  - System health monitoring
  - Fault detection
  - Intrusion detection.
- Two key approaches:
  - Supervised: must have labels
  - **Unsupervised** : labels not required

Source: license.umn.edu/technologies/20110023_anomaly-detection-algorithm-uses-hybrid-logic-to-predict-system-failure

More examples of applications

- Anomalies in sequence data, e.g., web logs, customer transactions, anomalous subsequences in sequences of animo-acids, network activities (intrusion)
- Spatiotemporal data, e.g., traffic data (anomalous behaviours of moving objects), meteorological data (anomalous weather patterns), disease outbreak data (anomalous trends), medical diagnosis from scans (onset of Alzheimer, sclerosis)
- Graphs: Anomalous compounds in a database of chemical compounds (represented as graphs); anomalous subgraphs in a large graph (web, social network)
- Time series: detection of abrupt change or novelty

Motivations of comparative studies

- What is an anomaly?
- Unsupervised anomaly detection tasks
- Lack of `good' benchmark datasets
- Impact of parameter choices of algorithms
- Lack of `good' performance measure
- Evaluations conducted by proposers of new methods could often be biased

Operational Definitions of Anomalies (1)

1. Distance-based or density-based definitions:

(i) D-Distance Anomalies are instances which have fewer than p neighboring points within a distance D.
(ii) kth NN Distance Anomalies are the top-ranked instances whose distance to the kth nearest neighbor is greatest.
(iii) Average kNN Distance Anomalies are the top-ranked instances whose average distance to the k nearest neighbors is greatest.
(iv) Density-based Anomalies are instances which are in regions of low density or low local/relative density.

Operational Definitions of Anomalies (2)

2. Isolation-based anomalies are instances which are most susceptible to isolation.

3. ZERO++: anomalies are instances which have the highest number of subspaces having zero appearances (over a set of subsamples).

Definition of outlier by Hawkins (1980):

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

Hawkins, D. (1980) Identification of outliers. *Monographs on Applied Probability and Statistics*.

Assessment without a ‘universal’ definition that can provide a `standard’ for performance assessment

- Methods using different operational definitions
- Rely on ground truth provided in the given dataset, and the ground truth is only used for the evaluation of how well a method performs.
- Reality: few or no ground truth

Factors influencing a performance assessment

- The nature of the anomaly detection problem
- The data properties of benchmark datasets: datasets chosen
- The characteristics of algorithms: number of parameters, sensitivity to parameter setting, ensemble or not; how it performs under different conditions
- Evaluation methodology: best performance, test accuracy
- A good measure in assessing the “goodness” of the ranking outcome

Comparative studies (not considered here)

- G. Orair, C. Teixeira, W. Meira Jr, Y. Wang, and S. Parthasarathy. Distance-Based Outlier Detection: Consolidation and Renewed Bearing. *Proceedings of the VLDB Endowment*, 3(1–2), pp. 1469–1480, 2010.
- M. Goldstein and S. Uchida. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PloS One*, 11(4), e0152173, 2016.
- G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B.Micenkova, E. Schubert, I.Assent, and M. E. Houle. On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. *Data Mining and Knowledge Discovery*, 30(4), pp. 891–927, 2016.
- C. C. Aggarwal and S. Sathe. Theoretical Foundations and Algorithms for Outlier Ensembles, *ACM SIGKDD Explorations*, 17(1), June 2015.

Key comparative studies considered here

1. Emmott, A. F., Das, S., Dietterich, T. G., Fern, A., Wong, W.-K. (2016). A Meta-Analysis of the Anomaly Detection Problem. *arXiv*:1503.01158

   - Focus: Assessment design

2. Aggarwal, C. C., Sathe, S. (2017) Outlier Ensembles: An Introduction. Springer International Publishing. (Chapter 6)

   - Focus: From the perspective of bias-variance: parameters, measure

3. 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.

   - Focus: From the analysis of learning curve

Study from Emmott et al (2016)

- Evaluation conducted based on real-world datasets, modified across four dimensions:
  - Point difficulty, relative frequency of anomalies, clusteredness of anomalies & relevant of features
- 8 Methods compared in the study
  - Robust kernel density estimation
  - Ensemble Gaussian Mixture Model
  - One-class SVM & Support vector data description
  - LOF & kNN Angle-based Outlier Detection
  - Isolation Forest (iForest)
  - Lightweight Online Detector of Anomalies (LODA)
- Measures: AUC and Average Precision
- Data properties: Data size: >1k ; #dim < 200;

Assessment design

- Scale of the study:
  - 19 mother datasets
  - 25,685 benchmark datasets
  - 205,480 micro-experiments (8 methods under comparison)

- From a mother dataset, different datasets with the following four dimensions (a few levels for each dimension) are generated:
  - Point difficulty (distance from boundary between normal points and anomalies)
  - Relative frequency (.001, .005, .01, .05, .1 anomalies)
  - Clusteredness (scattered versus clustered anomalies)
  - Feature irrelevance (add irrelevant attributes of different distance ratios: 3 levels)

Recommendations from Emmott et al (2016)

I. Isolation Forest [1] (in general and in large feature space of unknown quality)
II. ABOD [2] and LOF [3] for highly clustered anomalies
III. Isolation Forest [1] and LODA [4] scale well to large data sets while the other algorithms do not.

- ".. over-positive results reported in literature are not particularly helpful measuring progress."

[1] Liu, F. T., Ting, K. M., Zhou, Z .H. (2008) Isolation Forest. In *Proceedings of IEEE ICDM*, 413-422.
[2] Kriegel, H-P., Schubert, M., Zimek, A. (2008) Angle-based outlier detection in high-dimensional data. In *Proceedings of KDD*, 444-452.
[3] Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J. (2000) LOF: Identifying density-based local outliers. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 93-104
[4] Pevny, T. (2016) Loda: Lightweight On-line Detector of Anomalies. *Machine Learning*, 102(2), 275–304.

Study from Aggarwal & Sathe (2017)

- Based on bias-variance decomposition of error of an anomaly detector
- Focus on ensemble anomaly detectors
- 9 Methods compared
  - iForest (varying data subsampling)
  - kNN, Average kNN & LOF
  - Kernel Mahalanobis & Linear Mahalanobis
  - GASP (*Group-wise attribute selection and prediction)*
  - ALSO (*attribute-wise learning and scoring outliers*)
  - TRINITY
- 12 UCI datasets + a few artificial datasets:
  - Dara size: 150 – 6k; #dim: 6 – 166; %anomalies: 1.2% - 9.6%
- Measure: AUC (median)

Issues of parameter setting

- Parameter setting
  - '..compare the algorithms over a range of "reasonable" parameter values because the best values of the parameters are often not known.'
  - 'Excellent performance in a narrow and "hard-to-guess" range is not helpful for an analyst in the unsupervised setting.'
  - 'Using the best value of the parameters in each case will only favor the most *unstable* algorithm with multiple parameters.'
- The stability of the methods with different choices of *datasets*.
- How to compare especially when some parameters are data-size sensitive? Use variable subsampling, in order to vary the data size over a fixed set of parameters that are heuristically chosen.
- Compare box-plots instead of performances at individual parameter choices
- Measure: median AUC of box-plot

Summarised Result (Aggarwal & Sathe, 2017)

KNN	AKNN	MAH-L	MAH-N	IF-VS	TRINITY
0.866	0.842	0.845	0.850	0.873	0.874

Average over 12 datasets

Each dataset: median AUC

Recommendation from Aggarwal & Sathe (2017)

1. TRINITY [5] (a meta-ensemble of the following three base methods)
2. iForest
3. kNN
4. Kernel Mahalanobis – a soft version of kernel PCA [6]

Conclusions from Aggarwal & Sathe (2017):

- iForest do extremely well
- Distance-based detectors are robust
- No detector was able to consistently perform better than all other detectors.

[5] Aggarwal, C. C., Sathe, S. (2017) Outlier Ensembles: An Introduction. (Chapter 6)
[6] Hoffmann, H. (2007). Kernel PCA for Novelty Detection, *Pattern Recognition*, 40(3), 863–874.

Study from Ting et al (2017)

- The only study that explains the behaviour of anomaly detectors in terms of learning curve (increasing data sizes) using a theoretical analysis based on computational geometry.
- It shows that NN anomaly detectors have gravity-defiant behaviour, contrary to the conventional wisdom.
- Focus on theoretical analyses of the behaviour of NN anomaly detectors only: iNNE, aNNE and kNN
- The experiments are designed to verify the analytical result: NN anomaly detectors have gravity-defiant behaviour
- Methodology: 20 trials of train-and-test; measure: AUC;
- 8 datasets with data size: 5k – 300k; #dim: 3 – 5408; %anomalies: .03-2.3
- Variations in data properties: %anomalies, geometry, number of clusters

Analytical result in a nutshell

Detection accuracy of 1NN anomaly detector is influenced by three factors:

(1) the proportion of normal instances (or anomaly contamination rate),

(2) the nearest neighbour distances of anomalies in the dataset,

(3) sample size used by the anomaly detector where the geometry of normal instances and anomalies is best represented by the optimal sample size.

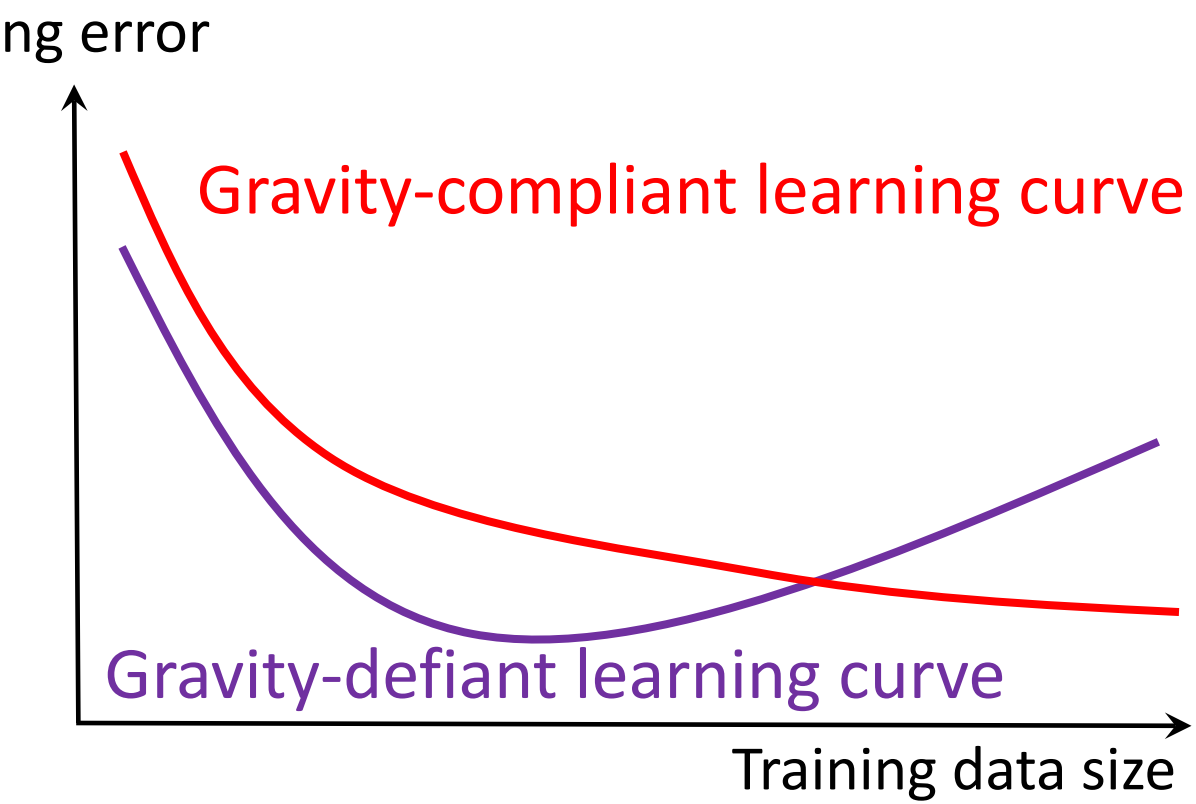
Analytical result in plain language

In sharp contrast to the conventional wisdom: more data the better, the analysis reveals that sample size has three impacts which have not been considered by the conventional wisdom.

1. Increasing sample size increases the likelihood of anomaly contamination in the sample;
2. Increasing the sample size decreases the average nearest neighbour distance to anomalies.
3. The optimal sample size depends on the data distribution.

Learning curve: gravity-defiant behaviour

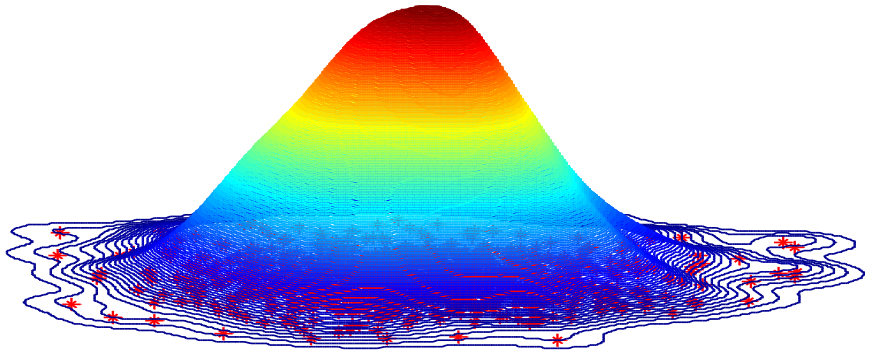
~~More data the better~~



- In a complex domain such as natural language processing, millions of additional data has been shown to continue to improve the performance of trained models (Halevy et al. 2009; Banko and Brill, 2001).

Intuition of why small data size can yield the best performing 1NN ensembles

- The gravity-defiant result does not imply that less data the better or in the limit zero-data does best.
- But it does imply that, under some data distribution, it is possible to have a good performing aNNE where each model is trained using one instance only!



- The geometry of normal instances and anomalies plays one of the key roles in determining the optimal sample size—that signifies the gravity-defiant behaviour of 1NN based anomaly detectors.

Recommendations from Ting et al (2017) [7]

- Recommend aNNE (an ensemble of 1NN) over kNN
  - Easier to tune ψ (aNNE) than k (kNN), as the latter is sensitive data size.
  - aNNE has significantly smaller optimal sample size than kNN.
  - Run significantly faster in large datasets
- Recommend iNNE over aNNE because the former reaches its optimal detection accuracy with a significantly smaller sample size.

- Note that the same gravity-defiant behaviour is observed in iForest too.

[7] 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.

Top recommended anomaly detectors from the three studies (1)

- Isolation based methods: iForest and iNNE [8]
  - All known weaknesses of iForest are overcome by iNNE, with higher time cost; but still has significantly lower time cost than kNN.
- Nearest neighbour-based methods: aNNE and kNN
  - Clustered anomalies: ABOD & LOF
  - iNNE [5] can do well in detecting clustered anomalies.
- Kernel Mahalanobis [6]: The key weakness is the time cost

[6] Hoffmann, H. (2007). Kernel PCA for Novelty Detection, *Pattern Recognition*, 40(3), 863–874.
[8] Bandaragoda, T. R., Ting, K. M., Albrecht, D., Liu F. T., Wells, J. R. (2018). Isolation-based Anomaly Detection using Nearest Neighbour Ensembles. *Computational Intelligence*. Doi:10.1111/coin.12156.

Top recommended anomaly detectors from the three studies (2)

- Among the compared methods, iForest and iNNE have the highest detection accuracy and also have the lowest time cost.
- Simple methods should be used as the baseline to justify any more complicated methods.

Contents

1. Introduction
2. Current comparative works and their recommendations
3. An analysis of strengths and weaknesses of current comparative works
   - Descriptions of the recommended methods
4. Current bias-variance analyses applied to anomaly detection
5. Dealing with high-dimensional datasets and subspace anomaly detection
6. Non-obvious applications of anomaly detectors
7. Potential future research directions
8. Factors to consider in choosing an anomaly detector

Strengths and limitations of individual studies

- Emmotte et al (2016)
  - Focus on average performance over different properties of datasets
  - Not meant to identify specific strengths and weaknesses of an algorithm
- Aggarwal & Sathe (2017)
  - From the perspective of bias and variance of ensembles; parameter issues
  - Not meant to identify specific strengths and weaknesses of an algorithm
- Ting et al (2017)
  - Focus on nearest neighbour algorithms from the perspective of their behaviour in terms of learning curve (as data size increases)

	Emmotte et al	Aggarwal & Sathe	Ting et al
Performance focus	Average over various data properties	Average over different parameter settings	Conditions under which a method works well/poorly
Behaviour under increasing data size	Not examined	Not examined	focus
Uniqueness	Huge variety of data properties are examined	Median performance	Geometry of data distribution are taken account
Parameter search	Default or limited search	Standard variable subsampling	A set range is searched.

Common limitations of all three studies

Datasets:

- Numeric datasets only
- Datasets derived from classification datasets
- The benchmark datasets used
- Almost all datasets are low dimensional datasets (< 500 dimensions)

Evaluation:

- Primary focus is on anomaly detection accuracy:
- Runtime and space complexity were not a key consideration

Algorithms:

- Algorithms which can deal with numeric attributes only.
- Not algorithms which can deal with high dimensional datasets (>1000 #dim)

The kind of comparative studies required

- Compare new methods with the best known current methods (baseline)
- The nature of the anomaly detection problem
- The data properties of benchmark datasets
  - Do better than `Classification datasets are converted into benchmark datasets'.
- The characteristics of algorithms: #parameters, sensitivity to parameter, how a method performs under different conditions
- Evaluation methodology: depends on what you want to achieve
  - Use the whole dataset for training and testing; or use train-and-test instead?
  - Automate parameter tuning: is this necessary?
- A good measure in assessing the "goodness" of the ranking outcome
  - not a key issue
- Algorithms which can deal with high dimensional (>1000) datasets

Characteristics of Ideal Anomaly Detectors

- Few parameters
  - parameter-free the best
  - Easy to tune; not too sensitive to parameter setting
- Fast runtime: can scale up to large datasets and high dimensional datasets
- Low space complexity
- Known behaviours under different data properties
- Can deal with different types of anomalies:
  - Clustered anomalies vs scattered anomalies
  - Local anomalies vs global anomalies

Descriptions of the recommended methods

- Isolation-based methods
- KNN methods

Isolation-based methods

- Isolation based methods: iForest, iNNE & aNNE
- They have all the characteristics one can wish for of an anomaly detector

Features:

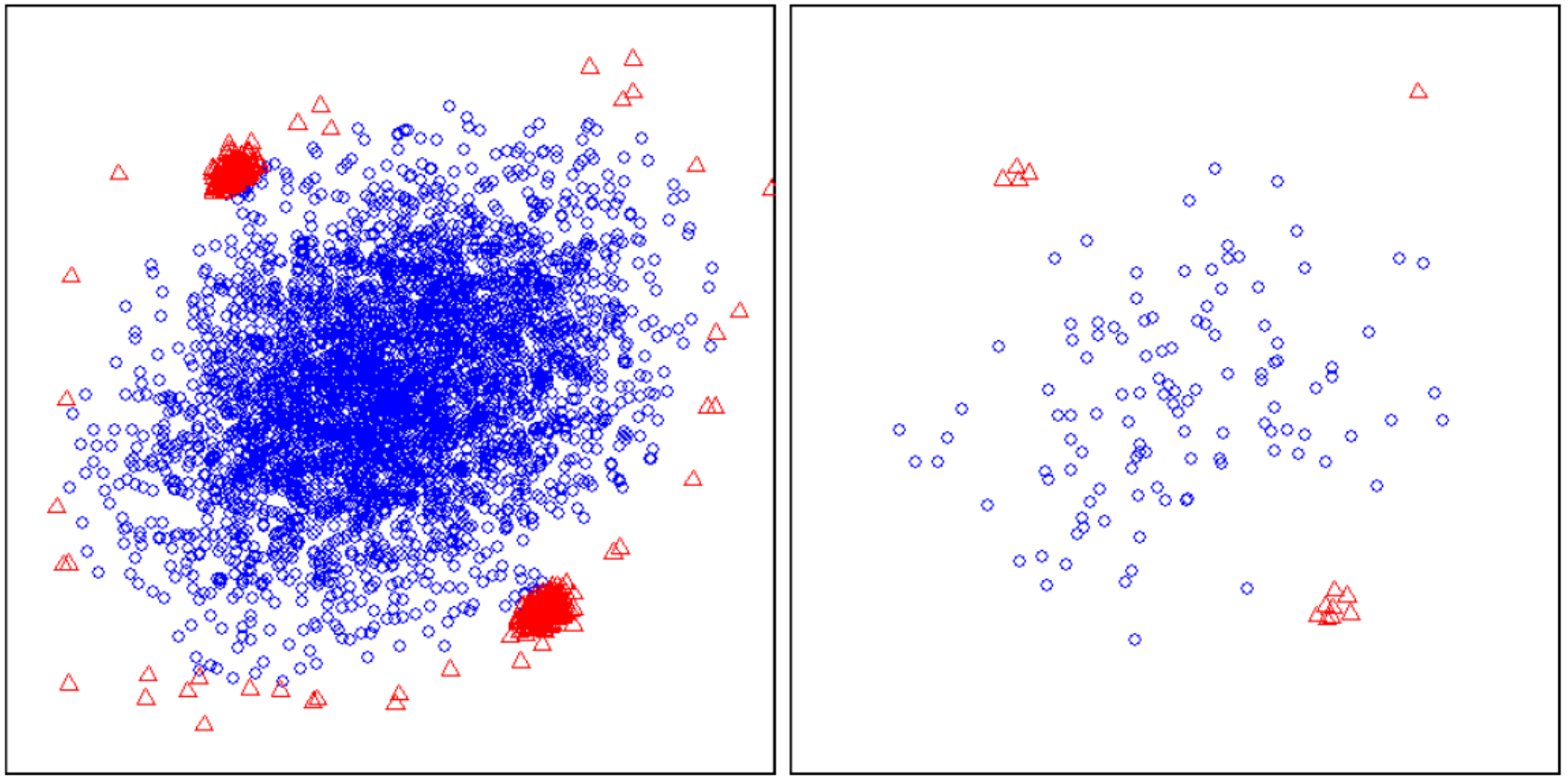
- Isolation mechanism to partition the data space (tree-based or NN-based)
- Data subsampling

Data subsampling

- Its significance in anomaly detection
- Start with iForest (ICDM2008), iNNE (workshop ICDM2014), aNNE (workshop ICDM2015)
- Theoretical Analyses: NIPS2013, MLJ2017
- Not just for efficiency as claimed by Aggarwal & Sathe (2017)

The effect of subsampling

- Reduce both the swamping and masking effects



(a) Original sample (4096 instances)

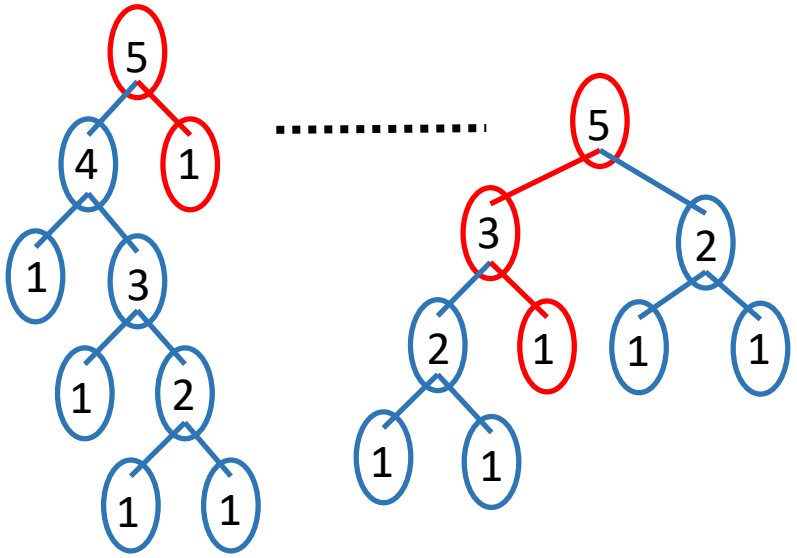
(b) Sub-sample (128 instances)

Source: Tony Liu's PhD Thesis – Anomaly Detection by Isolation

2. iForest (Liu et al ICDM 2008)

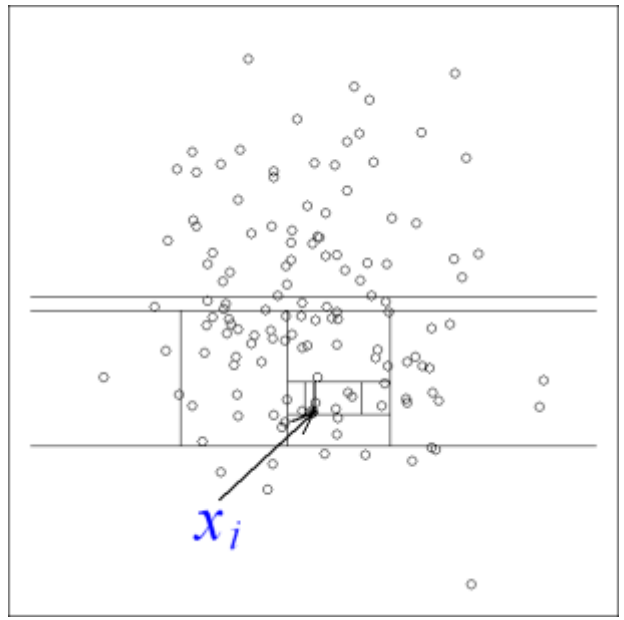
- A collection of isolation trees (iTrees)
- Each iTree isolates every instance from the rest of the instances in a given sample
- Anomalies are 'few and different'
  - More susceptible to isolation
  - Shorter average path

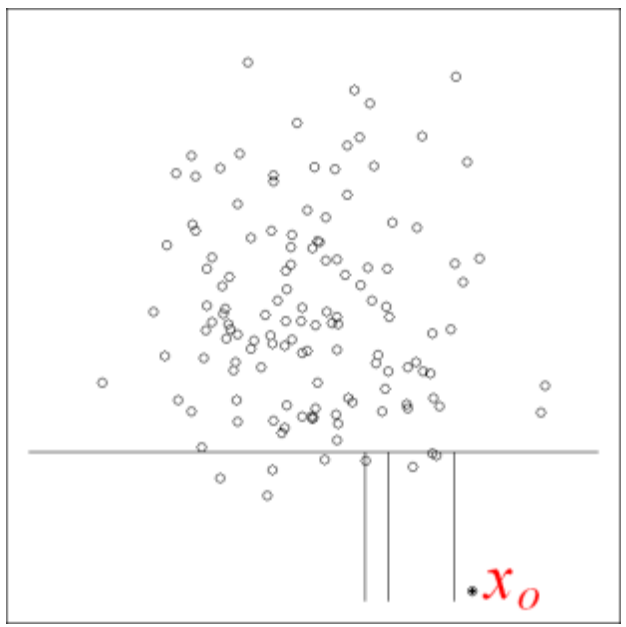
$$Score(\boldsymbol{x}) = \frac{1}{t}\sum_{i=1}^{t} \ell_i(\boldsymbol{x})$$

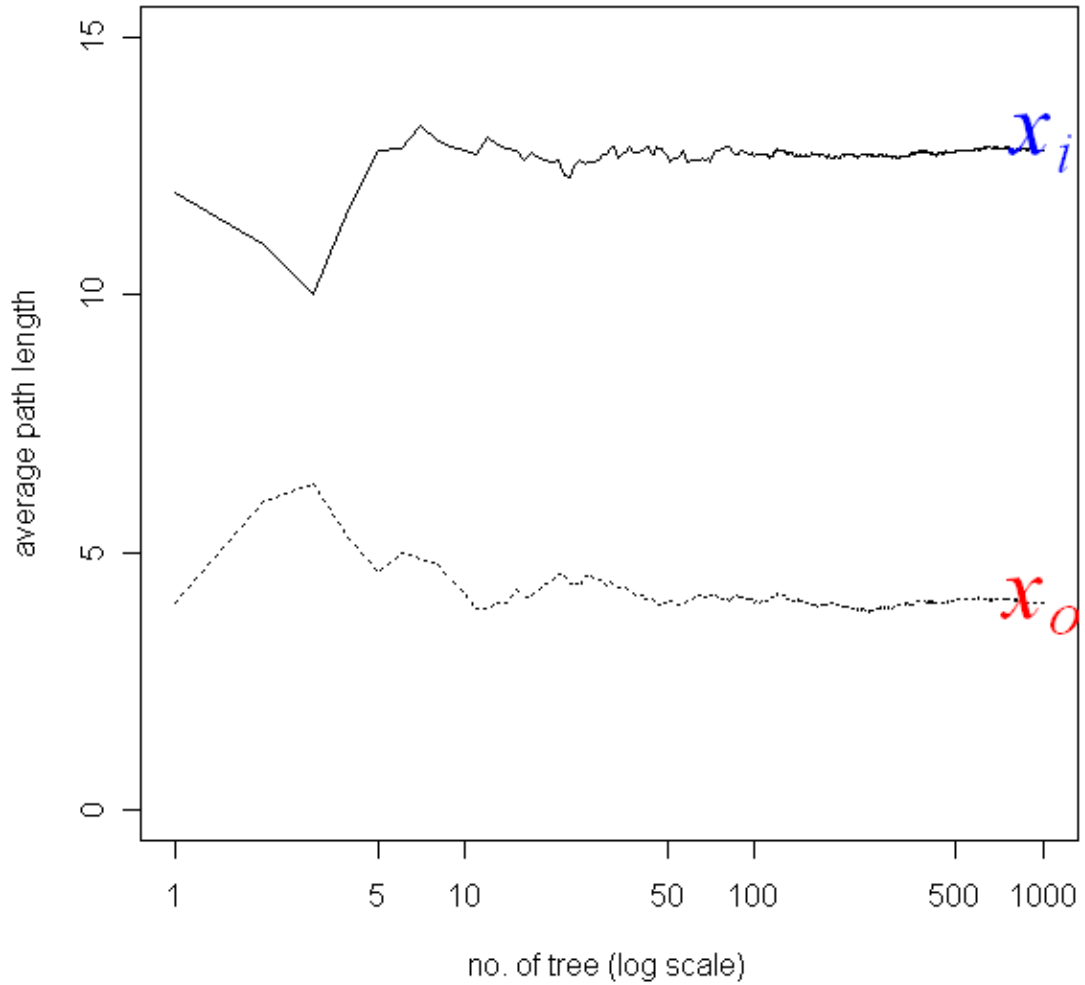


where $\ell_i(\boldsymbol{x})$ is the path length of $\boldsymbol{x}$ traversed in tree i

Isolation example







Source: Liu et al 2008

What iForest is not

- Random Forest or decision trees
- Density estimator

iForest versus iNNE
Conceptual comparison

- **Isolation mechanism**
  - hyper-rectangles vs hyper-balls
  - t sets of rectangles/balls
  - Each set is constructed from a small sample $(\psi \ll n)$
- Measure: path length vs Radius of Ball
- Relative measure
  - Ratio of mass at leaf and its immediate parent
  - Ratio of radii of two neighbouring Balls
- Anomaly Score: average measure over t sets

iNNE: Definitions (Bandaragoda et al, 2018)

Let $D \subset \Re^d$ be a given data set, and let $\|a - b\|$ denote the Euclidean distance between a and b, where $a, b \in \Re^d$.

Let $\mathcal{S} \subset D$ be a subsample of size ψ selected randomly without replacement from a dataset $D \subset \Re^d$; and η_x be the nearest neighbour of x.

Definition 1: A hypersphere $B(c)$ centred at c with radius $\tau(c) = \|c - \eta_c\|$, is defined to be $\{x : \|x - c\| < \tau(c)\}$, where $x \in \Re^d$ and $c,\ \eta_c \in \mathcal{S}$.

Definition 2: Isolation score for $x \in \Re^d$ based on $\mathcal{S}$ is defined as follows:

$$I(x) = \begin{cases} 1 - \frac{\tau(\eta_{cnn(x)})}{\tau(cnn(x))}, & \text{if } x \in \bigcup_{c \in \mathcal{S}} B(c) \\ 1, & \text{otherwise} \end{cases}$$

where $cnn(x) = \arg\min_{c \in \mathcal{S}} \{\tau(c) : x \in B(c)\}$.

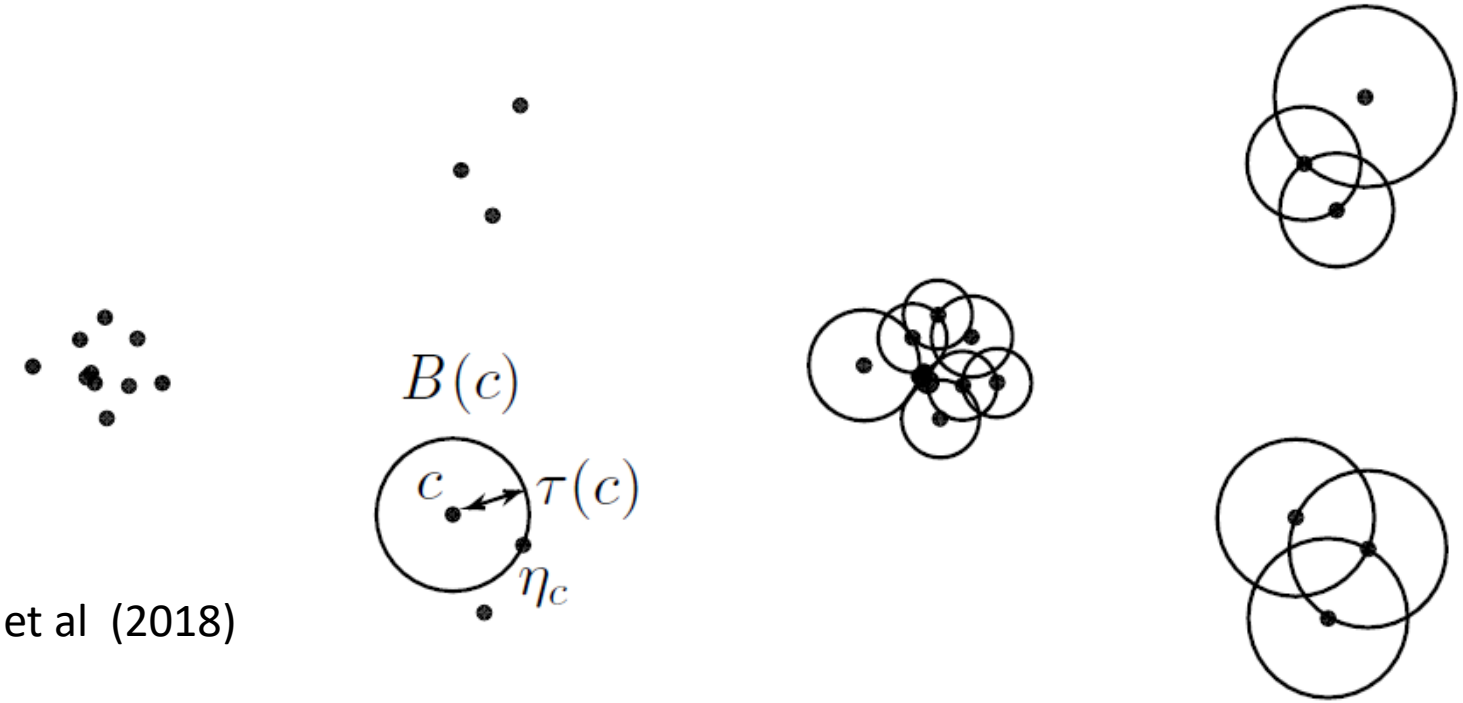
Definition 3: iNNE has a set of t sets of hyperspheres, generated from t subsamples S_i, defined as follows:

$$\left\{ \left\{ B(c) : c \in \mathcal{S}_i \right\} : i = 1, \ldots, t \right\}$$

Example: Constructing a set of Balls

- Sample S is selected randomly from the given dataset
- Ball $B(c)$ is created centering each $c \in S$
- Radius $\tau(c) = ||c - \eta_c||$

  η_c is the nearest neighbour of c where $c, \eta_c \in S$



Source: Bandaragoda et al (2018)

Anomaly score

- Average of isolation scores over an ensemble of size t

$$\bar{I}(x) = \frac{1}{t}\sum_{i=1}^{t} I_i(x)$$

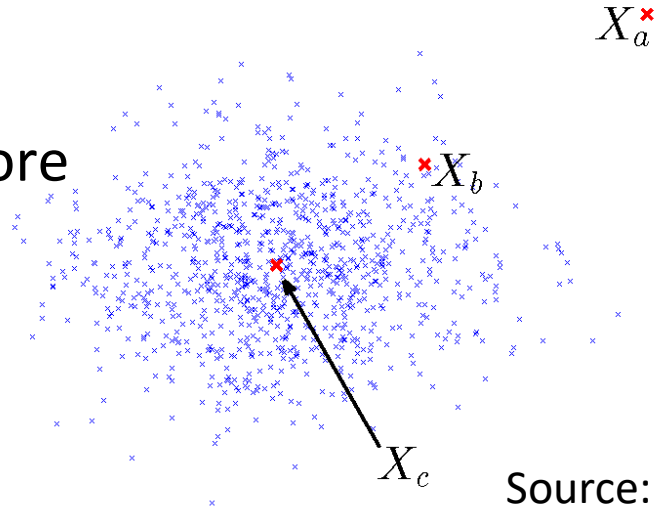
- Instances with high anomaly score are likely to be anomalies

- Accuracy of the anomaly score improve with t
  - t = 20 - 100 is sufficient (more stable than iForest)

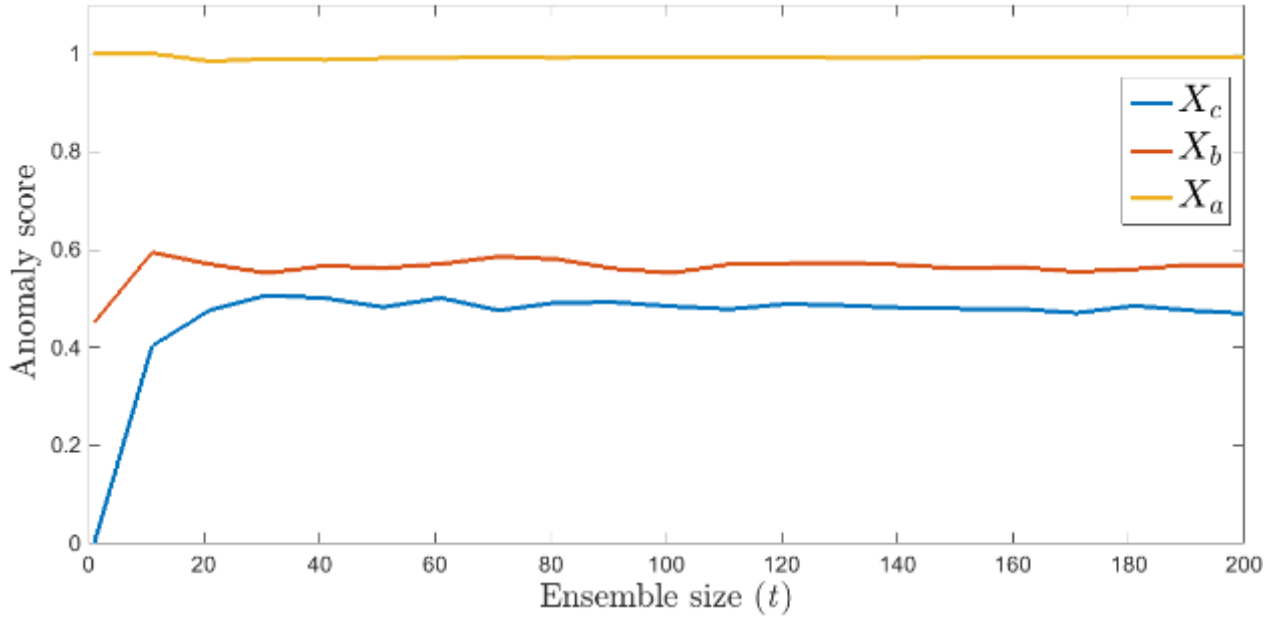
- Sample size, ψ, is a parameter setting
  - Similar to k in k-NN based methods
  - The sample size is usually in the range 2 - 128

Example

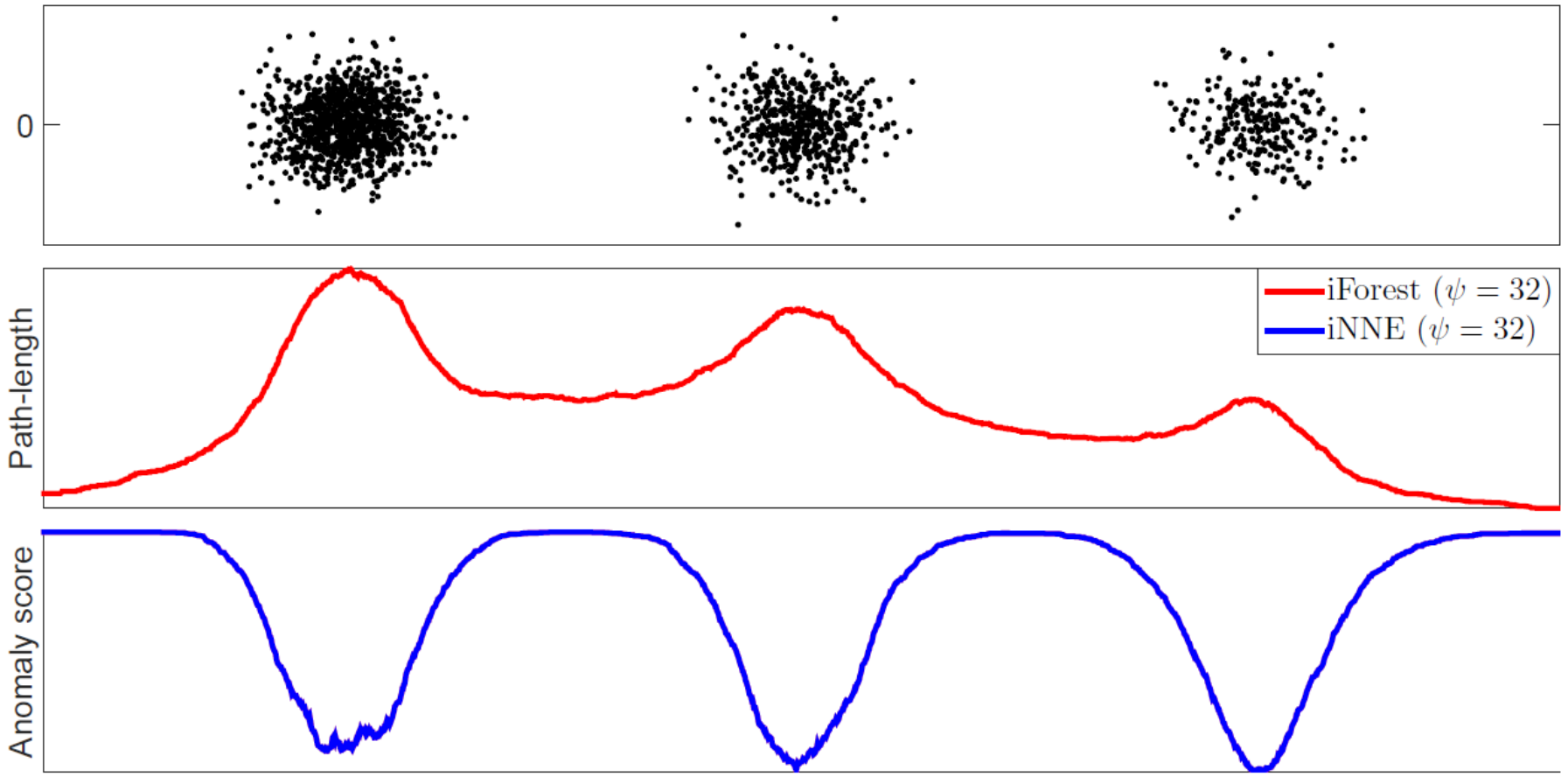
- X_a has the maximum anomaly score
- X_b has a lower anomaly score
- X_c has the lowest anomaly score



Source: Bandaragoda et al (2018)



Global score versus Local score



Source: Bandaragoda et al (2018)

Base NN anomaly detector of aNNE

Let a dataset D be sampled independently and randomly wrt F. Let $\mathcal{D}$ be a subsample set consisting of instances independently and randomly sampled from D; m is a distance measure

Nearest neighbour anomaly detector (1NN) using $\mathcal{D}$ has anomaly score for any $x \in D$ defined as follows:

$$q(x; \mathcal{D}) = \min_{y \in \mathcal{D}} m(x, y). \quad (1)$$

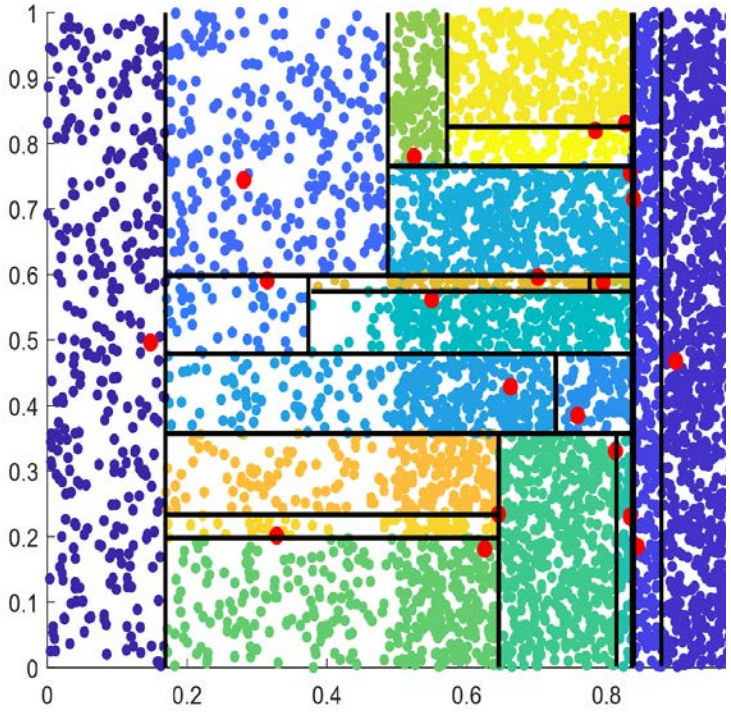
This anomaly detector has a decision rule where x is judged to be an anomaly if $q(x; \mathcal{D})$ is more than or equal to a threshold r.

Space Partitioning: iForest vs aNNE

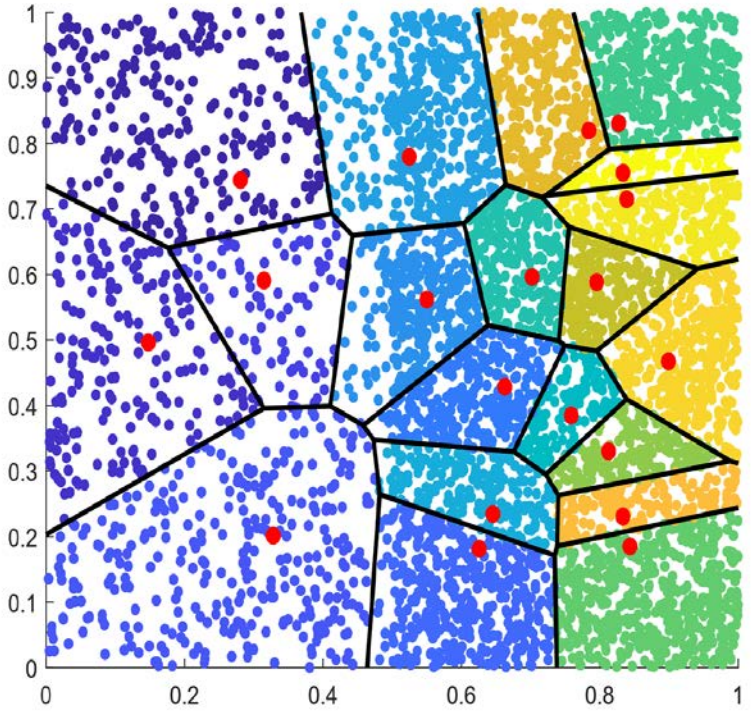
Shortcoming of iForest:

axis-parallel splits

Source: Qin et al [9]



Isolation-Tree Splitting



Nearest-Neighbor Splitting

[9] Qin Xiaoyu, Ting Kai Ming, Zhu Ye and Vincent Lee Cheng Siong (2019) Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. To appear in AAAI 2019.

aNNE is an isolation method (an unknown secret until Qin et al, 2019)

- Isolating partitions are cells in Voronoi diagram (does not need to be created explicitly).
- Anomaly score: Rather than using a proxy (e.g., path length in iForest, or radius of ball in iNNE), aNNE employs distance to the test point's nearest neighbour in the subsample.

k-Nearest neighbour-based methods

- kNN
  - Clustered anomalies: ABOD & LOF

Reasons not to use these methods (in comparison with aNNE and iNNE):

- Runtime is slower; and space complexity is high
- Parameter k is sensitive to data size

Potential reasons to use them:

- High number of clusters (kNN)
- High dimensional datasets (ABOD) --- have not compared

Recommendation and Caveat

All three studies lead to the recommendation:

Use Isolation-based methods (tree-based or NN-based) because

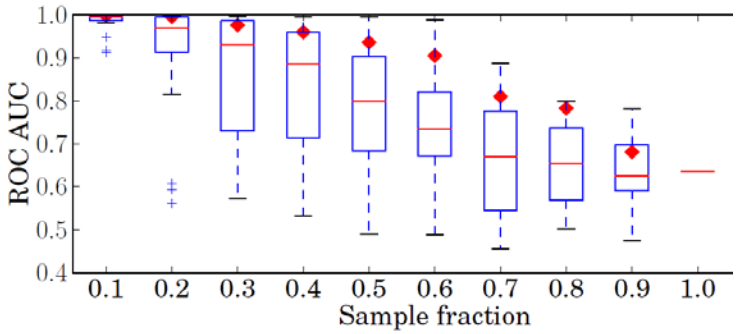
- Parameter: Easy to tune; not too sensitive to parameter setting
- Fast runtime: can scale up to large datasets and high dimensional datasets
- Low space complexity: linear to sample size and ensemble size.
- Known behaviours under different data properties
- Can deal with different types of anomalies

Caveat: All the limitations of these studies.

CONTENTS

Gravity Defiant Behavior of Learning Curve

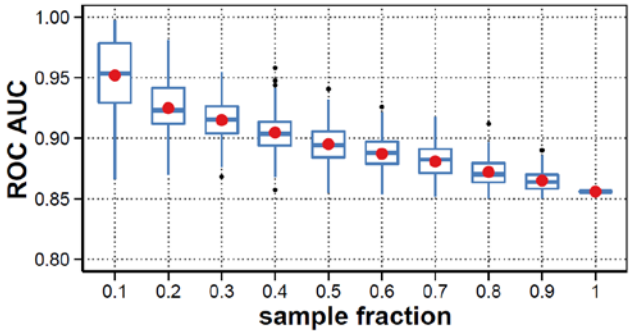
Recently, higher accuracy of k-NN based anomaly detection using less samples are reported in many papers.



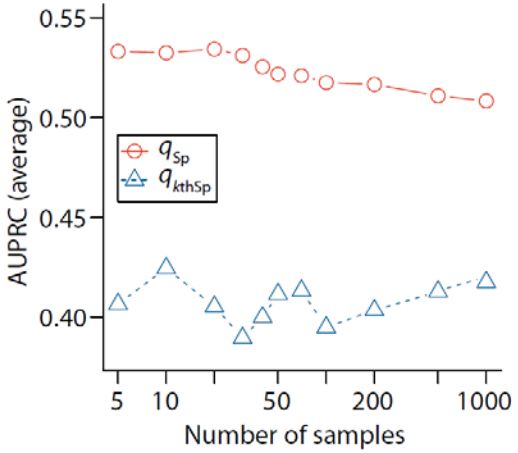
(b) LOF, $k = 2$

A. Zimek, A. Zimek, M. Gaudet, R.J.G.B. Campello, J. Sander, Proc. SIG-KDD2013, pp.428-436, 2013

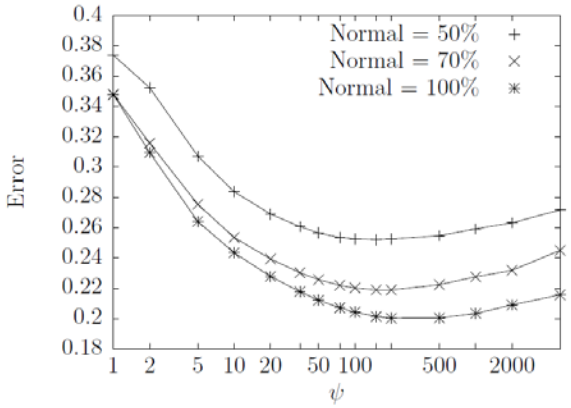
(b) Cardio: Subsampling with $k = 5$



C.C. Aggarwal and S. Sathe, ACM SIGKDD Explorations Newsletter, Vol.17, No.1, pp.24-47, 2015



M. Sugiyama and K.M. Borgwardt, Proc. NIPS'13, pp,467-475, 2013



(b) kNN

K.M. Ting, T. Washio, J.R. Wells, S. Aryal, Machine Learning, 2016

Outline

- “Higher accuracy with less samples” is an unusual statement in terms of the convention of statistics and machine learning.
- Currently, many arguments on this statement are ongoing.

1. A. Zimek, M. Gaudet, R.J.G.B. Campello, J. Sander, Subsampling for efficient and effective unsupervised outlier detection ensembles, Proc. SIG-KDD2013, pp.428-436, 2013

2. C.C. Aggarwal and S. Sathe, Theoretical Foundations and Algorithms for Outlier Ensembles, ACM SIGKDD Explorations Newsletter, Vol.17, No.1, pp.24-47, 2015

3. M. Sugiyama and K.M. Borgwardt, Rapid Distance-Based Outlier Detection via Sampling, Proc. NIPS'13, pp.467-475, 2013

4. K.M. Ting, T. Washio, J.R. Wells, S. Aryal, Defying the gravity of learning curve: a characteristic of nearest neighbour anomaly detectors, Machine Learning, Vol.106, Issue 1, pp.55-91, 2017

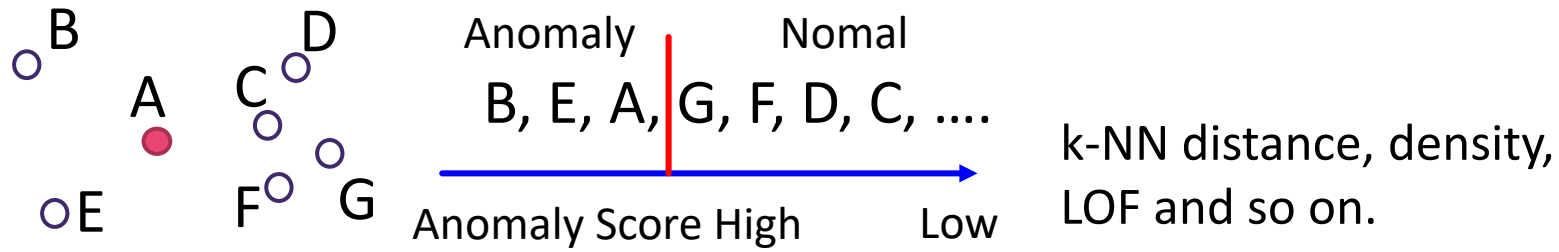
A. Zimek, M. Gaudet, R.J.G.B. Campello,
J. Sander, Proc. SIG-KDD2013, pp.428-436, 2013

The accuracy of the anomaly detection depends on the estimation resolution of the data density and thus the k-NN distance.

Let the k-NN distance of a point x be $q_k(x)$.

Density $\hat{p}_k(x) \propto \frac{k}{q_k^d(x)}$ ⇒ Error of density $e_k(x) = p(x) - \hat{p}_k(x)$

For example, "Ranking based Anomaly Decision"

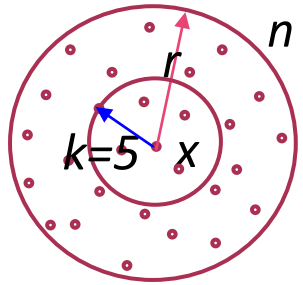


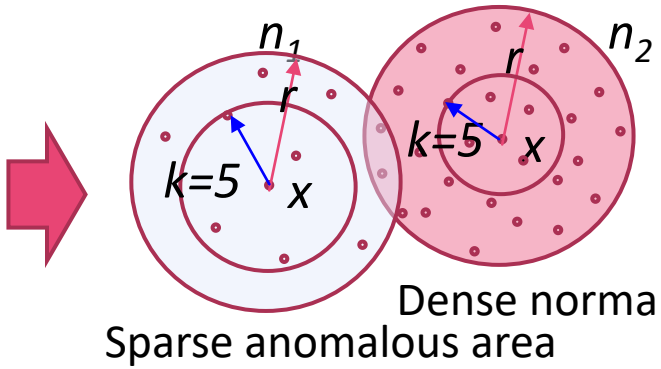
The estimation error (the space resolution) of k-NN distance dominates erroneous flips of the ranking order.

The space resolution of the k-NN distance increases when less samples are used. Accordingly, we observe higher accuracy of the anomaly detection with less samples.

The expected k nearest neighbour (k-NN) distance in a *d*-dimensional sphere containing *n* uniform points [M. M. Breunig et al., SIGMOD2001].

$$E[q_k \mid n] = r\left(\frac{k}{n}\right)^{1/d}$$





$$n_1 < n_2 \Rightarrow \Delta_{k,1} = E[q_k \mid n_1] - E[q_k \mid n_2] = r\left(\frac{k}{n_1}\right)^{1/d} - r\left(\frac{k}{n_2}\right)^{1/d}$$

By applying subsampling to randomly take a fraction f (<1) of the data sets, $\Delta_{k,f}$ is expanded from $\Delta_{k,1}$ by the factor $(1/f)^{1/d}$.

$$\Delta_{k,f} = r\left(\frac{k}{n_1 f}\right)^{1/d} - r\left(\frac{k}{n_2 f}\right)^{1/d}$$

Higher space resolution

An issue in the claim (Zimek et al, 2013)

The space resolution of the k-NN distance does not increase when less samples are used.

D. Evans et al., Asymptotic moments of near-neighbour distance distributions, Proc. Mathematical, Physical and Engineering Sciences, Vol.458, No.2028, pp.2839-2849, 2002

$$E[q_{k,n}] \cong \frac{r}{B_d^{1/d} n^{1/d} f^{1/d}} \frac{\Gamma(k+1/d)}{\Gamma(k)}$$ (B_d is a volume of a d-dimensional unit ball.)

$$Std[q_{k,n}] \cong \frac{r}{B_d^{1/d} n^{1/d} f^{1/d}} \sqrt{\frac{\Gamma(k+2/d)}{\Gamma(k)} - \left(\frac{\Gamma(k+1/d)}{\Gamma(k)}\right)^2}$$

By applying subsampling, the uncertainty of the distance difference also increases in the same manner. So, the stretch of $\Delta_{k,f}$ does not seem to increase the space resolution.

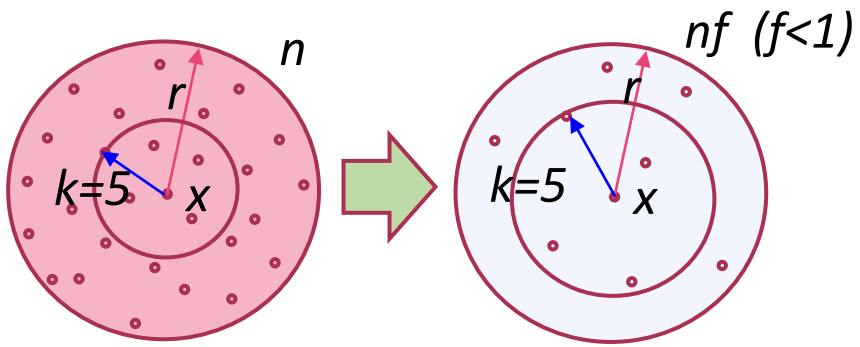
C.C. Aggarwal and S. Sathe, ACM SIGKDD Explorations Newsletter, Vol.17, No.1, pp.24–47, 2015

Given an oracle function of an anomaly score of a point x : $f(x)$, let the estimate of the function under a given data set : $\hat{f}(x)$.

$$E[MSE[\hat{f}]] = \int \{\hat{f}(x) - f(x)\}^2 p(x)dx$$

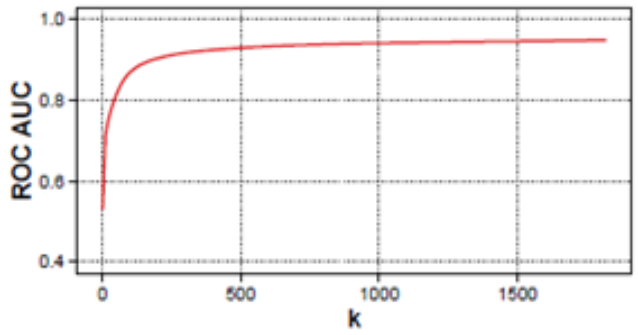
$$= \int \{f(x) - E[\hat{f}(x)]\}^2 p(x)dx + \int \{\hat{f}(x) - E[\hat{f}(x)]\}^2 p(x)dx$$

$$= \underbrace{E[\{f(x) - E[\hat{f}(x)]\}^2]}_{E[Bias[\hat{f}]]} + \underbrace{E[\{\hat{f}(x) - E[\hat{f}(x)]\}^2]}_{E[Var[\hat{f}]]}$$

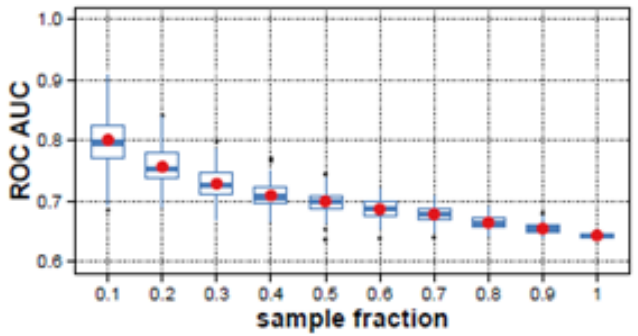


The subsampling with a fixed *k* increase the range of the k-NN density estimator: $\hat{f}(x)$. This affects both the bias and the variance.

In data sets *where the bias decreases with parameters of a* score-based anomaly detector, less subsamples with fixed *parameter values* will generally decrease the error of the detector.

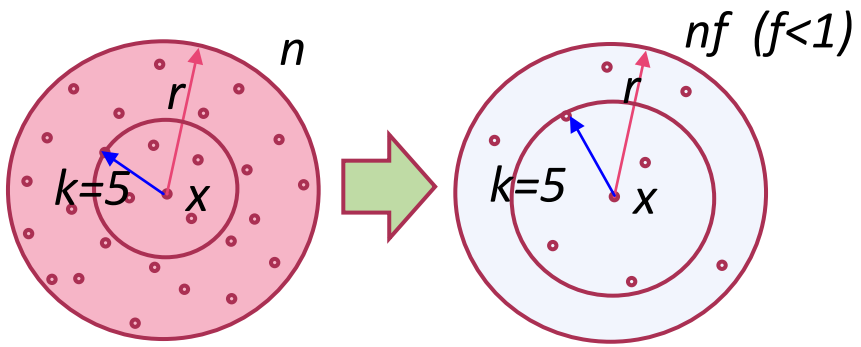


(a) Cardio: AUC vs. k



(b) Cardio: Subsampling with $k = 5$

Gravity defying behavior



Density $\hat{f}(x) = \hat{p}_k(x) \propto \frac{k}{q_k^d(x)}$

Their claim:
not limited to k-NN detector

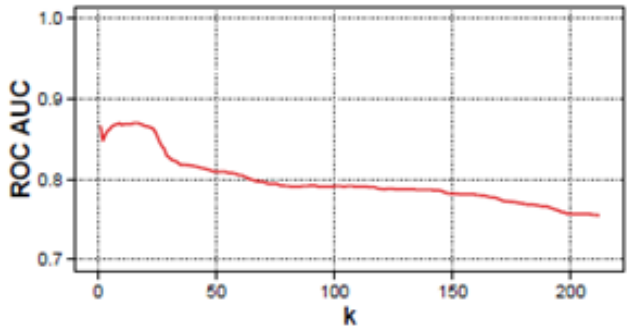
$$E[MSE[f]] =$$

$$E[Bias[\hat{f}]] + E[Var[\hat{f}]]$$

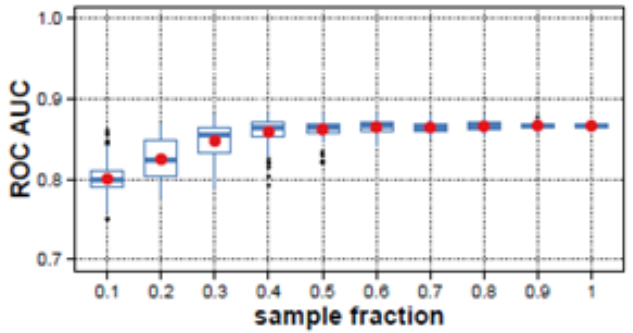
Decrease ↓ Increase ↑

may decrease in total ↓

In data sets *where the bias increases with parameters of a* score-based anomaly detector, less subsampling with fixed *parameter values* will generally significantly increase the error of the detector.

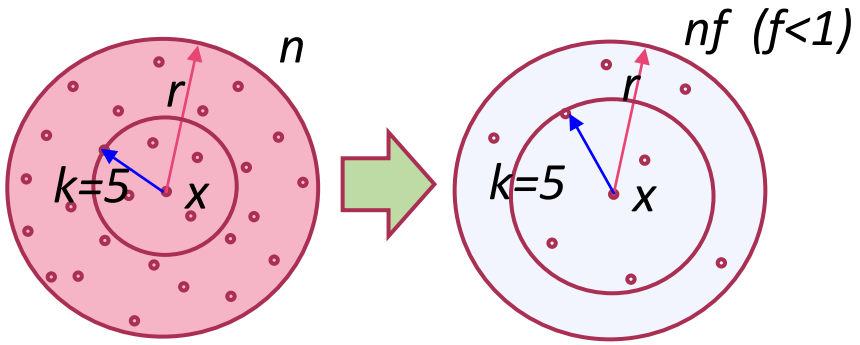


(a) Glass: AUC vs. k



(b) Glass: Subsampling with $k = 6$

Gravity compliant behavior



Density $\hat{f}(x) = \hat{p}_k(x) \propto \dfrac{k}{q_k^d(x)}$

Their claim:
not limited to k-NN detector

$$E[MSE[f]] =$$

$$E[Bias[\hat{f}]] + E[Var[\hat{f}]]$$

Increase Increase

significantly increase in total

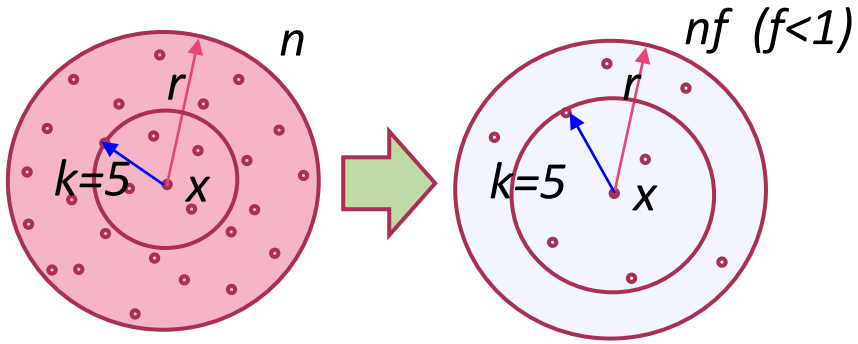
An issue in the claim (Aggarwal & Sathe, 2015)

D. Evans et al., Asymptotic moments of near-neighbour distance distributions, Proc. Mathematical, Physical and Engineering Sciences, Vol.458, No.2028, pp.2839-2849, 2002

$$E[q_{k,n}] \cong \frac{r}{B_d^{1/d} n^{1/d} f^{1/d}} \frac{\Gamma(k+1/d)}{\Gamma(k)}$$

$$Std[q_{k,n}] \cong \frac{r}{B_d^{1/d} n^{1/d} f^{1/d}} \sqrt{\frac{\Gamma(k+2/d)}{\Gamma(k)} - \left(\frac{\Gamma(k+1/d)}{\Gamma(k)}\right)^2}$$

The effect of the subsample fraction f on both the bias and the variance is independent of k.

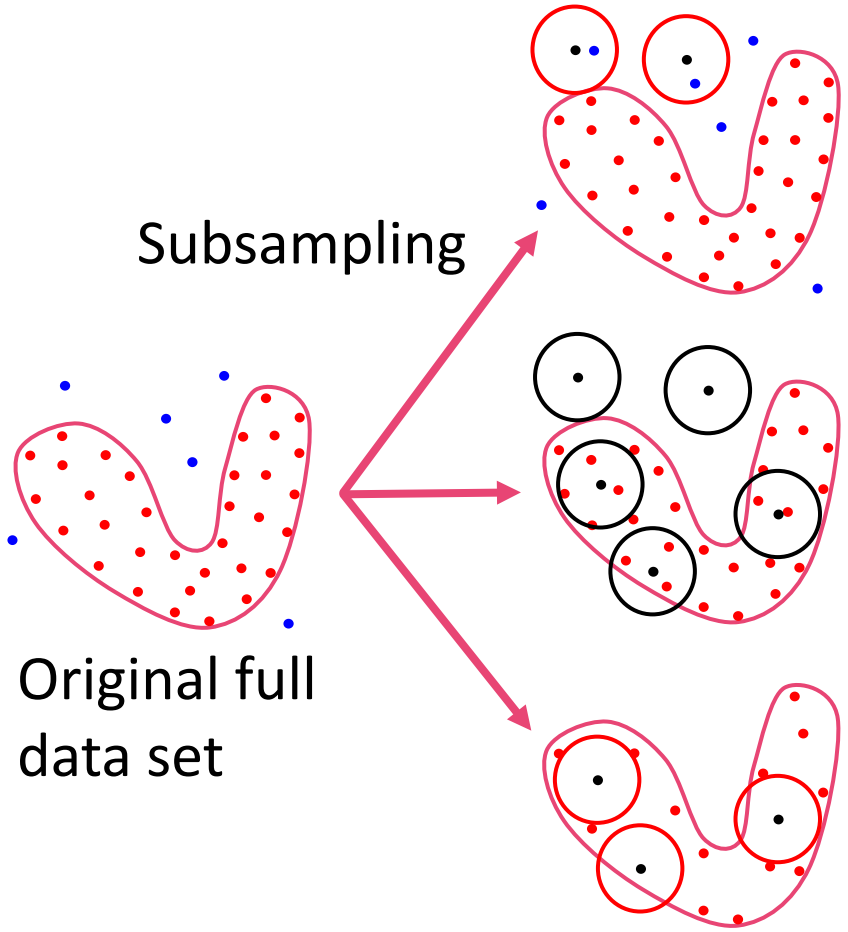


The less faction f should increase the bias and variance under a fixed k, and should significantly decrease the AUC.

However, we observe increasing or non-significantly decreasing AUC in practice. This suggests some other factors leading to AUC enhancement.

M. Sugiyama and K.M. Borgwardt, NIPS'13, 467–475, 2013

For the accurate anomaly detection, the subsamples must scattered over all normal area without covering any anomalous area. This nature originates the gravity defiant behavior of the learning curve.



When the subsampling fraction is large, some anomalies are not detected.

When the subsampling fraction is appropriate, anomalies are correctly detected.

When the subsampling fraction is small, some normal points are falsely detected as anomalies.

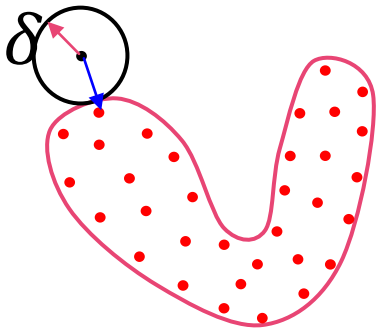
A nearest neighbor (1-NN) distance based anomaly detector has an optimum subsample size in theory.

They focused on Knorr and Ng's $\mathrm{DB}(\alpha, \delta)$-outliers model [Knorr and Ng,1998,2000]

Given a data set D, a set of anomaly points in D is defined as

$$D(\alpha;\delta) = \left\{ \forall x \in D \,\middle|\, \frac{|\{x' \in D \,|\, d(x,x') > \delta\}|}{|D|} > \alpha \right\}$$

α is a fraction of the normal points in the data set D.



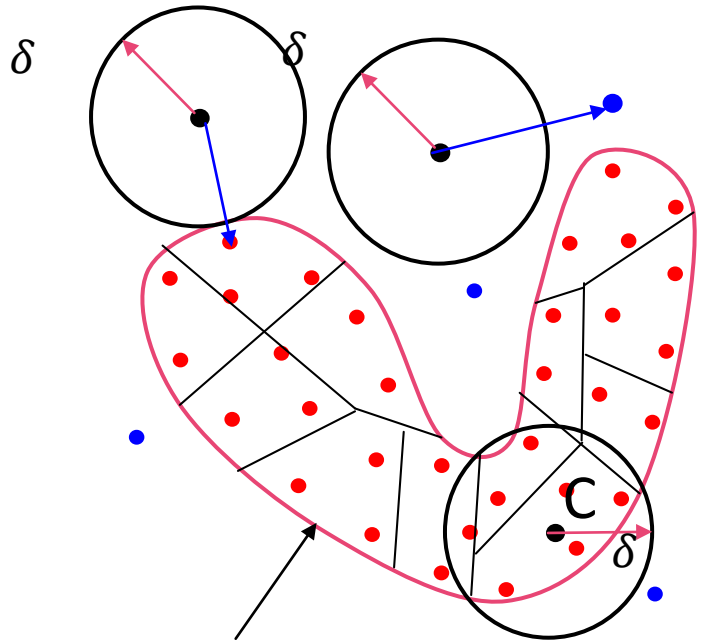
$|D| = 40$

$\alpha = 39/40$

The black point is anomaly.

They focused 1-NN distance $\mathrm{q}_1(\mathrm{x})$ based anomaly detector.

$$D(\alpha;\delta) = \{x \in D \,|\, q_1(x) > \delta\} \quad (\alpha = (|D| - 1)/|D|)$$



$\overline{D}(\alpha;\delta) = D - D(\alpha;\delta)$

The covering-radius of each cluster C is less than δ. Then, we randomly draw a subsample set $S \subset D$ of size $s = |S|$. Compute the 1-NN dist. $q_1(x)$ in S.

Theorem 1 *For an anomaly* $x \in D(\alpha;\delta)$ *and a cluster* $C \in P_\delta$, *we have*
$\Pr(\forall x' \in C, q_1(x) > q_1(x')) > \alpha^s(1-\beta_C{}^s)$
where $\beta_C = (|D| - |C|)/|D|$.

If all points in S are normal, $q_1(x) > \delta$ for anomaly x.

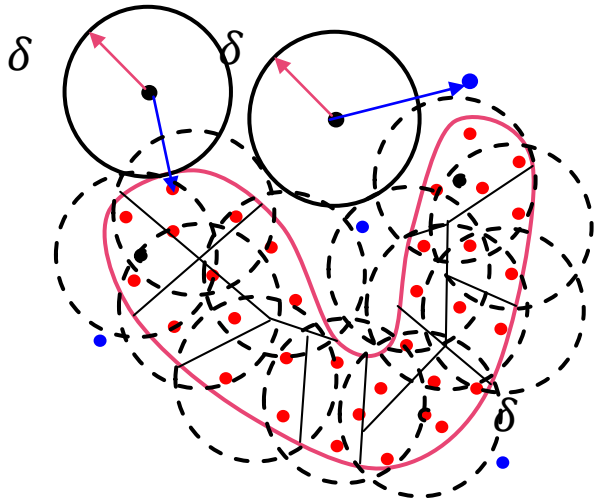
$\Pr = \alpha^s$ s.t. $\alpha = |\overline{D}(\alpha;\beta)|/|D|$ ⇒ $\Pr(q_1(x) > \delta) > \alpha^s$

If a point in S is from the cluster C, $q_1(x') < \delta$ for all normal $x' \in C$.

$\Pr = 1 - \beta_C{}^s$ s.t. $\beta_C = (|D| - |C|)/|D|$ ⇒ $\Pr(\forall x' \in C, q_1(x') < \delta) > 1 - \beta_C{}^s$

$\Pr(\forall x' \in C, q_1(x) > \delta > q_1(x')) > \alpha^s(1-\beta_C{}^s)$

Optimum subsample size



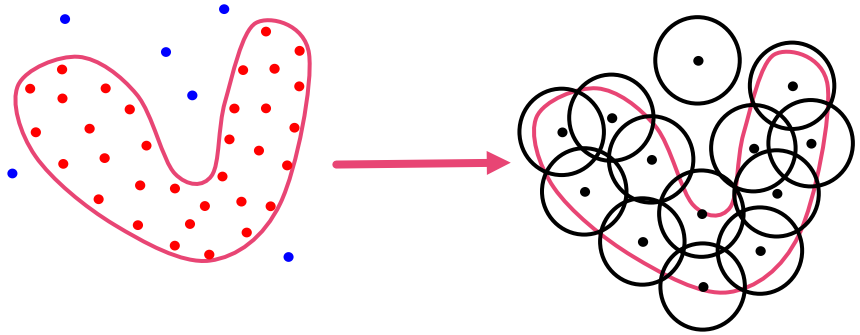
$$\Pr\left(\forall x' \in C, q_1(x) > q_1(x')\right) > \alpha^s(1-\beta_C{}^s) > \alpha^s(1-\beta^s) \rightarrow max \text{ for all C}$$

s maximizing the lower bound

$$s = \log_\beta\left(\frac{\log\alpha}{\log\alpha+\log\beta}\right)$$

This suboptimal sample size s is small for large α and small β.

For $(\alpha;\beta)$ = (0.99; 0.5), s=6.
For $(\alpha;\beta)$ = (0.999; 0.8), s=24.



When the subsampling fraction is appropriate, anomalies are correctly detected.

Issues in the claims (Sugiyama & Borgwardt, 2013)

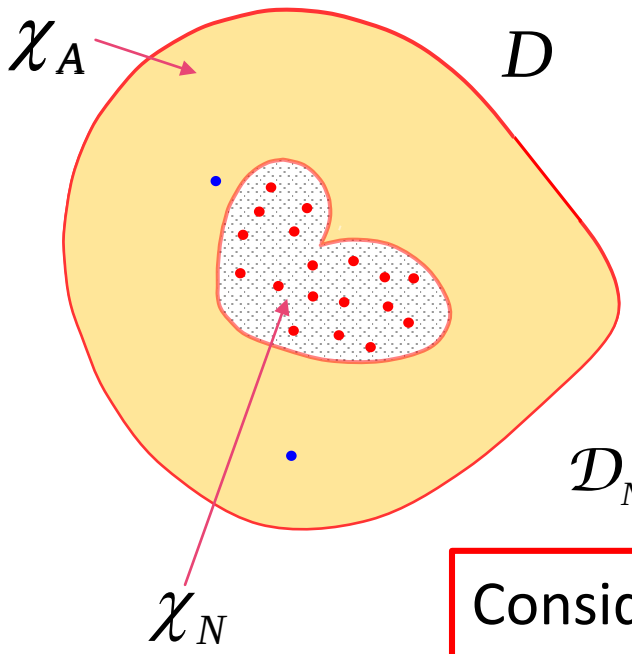
They provided a lower bound of the success probability of the 1-NN distance and subsample based anomaly detector.

1. The upper bound is not provided.
2. The bounds in terms of more rigorous accuracy indices like AUC is not provided.
3. The analysis is limited to the $\mathrm{DB}(\alpha, \delta)$-outliers model.

K.M. Ting, T. Washio, J.R. Wells, S. Aryal, Machine Learning, Vol 106, Issue 1, 55-91, 2017

Both upper and lower bounds of AUC of the 1-NN distance based anomaly detector is provided in simple closed forms.

A metric space (M, m)



A data set D drawn from *p(x)*,
D_N be the set of all points in χ_N ,
D_A be the set of all points in χ_A.

$$|D_A| << |D_N|$$

$\mathcal{D}$ is a data subset subsampled from D.

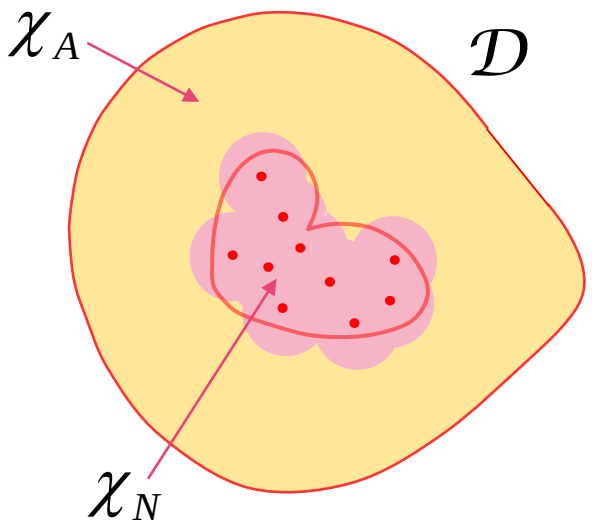
$\mathcal{D}_N$ and $\mathcal{D}_A$ are defined similarly to D_N and D_A .

Consider 1-NN distance based anomaly detector.

$$q(x;\mathcal{D}) = \min_{y \in \mathcal{D}} m(x,y) \geq r \Rightarrow x \text{ is anomaly.}$$

Representation of AUC in Anomaly Detection

AUC($\mathcal{D}$): AUC of the 1-NN anomaly detector using a subsampled data set $\mathcal{D}$. Since $|D_A| << |D_N|$, $D \cong D_N$, thus AUC($\mathcal{D}$) ≅ AUC($\mathcal{D}_N$).



PDF of true positive for distance s :

$$f(s;\mathcal{D}_N) = \int_{\{x \in \chi_A | q(x;\mathcal{D}_N)=s\}} p(x)dx$$

PDF of false positive for distance s :

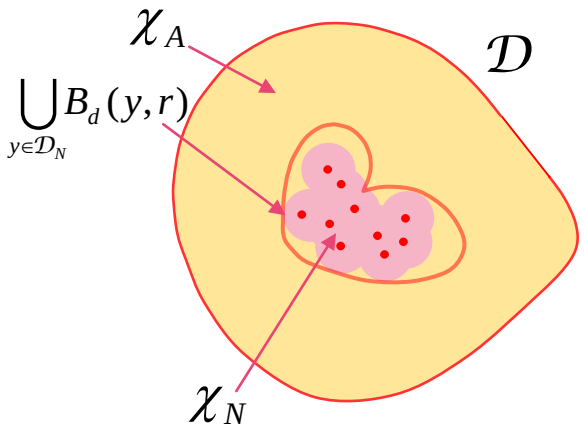
$$g(s;\mathcal{D}_N) = \int_{\{x \in \chi_N | q(x;\mathcal{D}_N)=s\}} p(x)dx$$

AUC($\mathcal{D}$) ≅ AUC($\mathcal{D}_N$) is represented as follows[Hand and Till 2001].

$$AUC(\mathcal{D}_N) = \int_0^1 G(r;\mathcal{D}_N)dF(r;\mathcal{D}_N) = \int_0^\infty G(r;\mathcal{D}_N)f(r;\mathcal{D}_N)dr$$

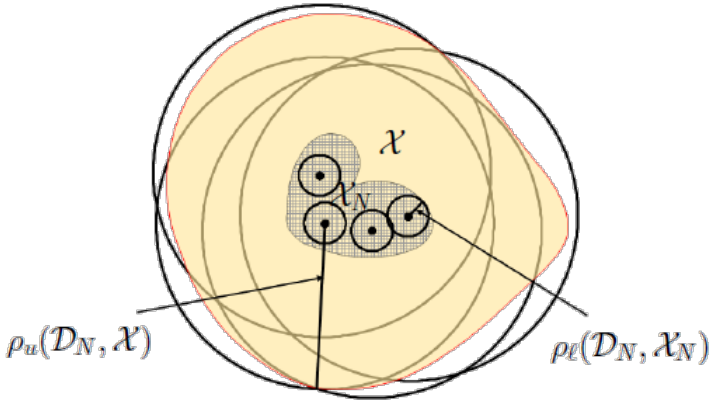
Inradius of χ_N and Covering radius of χ

Let $B_d(y,r)$ be a *d* dimensional ball centered at *y* with radius *r*.
The set of all points satisfying $q(x; \mathcal{D}_N) < r$ is $\bigcup_{y \in \mathcal{D}_N} B_d(y,r)$.



Inradius of $\chi_N : \rho_\ell(\mathcal{D}_N, \chi_N)$

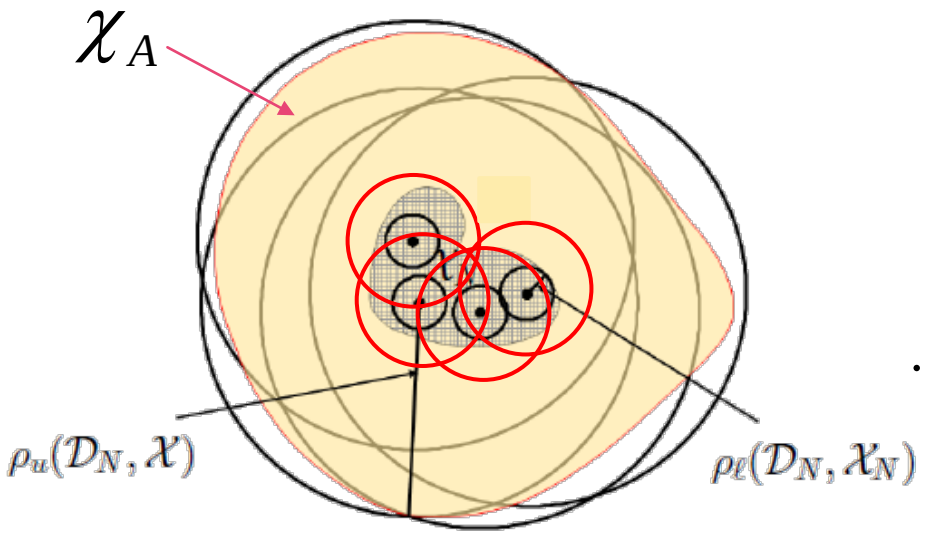
$$\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) = \sup \arg_r \left[\bigcup_{y \in \mathcal{D}_N} B_d(y,r) \subseteq \mathcal{X}_N \right]$$
$$= \sup \arg_r \left[\{x \in \mathcal{M} | q(x; \mathcal{D}_N) \leq r\} \subseteq \mathcal{X}_N \right]$$



Covering radius of $\chi : \rho_u(\mathcal{D}_N, \chi)$

$$\rho_u(\mathcal{D}_N, \mathcal{X}) = \inf \arg_r \left[\bigcup_{y \in \mathcal{D}_N} B_d(y,r) \supseteq \mathcal{X} \right]$$
$$= \inf \arg_r \left[\{x \in \mathcal{M} | q(x; \mathcal{D}_N) \leq r\} \supseteq \mathcal{X} \right]$$

The relation between AUC and these two radii



$q(x;\mathcal{D}) = r$ is the surface of the ball union $\bigcup_{y \in \mathcal{D}_N} B_d(y, r)$.

When $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) < r < \rho_u(\mathcal{D}_N, \mathcal{X})$, the surface intersects with $\mathcal{X}_A$.

$$\therefore f(r;\mathcal{D}_N) = \int_{\{x \in \mathcal{X}_A | q(x;\mathcal{D}_N) = r\}} p(x)dx \neq 0$$

Otherwise, the surface does not intersect with $\mathcal{X}_A$.

$$\therefore f(r;\mathcal{D}_N) = \int_{\{x \in \mathcal{X}_A | q(x;\mathcal{D}_N) = r\}} p(x)dx = 0$$

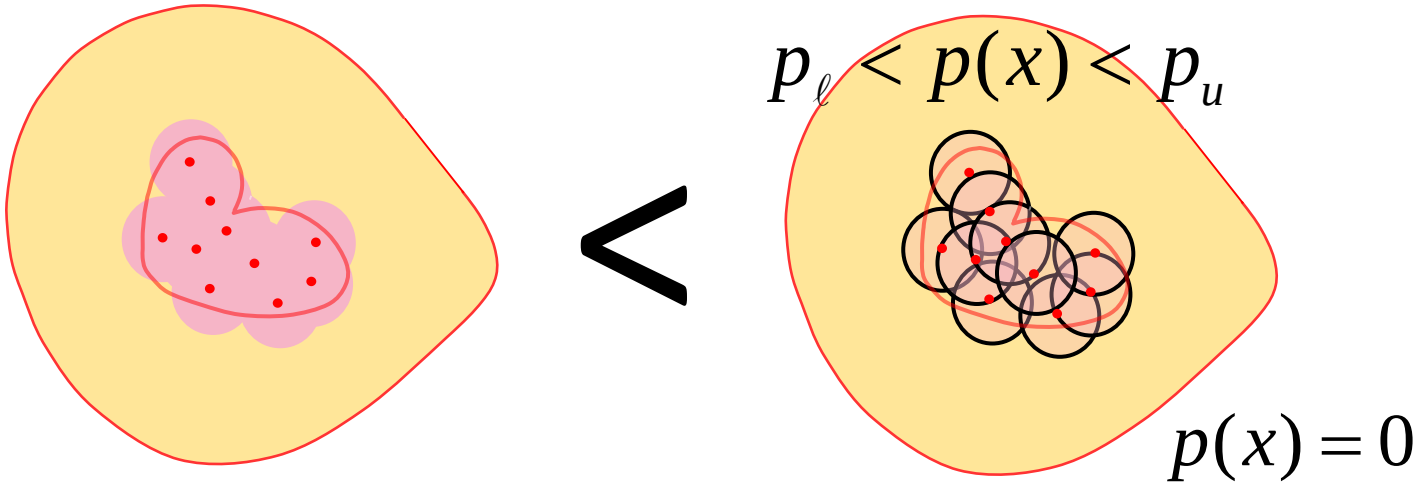
$$\therefore AUC(\mathcal{D}_N) = \int_0^\infty G(r;\mathcal{D}_N) f(r;\mathcal{D}_N) dr = \int_{\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)}^{\rho_u(\mathcal{D}_N, \mathcal{X})} G(r;\mathcal{D}_N) f(r;\mathcal{D}_N) dr$$

Upper bounds of $f(r;\mathcal{D}_N)$

Let $|\mathcal{D}| = \psi$ and $|\mathcal{D}_N| = h$, respectively.

Lemma 1 *$f(r;\mathcal{D}_N)$ is upper bounded for $r \in \mathcal{R}^+$ as follows.*

$$f(r;\mathcal{D}_N) < p_u S_d h r^{d-1} \quad \text{if } \rho_\ell(\mathcal{D}_N, \mathcal{X}_N) \le r \le \rho_u(\mathcal{D}_N, \mathcal{X}),$$
$$f(r;\mathcal{D}_N) = 0 \quad \text{otherwise.}$$



$$f(r;\mathcal{D}_N) = \int_{\{x \in \chi_A | q(x;\mathcal{D}_N) = r\}} p(x)dx$$

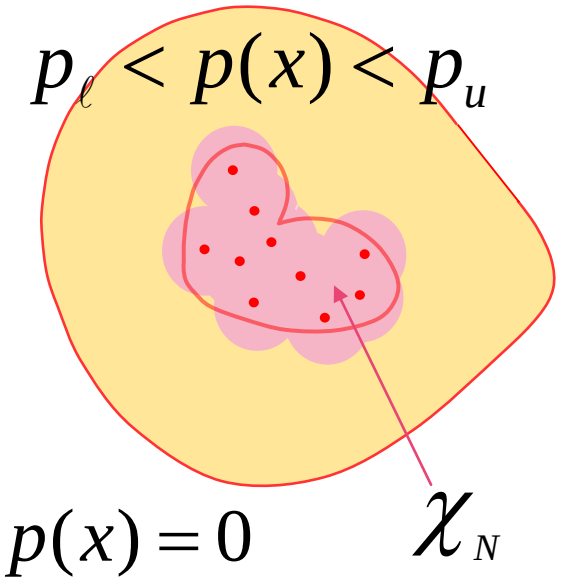
$$\sum_{y \in S} \int_{S_d(y,r)} p(x)dx < \sum_{y \in S} \int_{S_d(y,r)} p_u dx = p_u S_d h r^{d-1}$$

Upper bounds of $G(r;\mathcal{D}_N)$

Lemma 2 *$G(r;\mathcal{D}_N)$ is upper bounded for $r \in \mathcal{R}^+$ as*

$$G(r;\mathcal{D}_N) < p_u|\mathcal{X}_N|,$$

where $|\mathcal{X}_N|$ is the volume of $\mathcal{X}_N$.



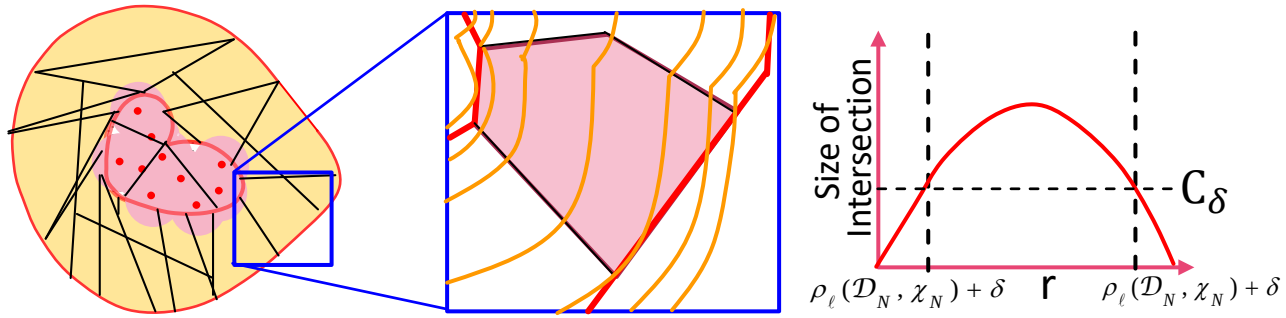
$$g(s;\mathcal{D}_N) = \int_{\{x \in \chi_N | q(x;\mathcal{D}_N)=s\}} p(x)dx$$

$$G(r;\mathcal{D}_N) = \int_0^r g(s;\mathcal{D}_N)ds < p_u|\chi_N|$$

Lower bounds of $f(r;\mathcal{D}_N)$ and $G(r;\mathcal{D}_N)$

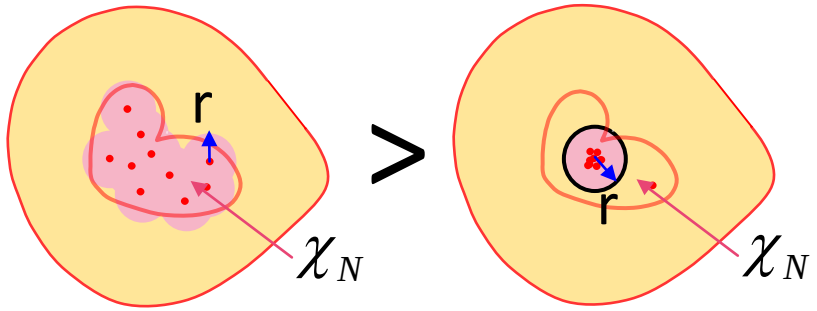
Lemma 3 *There exists some constant $\delta_0 \in \mathcal{R}^+$ such that a constant $C_\delta \in \mathcal{R}^+$ lower bounds $f(r;\mathcal{D}_N)$ for every $\delta \in (0, \delta_0)$ and $r \in \mathcal{R}^+$ as follows.*

$$f(r;\mathcal{D}_N) > p_\ell C_\delta \quad \text{if } \rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta \le r \le \rho_u(\mathcal{D}_N, \mathcal{X}) - \delta,$$
$$f(r;\mathcal{D}_N) \ge 0 \quad \text{otherwise.}$$



Lemma 4 *There exists some constant $C_\ell \in \mathcal{R}^+$ such that $G(r;\mathcal{D}_N)$ is lower bounded for $r \in \mathcal{R}^+$ as follows.*

$$G(r;\mathcal{D}_N) > p_\ell C_\ell r^d \quad \text{if } 0 < r \le \rho_u(\mathcal{D}_N, \mathcal{X}),$$
$$G(r;\mathcal{D}_N) > p_\ell |\mathcal{X}_N| \quad \text{otherwise.}$$



Bounds of AUC expected for all possible $\mathcal{D}_N$ drawn from D

$\mathcal{D} \cong \mathcal{D}_N \Leftrightarrow h \cong \psi$. Given $\alpha = |D_N|/|D|$, $p(\mathcal{D}=\mathcal{D}_N)=\alpha^{\psi}$.

+

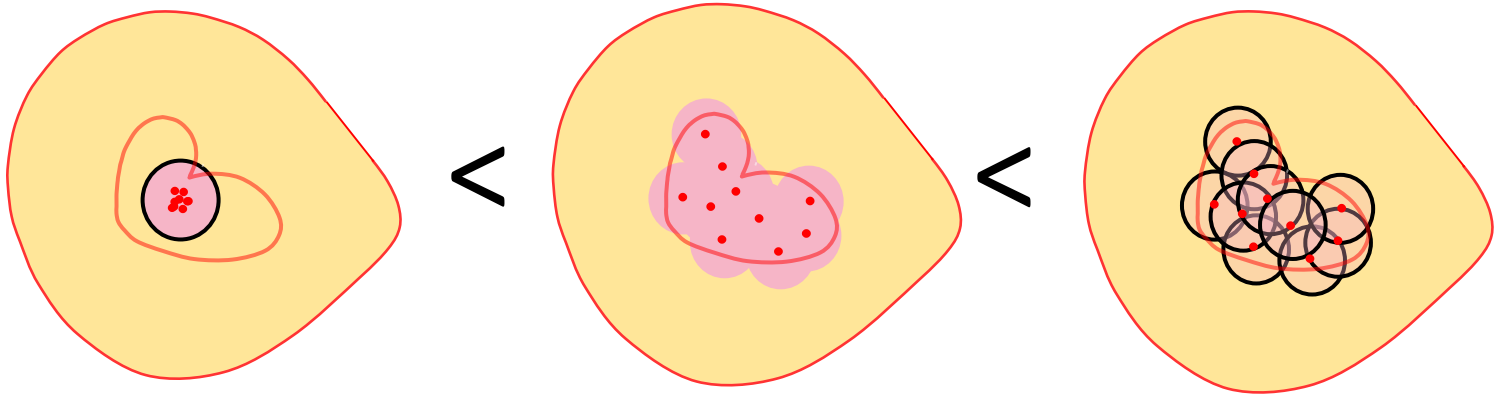
The bounds of $f(r;\mathcal{D}_N)$ and $G(r;\mathcal{D}_N)$ in Lemma 1-4, and

$$AUC(\mathcal{D}_N) = \int_{\rho_\ell(\mathcal{D}_N,\chi_N)}^{\rho_u(\mathcal{D}_N,\chi)} G(r;\mathcal{D}_N) f(r;\mathcal{D}_N) dr$$

Theorem 1 The expected AUC bounds in closed forms

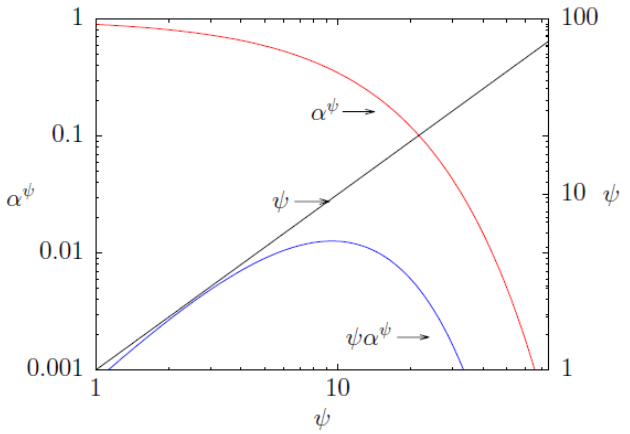
$$\langle AUC \rangle < \frac{p_u^2 |\mathcal{X}_N| S_d}{d} \psi\alpha^{\psi} \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_{\psi} - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_{\psi} \right) + O((1-\alpha)\psi^2), \text{ and}$$

$$\langle AUC \rangle > \frac{p_\ell^2 C_\ell C_\delta}{d+1} \alpha^{\psi} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_{\psi} - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_{\psi} \right\} + O((1-\alpha)\psi).$$

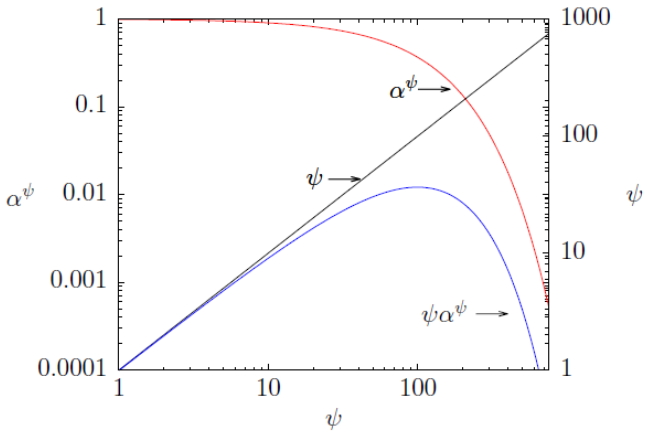


Gravity Defiant Behavior

Term $\psi\alpha^{\psi}$



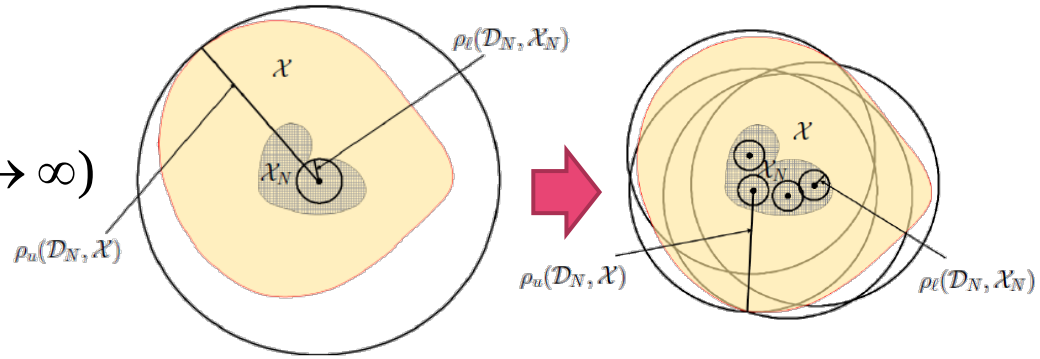
(a) $\alpha = 0.9$: $\psi\alpha^{\psi}$ has $\psi_{opt} = 10$.



(b) $\alpha = 0.99$: $\psi\alpha^{\psi}$ has $\psi_{opt} = 100$.

Term $\left\langle \rho_u(\chi)^d \right\rangle_{\psi} - \left\langle \rho_\ell(\chi_N)^d \right\rangle_{\psi}$

$$\approx \left\langle \rho_u(\chi)^d \right\rangle_{\psi} \to \text{small}\,(\psi \to \infty)$$



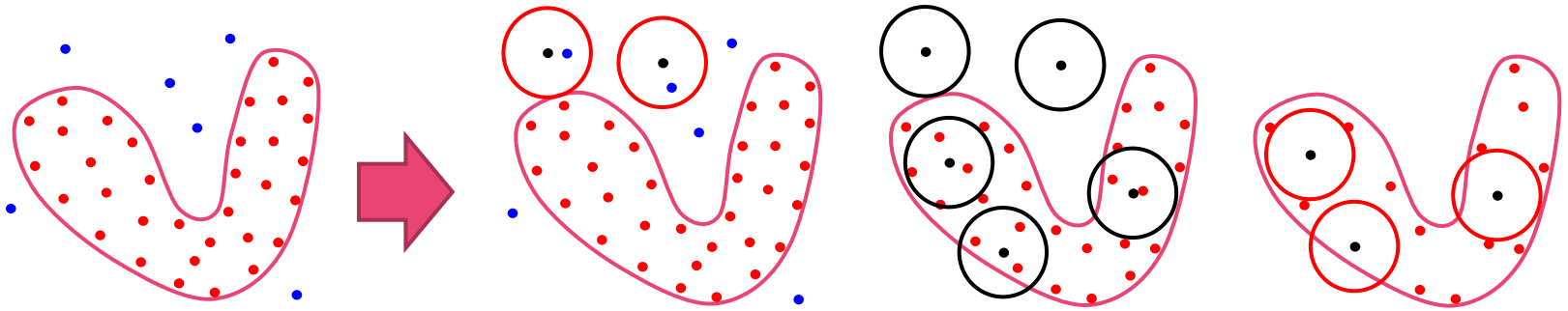
When $\psi \to \infty$, both upper and lower bounds decrease.

$$\langle AUC \rangle < \frac{p_u^2 |\mathcal{X}_N| S_d}{d} \psi\alpha^{\psi} \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_{\psi} - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_{\psi} \right) + O((1-\alpha)\psi^2), \text{ and}$$

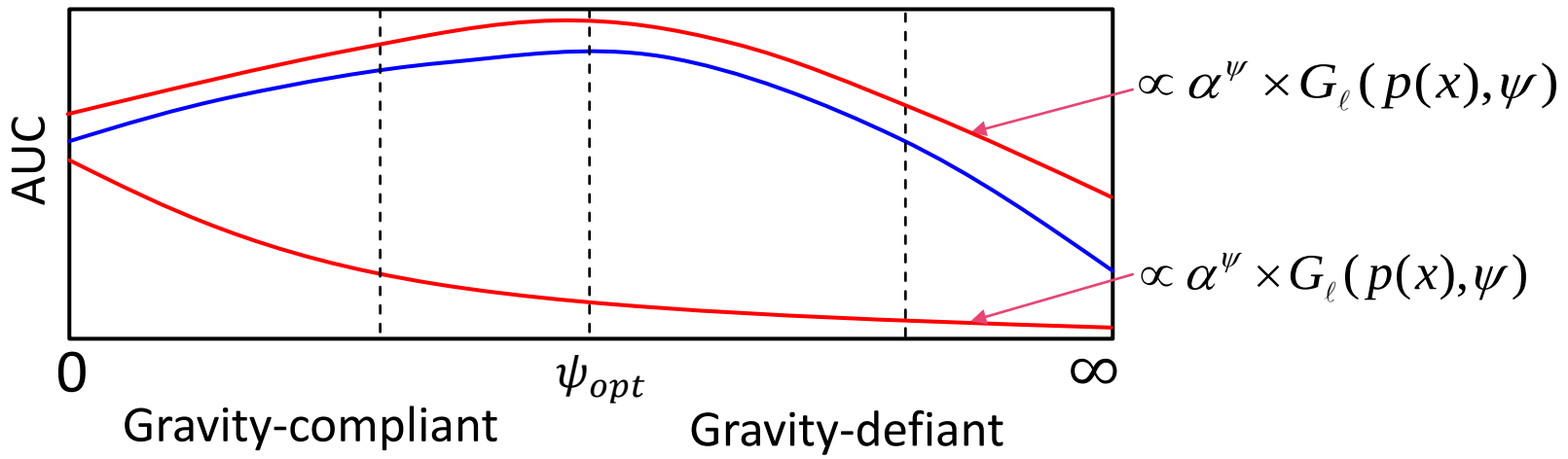
$$\langle AUC \rangle > \frac{p_\ell^2 C_\ell C_\delta}{d+1} \alpha^{\psi} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_{\psi} - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_{\psi} \right\} + O((1-\alpha)\psi).$$

Gravity Defiant Behavior

Mechanism: Subsampling, false negative (FN) and false positive (FP)



Accuracy of nearest neighbour anomaly detectors



Summary of Theory and Its Experimental Validation

Theoretical expectation based on Theorem 1

Table 1: Changes in AUC and ψ_{opt} as one data characteristic (α, $\rho_u(\mathcal{X})$ or $\rho_\ell(\mathcal{X}_N)$) changes. Δ_A is the nearest neighbour distances of anomalies.

Change in one data characteristic	$\rho_u(\mathcal{X})$	$\rho_\ell(\mathcal{X}_N)$	Δ_A	AUC	ψ_{opt}
a) α increases	$=$	$=$	$=$	$\Uparrow$	$\Uparrow$
b1) $\mathcal{X}_A$ becomes bigger	$\Uparrow$	$=$	$\Uparrow$	$\Uparrow$	*
b2) $\mathcal{X}_N$ becomes bigger	$\Downarrow$	$\Uparrow$	$\Downarrow$	$\Downarrow$	*
b3) Number of clusters in $\mathcal{X}_N$ increases	$\Downarrow$	$=$	$\Downarrow$	$\Downarrow$	*

* The direction of ψ_{opt} depends on the geometry of $\mathcal{X}$ and $\mathcal{X}_N$.

Experimental results

Table 4: Changes in error(=1-AUC), ψ_{opt} and Δ_A as data characteristics (α, $\rho_u(\mathcal{X})$, $\rho_\ell(\mathcal{X}_N)$) change, where Δ_A and Δ_N are the average nearest neighbour distance of anomalies and normal instances, respectively.

Section	Change in Data Characteristic	α	$\rho_u(\mathcal{X})$	$\rho_\ell(\mathcal{X}_N)$	Δ_A	Error	ψ_{opt}
7.2	$\mathcal{X}_A$ becomes bigger	$=$	$\Uparrow$	$=$	$\Uparrow$	$\Downarrow$	$\Downarrow$
7.3(i)	$\mathcal{X}_N$ becomes bigger	$=$	$\Downarrow$	$\Uparrow$	$\diamondsuit$	$\Uparrow$	$\Uparrow$
7.3(ii)	Number of clusters in $\mathcal{X}_N$ increases	$\uparrow$	$\Downarrow$	$\cong$	$\Downarrow$	$\Uparrow$	$\Uparrow$
7.4	Increase number of anomalies	$\Downarrow$	$\cong$	$=$	$\Downarrow$	$\Uparrow$	$\Downarrow$
7.5	Increase number of normal instances	$\Uparrow$	$=$	$\cong$	$\cong$	$\Downarrow$	$\Uparrow$

$\diamondsuit$ The increased Δ_N is more pronounced than the decreased Δ_A in Section 7.3(i).

The gravity defiant behavior is characterized by the simple theoretical AUC bounds, and it is experimentally validated.

Summary of Arguments on Gravity Defiance Behaviour

1. Zimek, Gaudet, Campello, Sander, SIG-KDD2013, 2013
   The space resolution of k-NN dist. over the change of #samples
   - ✓ The space resolution of the k-NN distance does not increase when the #samples reduces.
2. Aggarwal and Sathe, SIGKDD Explorations Newsletter, 2015
   The balance between the bias and the variance of the score
   - ✓ The balance depends on the data and the anomaly detector. It does not explain the behavior of 1-NN detector.
3. Sugiyama and Borgwardt, NIPS'13, 2013
   The optimum #samples to cover all normal areas but not any anomalous areas
   - ✓ The analysis captured a mechanism but did not include the geometrical effect of the data distribution.
4. Ting, Washio, Wells, and Aryal, Machine Learning J., 2017
   It revealed the mechanisms of both Sugiyama and Borgwardt's finding and the geometrical effect.

Contents

1. Introduction
2. Current comparative works and their recommendations
3. An analysis of strengths and weaknesses of current comparative works
4. Current bias-variance analyses applied to anomaly detection
5. Dealing with high-dimensional datasets and subspace anomaly detection
6. Non-obvious applications of anomaly detectors
7. Potential future research directions
8. Factors to consider in choosing an anomaly detector

High-dimensional Anomaly Detection

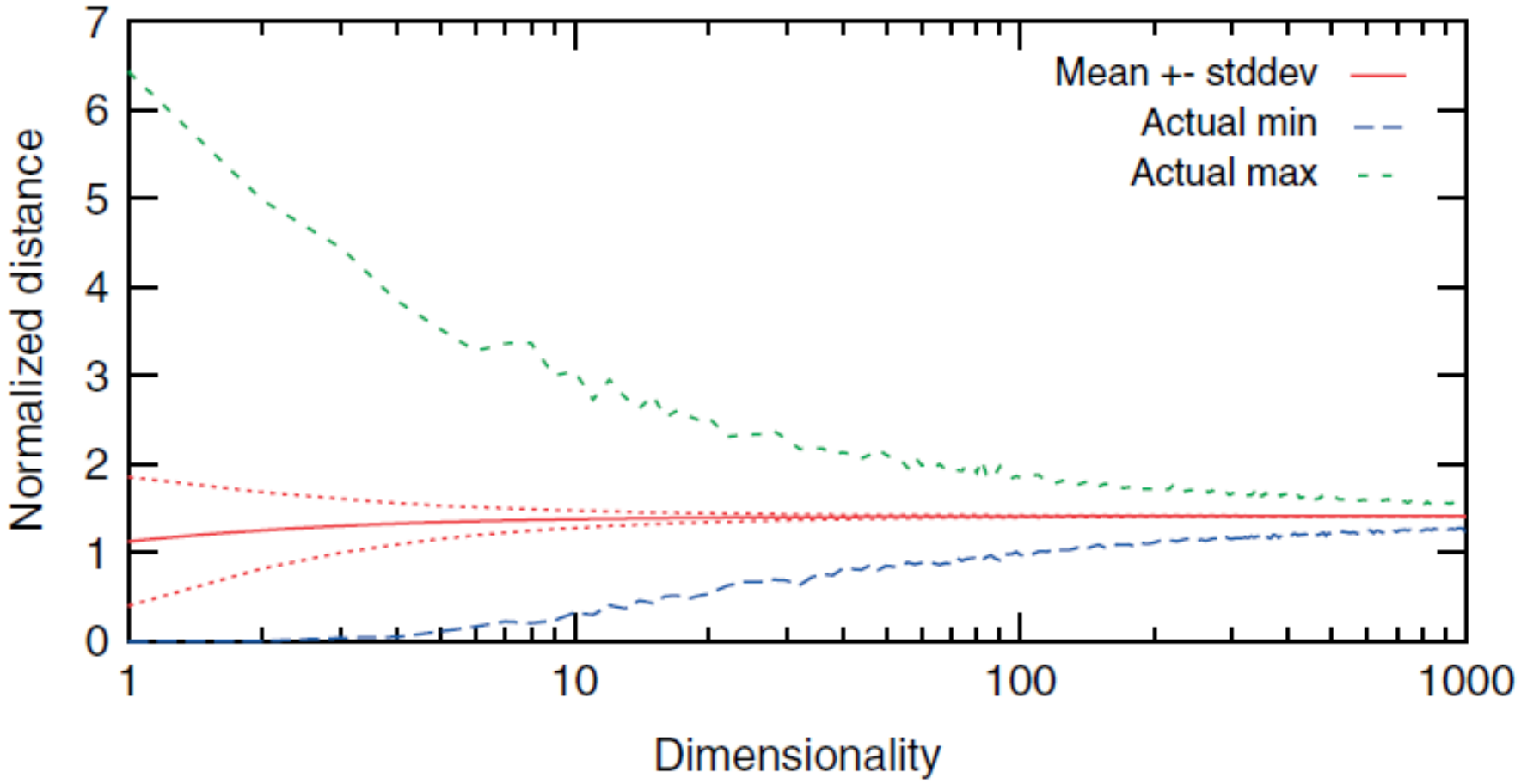
- Applications with hundreds of thousands to millions of dimensions
  - E.g., Bioinformatics, Corporate fraud detection, Malicious URL detection etc [10].
- Curse of dimensionality
  - data becomes sparse, and outliers are masked
  - all pairs of data points are almost equidistant – distance concentration
  - the concept of NN becomes meaningless
- Not all dimensions are relevant to detect all anomalies
  - Irrelevant or noisy dimensions affect the anomaly score

[10] Pang, G., Cao, L., Chen, L. and Liu, H. (2018). Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection, SIGKDD. London, UK.

Concentration effect of distance

Sample of 10^5 instances drawn from a Gaussian [0, 1], pairwise distance is normalized by $1/\sqrt{d}$

(Figure from Zimek et al [11])



[11] Zimek, A., Schubert, E., Kriegel, H.-P. (2012) A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining, vol. 5 (5) 2012, 363-387

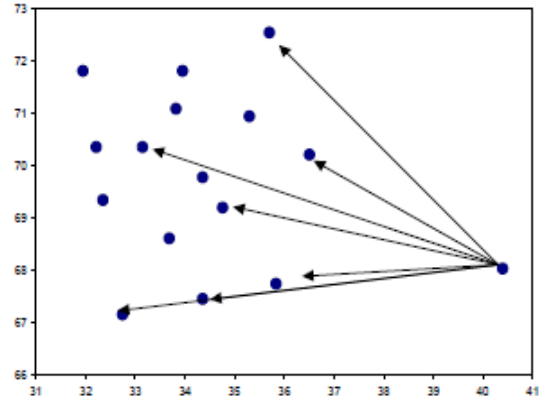
Key High dimensional anomaly detectors

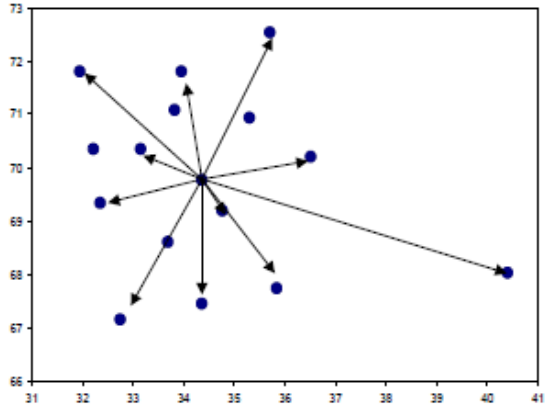
H.-P. Kriegel, M. Schubert, and A. Zimek. (2008). Angle-based outlier detection in high-dimensional data. In Proceedings KDD'08, pages 444–452

De Vries, T., Chawla, S., and Houle, M. E. (2010). Finding local anomalies in very high dimensional space, In Proceedings of the IEEE ICDM, Sydney, Australia, 128 – 137

Pang, G., Cao, L., Chen, L. and Liu, H. (2018). Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection, The 24th SIGKDD Conference on Knowledge Discovery and Data Mining. London, UK.

Angle-based anomaly detection





- Angle-based Outlier Factor (ABOF) [12]
  - variance of angles of a data point with all other points
- Claimed to be more suitable in high-dimensional datasets
  - angles are more stable than distances
  - datasets used: d=1000 (synthetic) and d=200 (real)
- Angle based measures are not immune to the curse of dimensionality [13]
  - concentration effect of the cosine measure [14]

[12] H.-P. Kriegel, M. Schubert, and A. Zimek. (2008). Angle-based outlier detection in high-dimensional data. In Proceedings KDD'08, pages 444–452

[13] Aggarwal, C. C. (2017). Outlier Analysis. Second edition. Springer International Publishing

[14] Radovanovic, M., Nanopoulos, A. and Ivanovic, M. (2010). On the Existence of Obstinate Results in Vector Space Models. ACM SIGIR Conference

PINN: Projection Index Nearest Neighbours

(De Vies et al [16, 17])

- Random Projection (RP) [15]

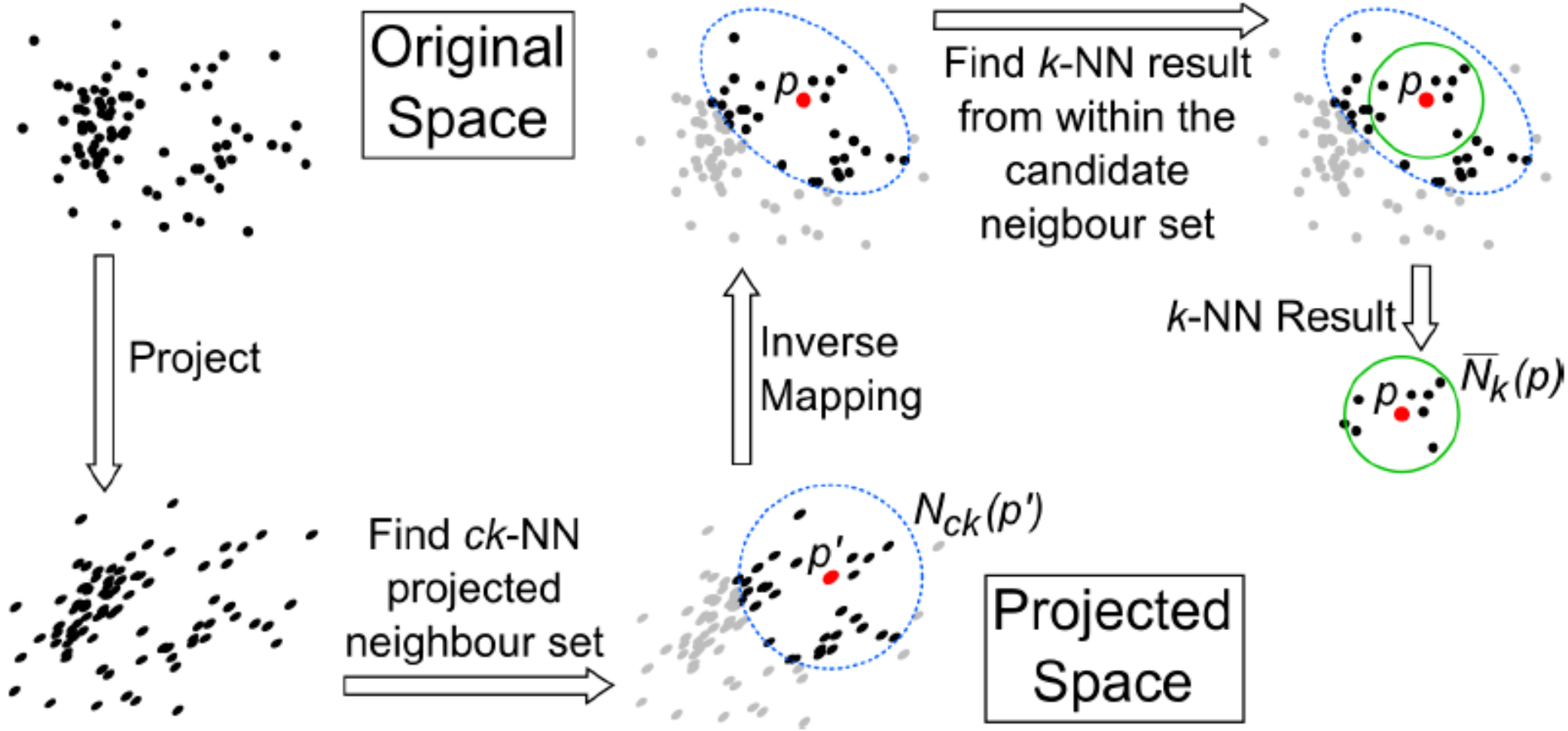
  - generates candidate kNN set efficiently in a lower-dimensional space
- Finds kNNs from the candidate set in the original space
- Uses approximate kNNs in LOF

[15]Achlioptas, D. (2001). Database-friendly random projections, In Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART, Santa Barbara, CA.

[16]De Vries, T., Chawla, S., and Houle, M. E. (2010). Finding local anomalies in very high dimensional space, In Proceedings of the IEEE ICDM, Sydney, Australia, 128 – 137.

[17]De Vries, T., Chawla, S. and Houle, M. E. (2012). Density-preserving projections for large-scale local anomaly detection, Knowledge and Information System 32(1): 25–52.

PINN Framework



(Figure by M. E. Houle)

PINN Results

- Used datasets with d up to more than 100,000
- Show that RP+PINN+LOF produces better results than RP+LOF and PCA+LOF
- This comparison is not fair because using PINN, kNNs are searched from the candidate set in the original space whereas in RP and PCA everything is done in lower-dimensional spaces
- Evaluation is based on the ground truth using standard LOF in the original space, not sure how does RP+PINN+LOF compares against the standard LOF in the original space
- Main claim is in terms of efficiency.

REPEN: Representation learning for random distance-based outlier detection (Pang et al [18])

- Learns a low-dimensional representation tailored for the fast random distance-based anomaly detector called Sp [19]
  - combines representation learning with outlier detector
  - maintains irregularities in the low-dimensional space as opposed to maintaining data regularities in general representation learning
  - can leverage prior knowledge or known anomalies

[18] Pang, G., Cao, L., Chen, L. and Liu, H. (2018). Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection, The 24th SIGKDD Conference on Knowledge Discovery and Data Mining. London, UK.

[19] Sugiyama, M. and Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. In NIPS. 467–475.

REPEN Framework

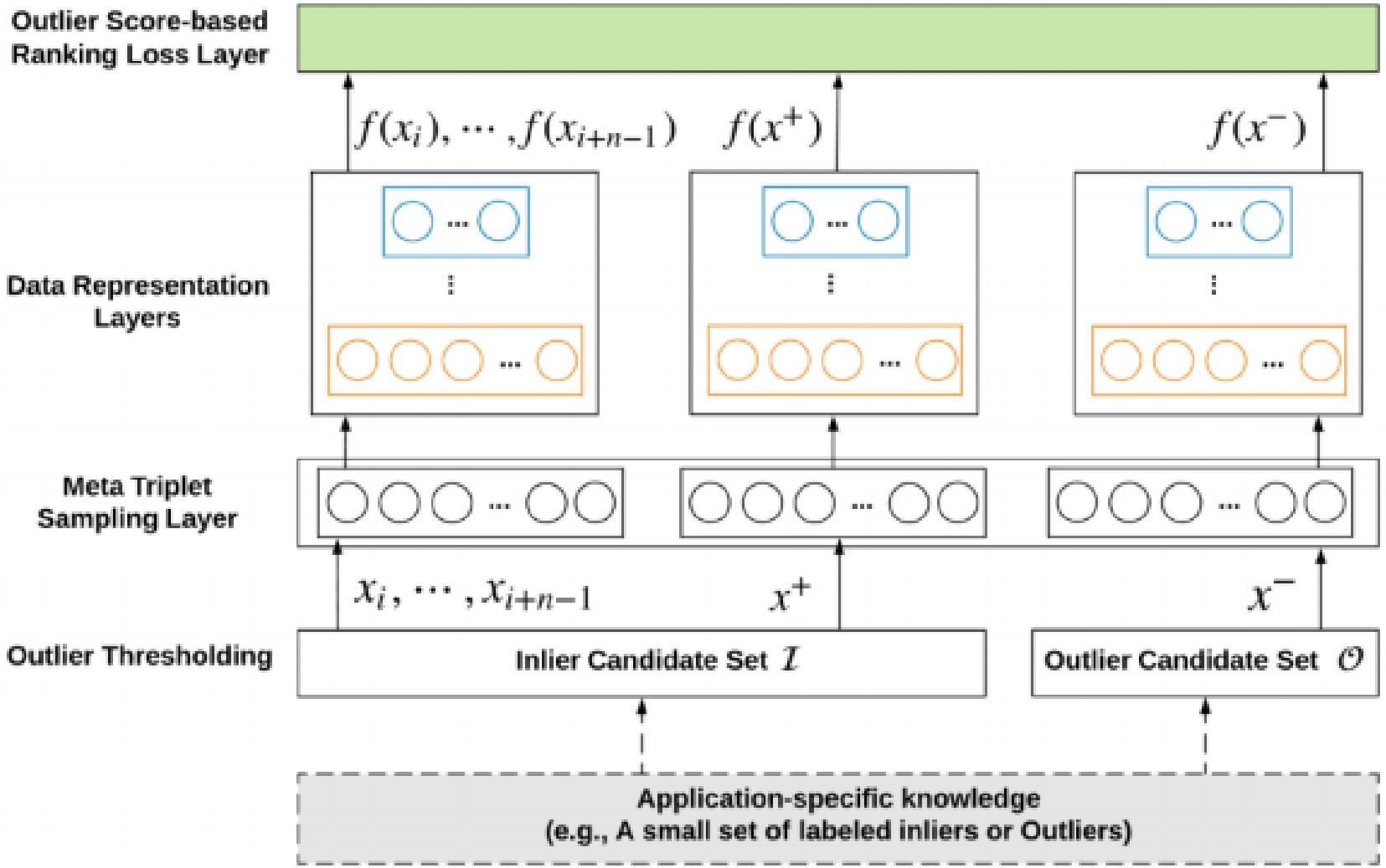


Figure from Pang et al (2018)

REPEN Results

- Used datasets with d up to more than 200,000 and n up to more than 16 millions
- Produces better results than using Sp in the original space
- Produces better results in compare to other unsupervised representation learning methods
- Significantly improves the AUC with a small number of labelled anomalies if available

Subspace anomaly detection

- Look for anomalies in lower dimensional subspaces
  - run anomaly detector in subspaces
- Useful in high-dimensional problems
  - can alleviate the curse of dimensionality
  - can help to detect anomalies masked by irrelevant dimension – some data may look anomalous in some dimensions only
- Challenges:
  - finding relevant subspaces – number of subspaces grow exponentially with d
  - different subspaces can be relevant to identify different anomalies

Key subspace anomaly detectors

Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection. In Proceedings of the SIGKDD.

Sathe, S. and Aggarwal, C. (2016). Subspace Outlier Detection in Linear Time with Randomized Hashing. In Proceedings of the ICDM.

Keller, F., Muller, E. and Bohm, K. (2012). HiCS: high contrast subspaces for density-based outlier ranking, In Proceedings of the ICDE,

Sathe, S. and Aggarwal, C. (2016). LODES: Local Density Meets Spectral Outlier Detection. In Proceedings of SIAM Conference on Data Mining.

Muller, E., Assent, I., Iglesias, P., Mulle, Y. and Bohm, K. (2012). Outlier Analysis via Subspace Analysis in Multiple Views of the Data. In Proceedings of the ICDM Conference.

Feature Bagging (Lazarevic and Kumar [20])

- Anomaly detectors are applied using different subsets of features
- Combines scores from all anomaly detectors
- Can be useful in high-dimensional problems
  - efficiency – computations in lower dimensional subspaces
  - effectiveness – ensemble of multiple subspaces
- Experiments:
  - LOF as the base anomaly detector
  - Datasets with d up to 617
- Issues:
  - outlier score in different subspaces scale differently – dimensionality bias

[20] Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection, In Proceedings of the SIGKDD, 157–166.

RS-Hash (Sathe & Aggarwal [21])

- Uses subspace grids and randomised hashing in ensemble
- For each model in an ensemble
  - a grid is constructed using subsets of features and data
  - random hashing is used to record data counts in grid cells
  - anomaly score of a data point is the log of the frequency in its hashed bins
- Results: used datasets with d up to 166
- Issue:
  - dimensionality bias

[21] Sathe, S. and Aggarwal, C. (2016). Subspace Outlier Detection in Linear Time with Randomized Hashing. ICDM Conference

HiCS: High Contrast Subspace (Keller et al [22])

- Rather than selecting subspaces at random, it uses statistical test to carefully select subspaces
  - selects subspaces with high contrast – Apriori-like exploration
  - contrast is measured as the deviation of observed and expected pdf in a subspace
  - expected pdf is estimated with dimension independence assumption
  - uses Monte Carlo samples to compute pdfs
- Discussed as a preprocessing approach
- Used LOF in the selected subspaces (other algorithms can be used)

[22] Keller, F., Muller, E. and Bohm, K. (2012). HiCS: high contrast subspaces for density-based outlier ranking, In Proceedings of the ICDE, Washington, DC,

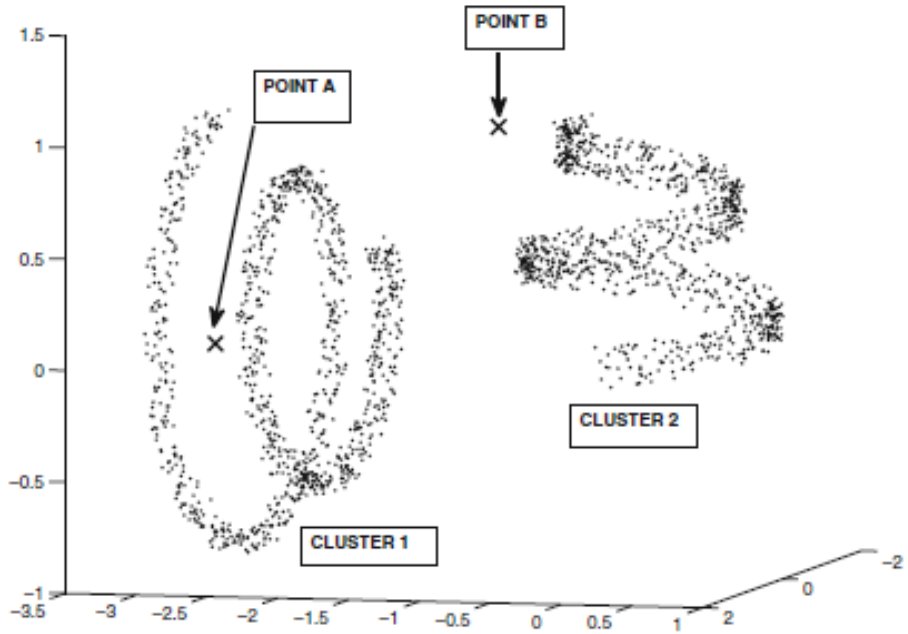
HiCS: Results and Issues

- Experimental results:
  - datasets used have d in the order of couple of hundreds

- Issues:
  - Dimensionality biasness of anomaly scores
  - Decoupling subspace search from anomaly detector to be used
  - Subspaces relevant for one algorithm may not be good for others
  - May be the selection based on contrast is good for density-based methods and not for others

Spectral Methods: Anomaly detection in nonlinear subspaces (Dang et al [25], Sathe & Aggarwal [26])

- To detect anomalies in the case where data lies in a low dimensional manifold
- Low-dimensional spectral embedding of data is learned
  - by extracting top eigenvectors of kNN graph
- Experiments: d less than 50 [26]
  d up to 960 [25]



(Figure from Sathe & Aggarwal [26])

[25] Dang, X., Misenkova, B., Assent, I., Ng, R. (2013) Outlier detection with space transformation and spectral analysis. In Proceedings of SIAM Conference on Data Mining

[26] Sathe, S. and Aggarwal, C. (2016). LODES: Local Density Meets Spectral Outlier Detection. In Proceedings of SIAM Conference on Data Mining

OutRank: Subspace anomaly detection using ensembles of subspace clustering (Muller et al [27])

- Each model in the ensemble uses the following two steps
  1. Execute a subspace clustering algorithm such as PROCLUS [28]on the data in order to create clusters in a subspace.
  2. Score a data point based on its frequency of presence in a cluster, the size and dimensionality of the cluster it belongs, and its distance to the cluster centroid.
- Data Points belonging to many clusters with a large number of points and many dimensions are unlikely to be outliers.
- Experiments – used data sets with dimensionality up to 128.

[27] Muller, E., Assent, I., Iglesias, P., Mulle, Y. and Bohm, K. (2012). Outlier Analysis via Subspace Analysis in Multiple Views of the Data. ICDM Conference.

[28] Aggarwal, C., Procopiuc, C., Wolf, J., Yu, P. and Park, J. (1999). Fast Algorithms for Projected Clustering. ACM SIGMOD Conference.

iForest (Liu et al [23])

- Can be viewed as a subspace method
- Uses ensemble of isolation trees
- Each tree is an unbalanced trees constructed from a small subsamples
  - only a small subset of dimensions are used to partition the space
  - regions (leaves) are defined using a different subset of dimensions
- Experiments: used datasets with d up to couple of hundreds

[23] Liu, F. T., Ting, K. M., Zhou, Z. H. (2008) Isolation Forest. In Proceedings of IEEE ICDM, 413-422.

ZERO++ (Pang et al [24])

- Ensemble and subspace based method
- Each model is a grid is constructed using a subset of features and a small subsamples
- Anomaly score is computed as the number of models where data lies in zero mass regions
- Runs faster compared to other subspace methods using LOF
- Experiments: Used datasets with d up to 649

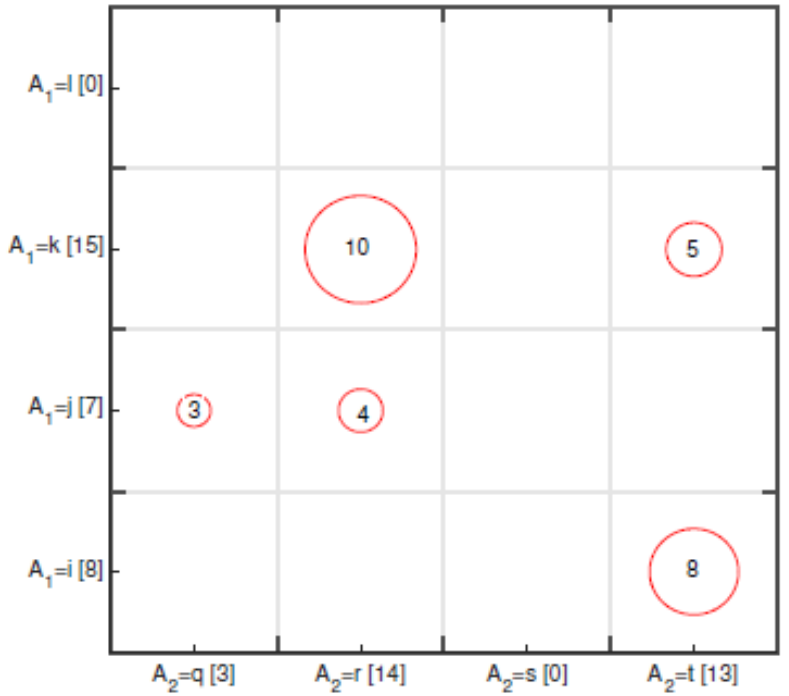


Figure from Pang et al (2016): Two-dimensional grid with 30 samples, where the size of the circle indicates the number of samples in a region; and the number in [ ] indicates the number of samples in one-dimensional subspace.

[24] Pang, G., Ting, K. M., Albrecht, D., Jin, H., (2016). ZERO++: Harnessing the power of zero appearances to detect anomalies. Journal of Artificial Intelligence Research. Vol 57, 593-620

Subspace anomaly detection: Summary

- Subspace methods identify subspaces for the entire dataset
  - do not identify subspaces for individual anomalies or groups.
- Typically employ feature sampling
- This approach is aimed for high dimensional datasets
  - where anomalies are assumed to able to identify using a small number of dimensions
  - demonstrated to work with dimensions in hundreds only
  - can not handle dimensionality in hundreds of thousands and millions
- Most of these methods uses distance-based anomaly detector and hence computationally expensive in large datasets – can not be used in datasets with millions of instances

Outlying Aspect Mining

- Given a data instance, the task is to find subspaces where it is an anomaly
- Identify relevant subspaces for an individual anomaly
- Requires evaluation of subspaces
- Examples of methods: HOS-Miner [29], OAMiner [30], Beam [31], sGrid [32]

[29] Zhang, J., Lou, M., Ling, T. W., and Wang, H. (2004). HOS-Miner: A system for detecting outlying subspaces of high-dimensional data. In Proceedings of the Thirtieth International Conference on VLDB, Toronto, Canada. 1265–1268.

[30] Duan L., Tang, G., Pei, J., Bailey, J., Campbell, A. and Tang, C. (2015). Mining outlying aspects on numeric data. Data Mining and Knowledge Discovery, 29(5):1116–1151.

[31] Vinh, N. X., Chan, J., Romano, S., Bailey, J., Leckie, C., Ramamohanarao, K. and Pei, J. (2016). Discovering outlying aspects in large datasets. Data Mining and Knowledge Discovery, 1–36

[32] Wells, J. R. and Ting, K. M. (2017). A simple efficient density estimator that enables fast systematic search. CoRR, abs/1707.00783.

Overall summary of high dimensional anomaly detection

- Most of them can not handle more than a thousand dimensions.
  - exceptions are PINN and REPEN
- Many of them use NNs search in the entire data and do not scale well to a very large data size (millions of data)
  - exceptions are RS-Hash, iForest, ZERO++, PINN and REPEN
- PINN and REPEN results suggest following promising avenues:
  - using approximate NNs (NNs search in a small subsamples) in projected low-dimensional representation
  - learning low-dimensional representation suitable for the base detector to be used
  - coupling representation learning and anomaly detector as opposed to decoupling them in many subspace methods

Needs more work

- Methods to deal hundreds of thousands to millions of dimensions
- Subspace methods:

  - dimensionality biasness of anomaly scores

  - systematic exploration of subspaces – infeasible in the case of large d
- Comprehensive survey and comparison of existing high-dimensional and/or subspace methods

  - strengths and weaknesses of each method are not clear
- Dealing with high dimensionality and large data size – ‘big data’

Contents

1. Introduction
2. Current comparative works and their recommendations
3. An analysis of strengths and weaknesses of current comparative works
4. Current bias-variance analyses applied to anomaly detection
5. Dealing with high-dimensional datasets and subspace anomaly detection
6. Non-obvious applications of anomaly detectors
7. Potential future research directions
8. Factors to consider in choosing an anomaly detector

Non-obvious applications of anomaly detectors

1. Classification under streaming emerging new class (Mu et al, TKDE 2017)
2. Measuring similarity between two points (Ting et al, KDD2016; 2018; Qin et al, AAAI 2019)
3. Ranking, e.g., in information retrieval (Zhou et al, PR 2012)

Non-obvious application #1

Classification under Streaming Emerging New Classes:

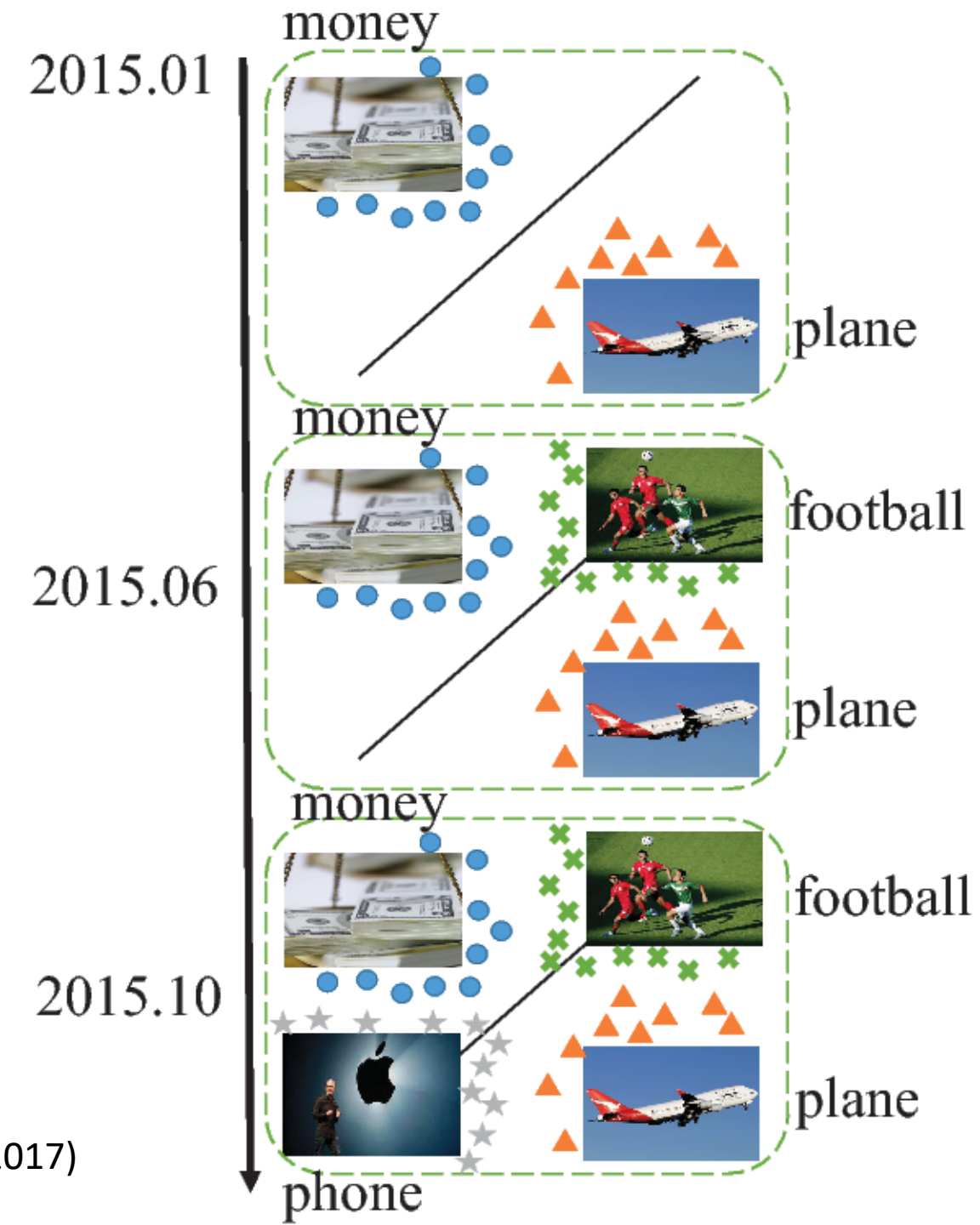
A Solution using Completely-random Trees.

Xin Mu, Kai Ming Ting and Zhi-Hua Zhou (2017). *IEEE Transactions on Knowledge and Data Engineering*, Vol 29, 1605-1618.

Problem in Data Streams

- Classification under Streaming Emerging New Classes:

  The SENC problem



(Source: Xin et al, 2017)

Three subproblems in SENC

1. Detecting emerging new classes
2. Classifying known classes
3. Updating models to integrate each new class as part of the known classes

Existing common approaches

- Treat the SENC problem as a classification problem i.e., focus on the second and the third subproblems.
- Solve it using either a supervised learner or a semi-supervised learner.
- Use a clustering method to identify new classes.

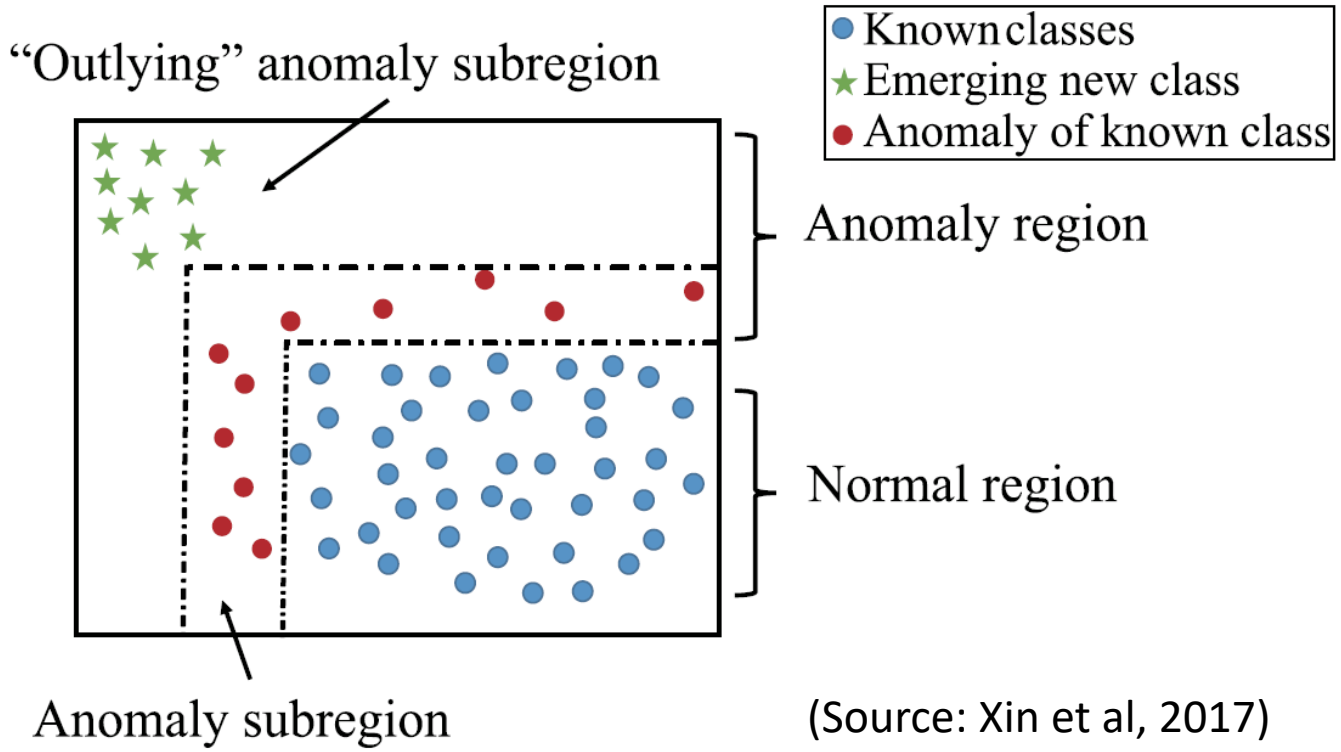
An alternative approach (Mu et al, TKDE 2017)

- Uses an unsupervised anomaly detector as the basis to solve the whole SENC problem
- Employs completely-random trees as a single common core to solve all three subproblems

Ideas

1. Intuition: instances of **emerging new classes are 'outlying anomalies'** w.r.t. the instances of known classes. The assumption is that anomalies of the known classes are more "normal" than the "outlying" anomalies.

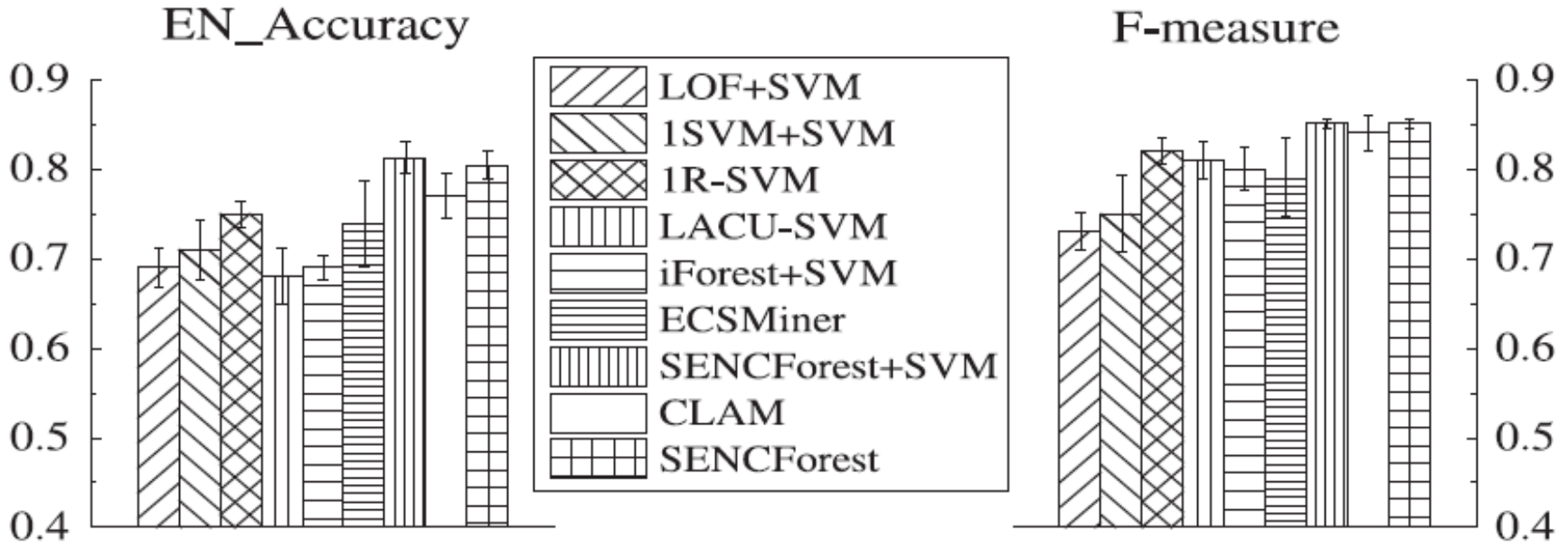
2. **Completely-randomly trees** have been shown to be competitive classifier and anomaly detector independently. This provides an opportunity to use the trees **as a common core** to solve all three subproblems in SENC.



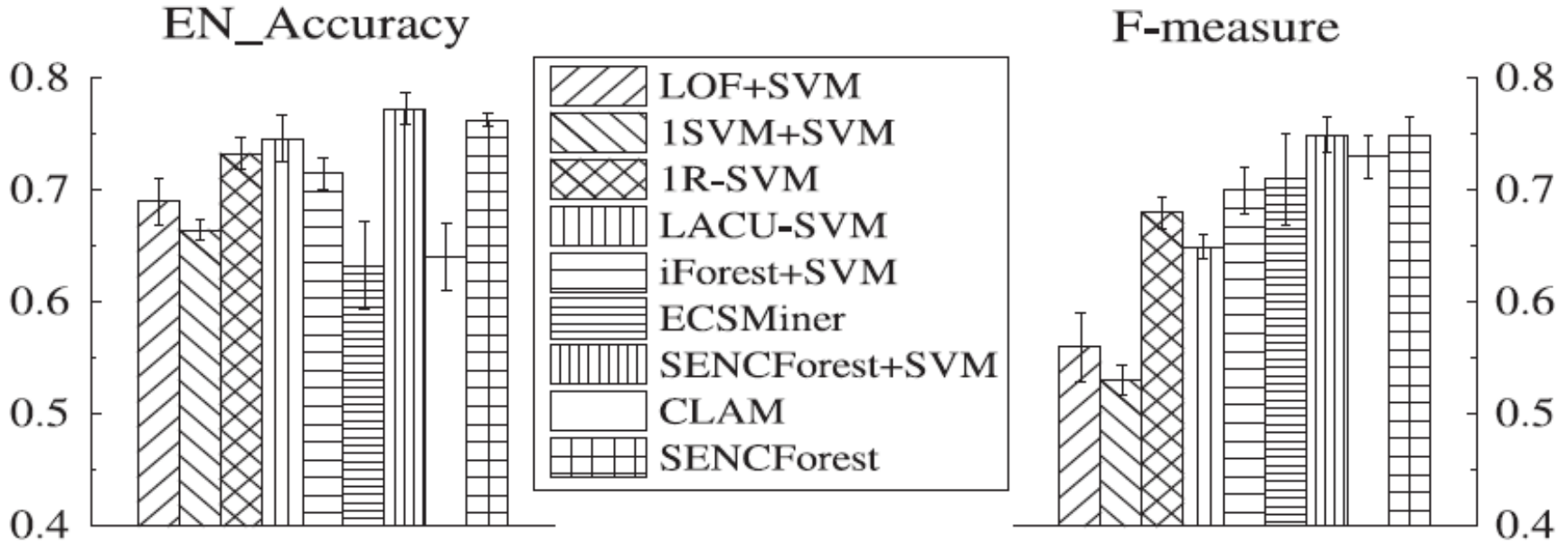
(Source: Xin et al, 2017)

Evaluation Result

Data stream without true labels
Average over 10 trials



(a) KDDCup 99



(b) Forest Cover

(Source: Xin et al, 2017)

Summary

- This work decomposes the SENC problem into three subproblems and posits that the ability to tackle the first subproblem of detecting emerging new classes is crucial for the whole problem.
- It shows that the unsupervised-anomaly-detection-focused approach, coupled with using completely-random trees, as a common core, provides a solution for the entire SENC problem.
- The empirical evaluation shows that SENCForest outperforms eight existing methods, despite the fact that it was not given the true class labels in the entire data stream; and other methods were given the true class labels at each model update. In addition, it works effectively in long streams under the limited memory environment.

Non-obvious application #2

Measuring similarity between two points

Kai Ming Ting, Ye Zhu, Mark James Carman, Yue Zhu, Zhi-Hua Zhou (2016). Overcoming Key Weaknesses of Distance-based Neighbourhood Methods using a Data Dependent Dissimilarity Measure. *Proceedings of The ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1205-1214.

Kai Ming Ting, Yue Zhu, Zhi-Hua Zhou (2018). Isolation Kernel and Its Effect on SVM. *Proceedings of The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2329-2337.

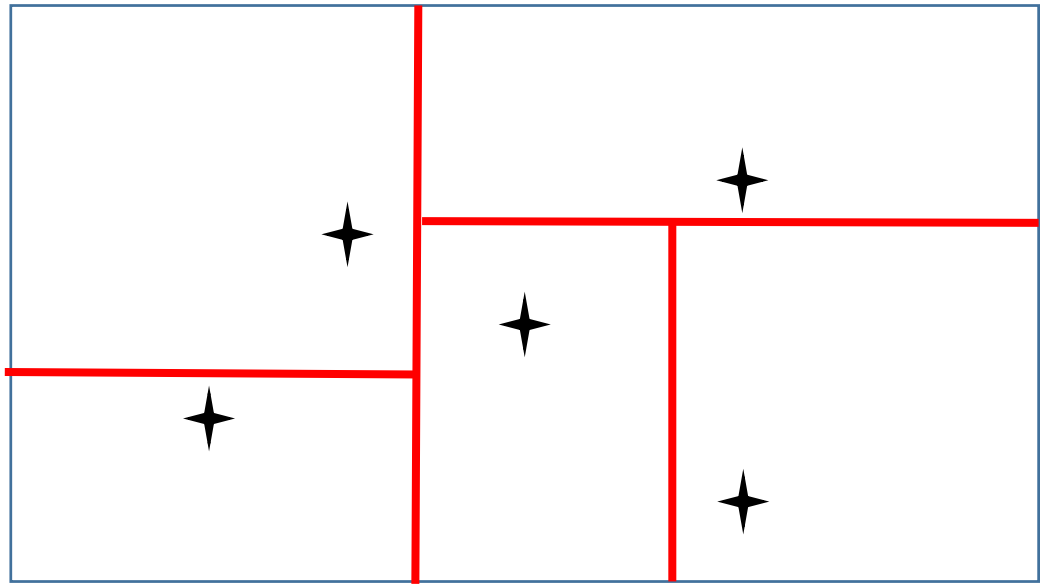
Xiaoyu Qin, Kai Ming Ting, Ye Zhu and Vincent Cheng Siong Lee (2019) Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. To appear in AAAI 2019.

Kernel choice/design issue in kernel-based methods

- Kernel-based methods: A poorly chosen kernel produces poor task-specific performances
- Three approaches:
  A. Multiple Kernel Learning – reduce the risk of choosing an unsuitable kernel by using multiple kernels (learn weights)
  B. Distance metric learning – adapt similarity measurements to the task automatically from data (space transformation)
  C. Produce data dependent kernel directly from data
- For classification tasks, all current methods rely on class information

Space Partitioning mechanism

- Space partitioning mechanism produces a partitioning H from a given sample of ψ points that partitions the data space into ψ isolating partitions, which are non-overlapping partitions covering the whole space.
- Each isolating partition $\theta \in H$ isolates one point from the rest of the points in the sample.



Definitions: Isolation Kernel (1)

Let $D = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^d$ be a dataset sampled from an unknown probability density function $x_i \sim F$. Let $\mathcal{H}_\psi(D)$ denote the set of all partitions H that are admissible under D where each isolating partition $\theta \in H$ isolates one data point from the rest of the points in a random subset $\mathcal{D} \subset D$, and $|\mathcal{D}| = \psi$.

Definition 1 *For any two points $x, y \in \mathbb{R}^d$, Isolation Kernel of x and y wrt D is defined to be the expectation taken over the probability distribution on all partitioning $H \in \mathcal{H}_\psi(D)$ that both x and y fall into the same isolating partition $\theta \in H$:*

$$K_\psi(x, y|D) = \mathbb{E}_{\mathcal{H}_\psi(D)}[\mathbb{I}(x, y \in \theta \mid \theta \in H)] \qquad (1)$$

where $\mathbb{I}(B)$ is the indicator function which outputs 1 if B is true; otherwise, $\mathbb{I}(B) = 0$.

Definitions: Isolation Kernel (2)

In practice, Isolation Kernel would be estimated from a finite number of partitionings $H_i \in \mathcal{H}_\psi(D), i = 1, \ldots, t$ as follows:

$$K_\psi(\mathbf{x}, \mathbf{y}|D) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{I}(\mathbf{x}, \mathbf{y} \in \theta \mid \theta \in H_i) \qquad (2)$$

Recall that ψ is the sample size used to generate each H_i which has ψ isolating partitions.

Require a space partitioning mechanism that produces non-overlapping isolating partitions from a sample.

A key property of human-judged similarity

Despite the widespread use of distance measures, research in psychology has pointed out since 1970's that distance measures do not possess a key property of similarity as judged by humans, i.e., the characteristic where **two instances in a dense region are less similar to each other than two instances (of the same inter point distance) in a sparse region**.

For example, two Caucasians will be judged as less similar when compared in Europe than in Asia.

Human Judged similarity:

$sim(\textbf{apple}, \textbf{apple}|Apples\ only)$

$sim(\textbf{apple}, \textbf{apple}|Pears\ \&\ Apples)$

The required space partitioning mechanism

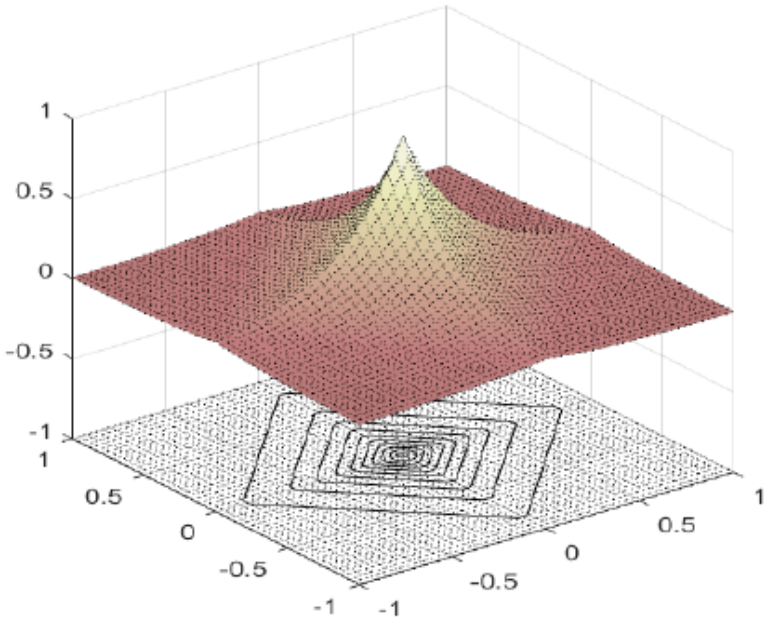
To obtain the above kernel characteristic, the required property of the space partitioning mechanism is to create partitioning *H* having

- large partitions in sparse region and
- small partitions in dense region.

Isolation Forest [1] is a space partitioning mechanism which has this property.

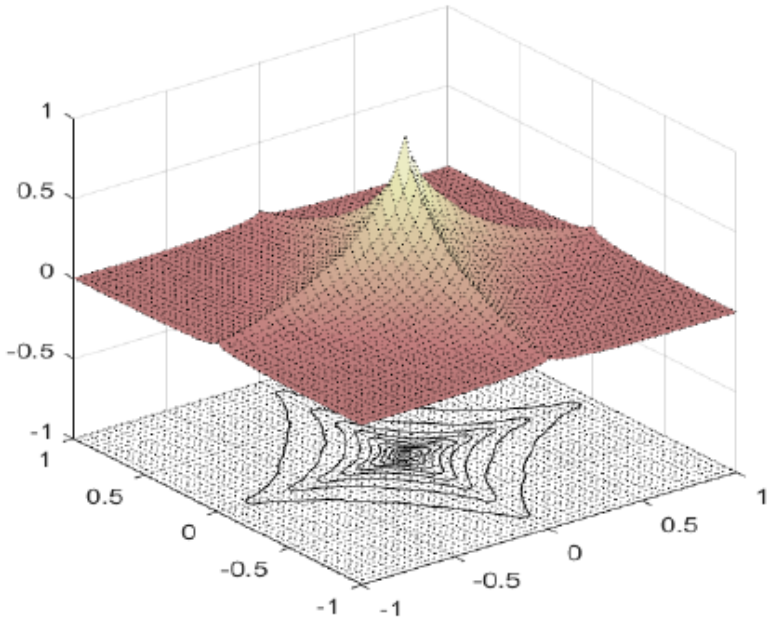
[1] Liu, F. T., Ting, K. M., Zhou, Z .H. (2008) Isolation Forest. In *Proceedings of IEEE ICDM*, 413-422.

Isolation kernel approximates Laplacian kernel under uniform density distribution



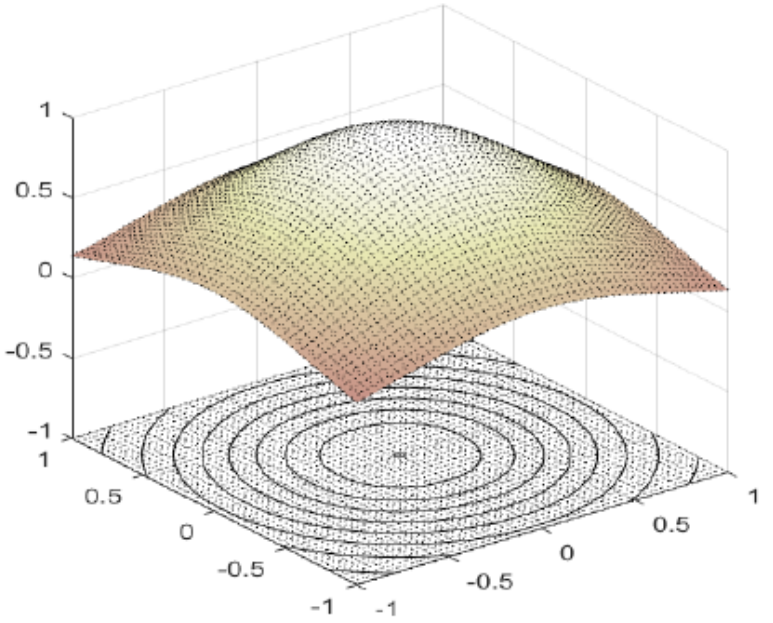
(a) Laplacian kernel

$L(x, y) = \exp(-\lambda||x - y||)$



(b) Isolation Kernel

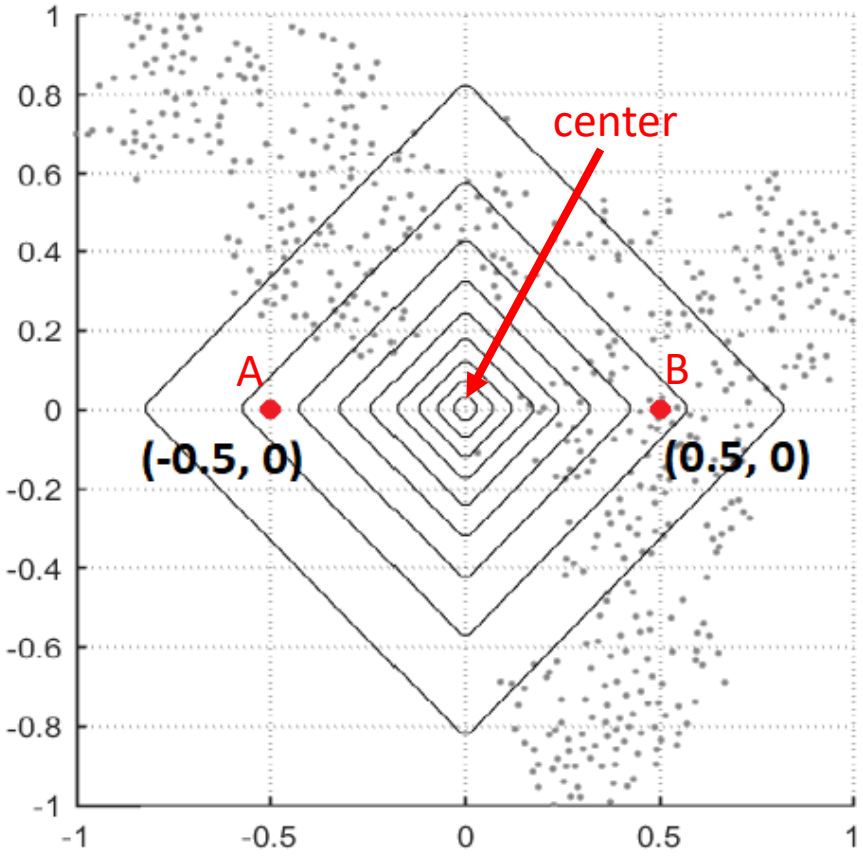
Induced from data (ψ)



(c) RBF kernel

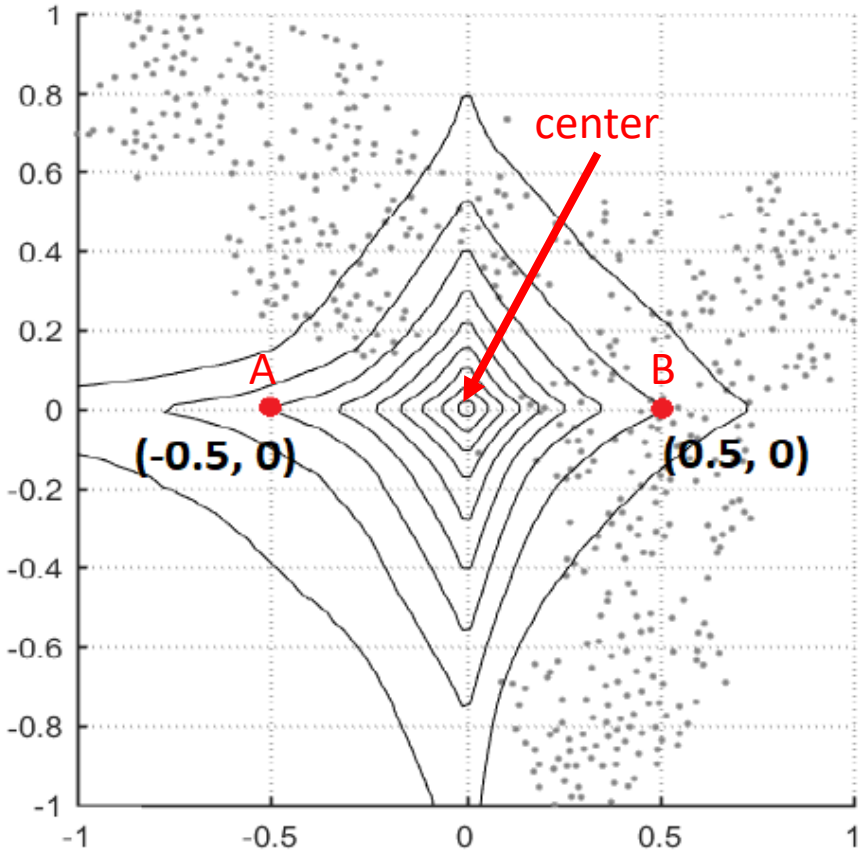
$\mathsf{G}(x, y) = \exp(-\gamma||x - y||^2)$

Isolation kernel
is adaptive to different densities



Laplacian Kernel:

$L(x, y) = \exp(-\lambda||x - y||)$



Isolation kernel:

Induced from data

Example outcome of using aNNE to implement Isolation Kernel

Nearest-Neighbour-Induced Isolation Similarity and

Its Impact on Density-Based Clustering [33]

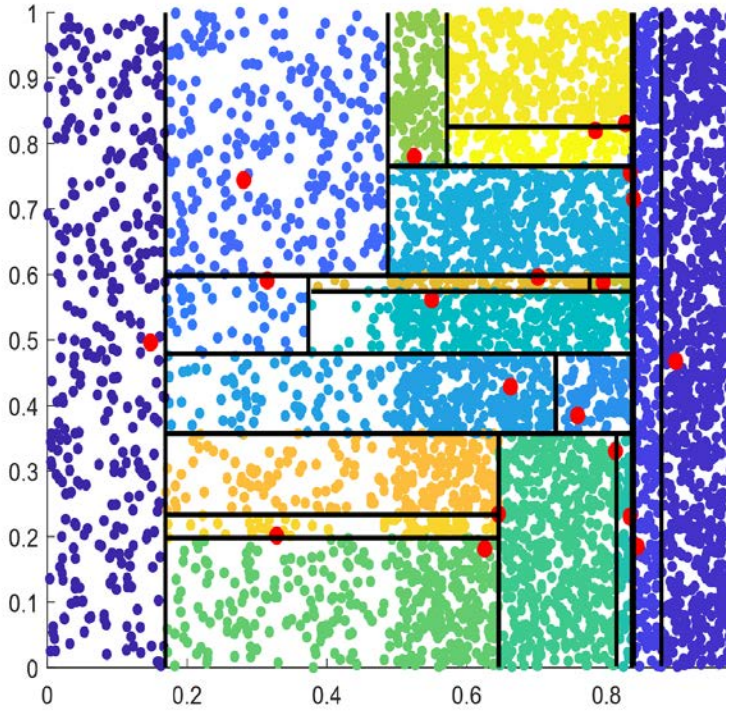
- Uses aNNE to implement Isolation Kernel
- Simply replace distance measure with Isolation Kernel in density-based clustering algorithm DBSCAN; the rest of the procedure unchanged.

[33] Qin Xiaoyu, Ting Kai Ming, Zhu Ye and Vincent Lee Cheng Siong (2019) Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. To appear in AAAI 2019.

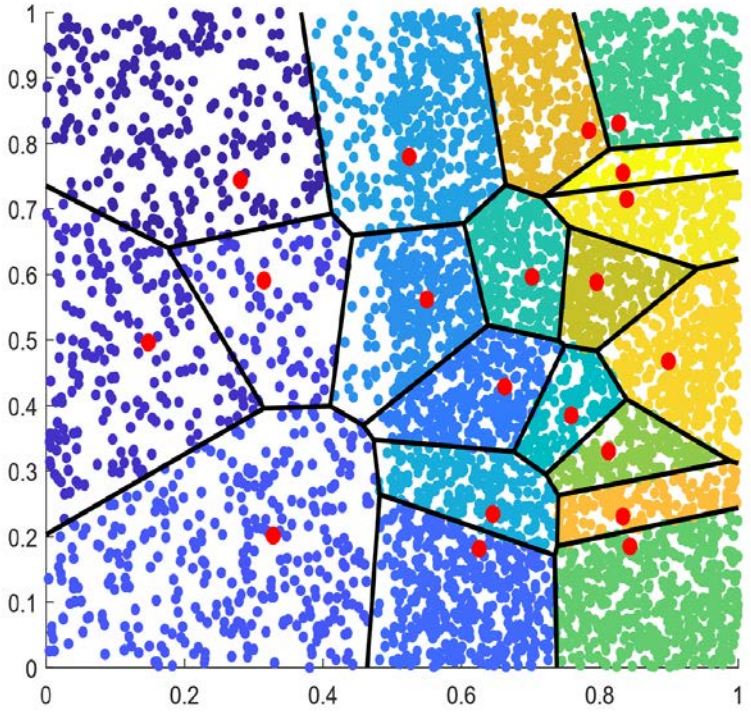
Space Partitioning: iForest vs aNNE

Shortcomings of iForest:

axis-parallel splits

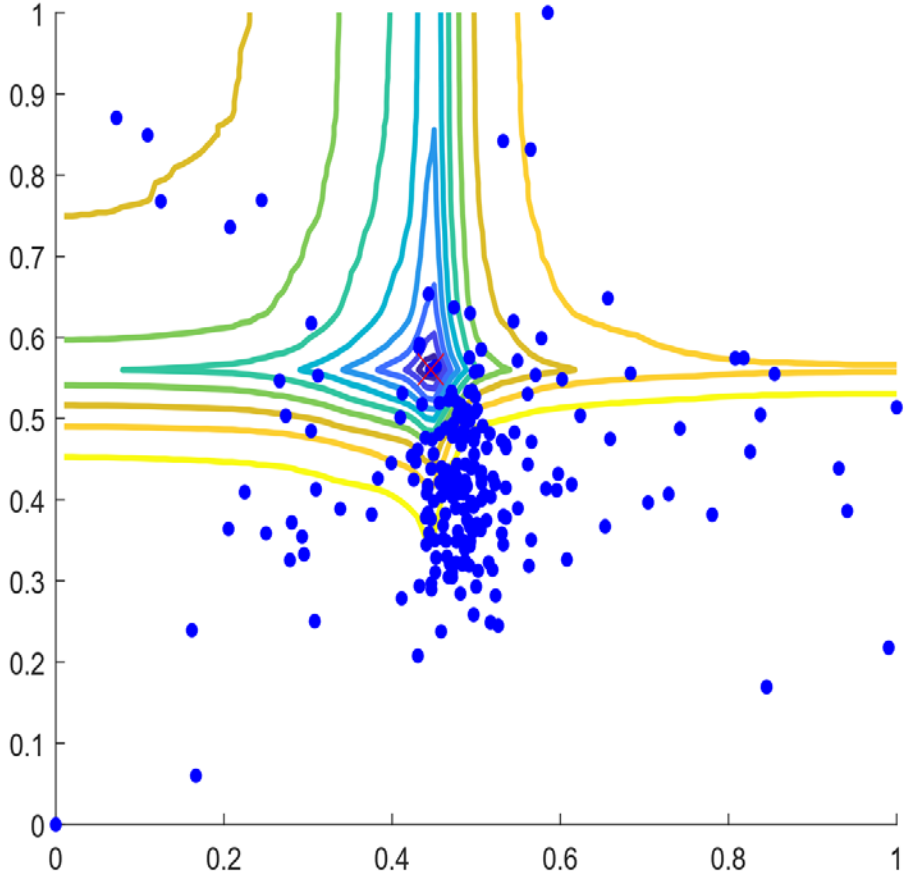


Isolation-Tree Splitting

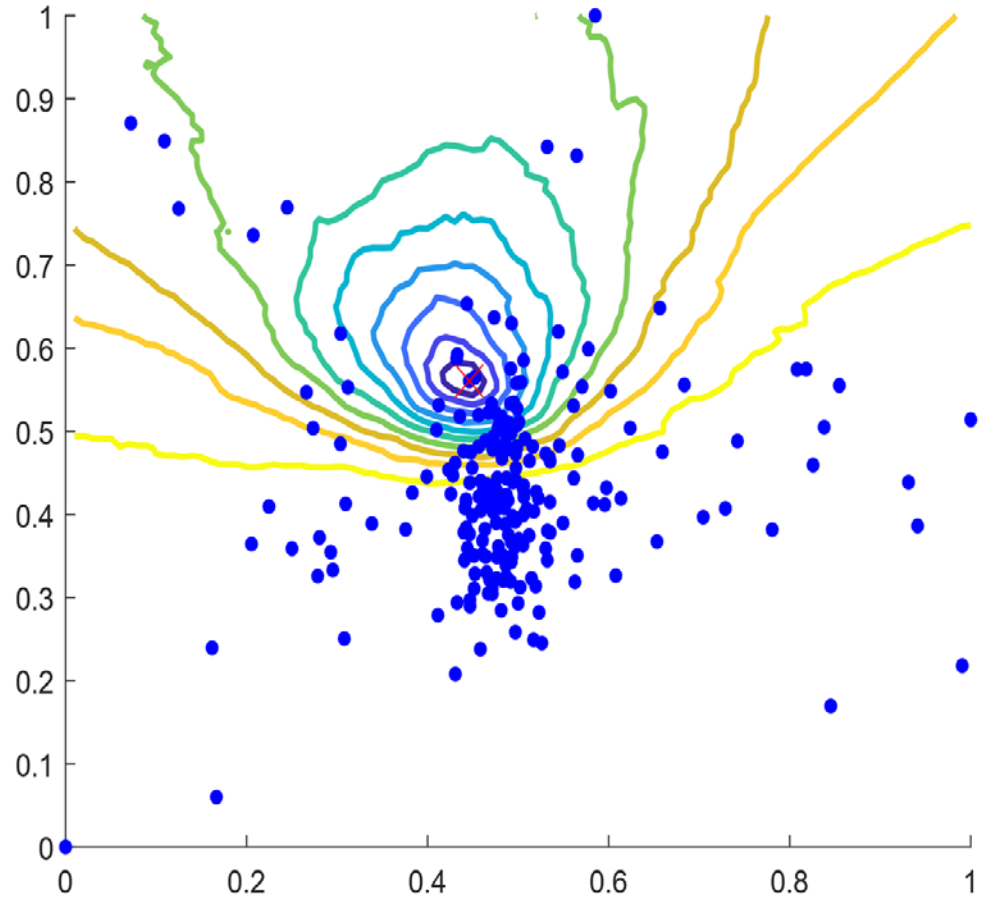


Nearest-Neighbor Splitting

Contour Maps of Isolation Kernel



iForest



aNNE

DBSCAN vs MBSCAN

Neighborhood density function

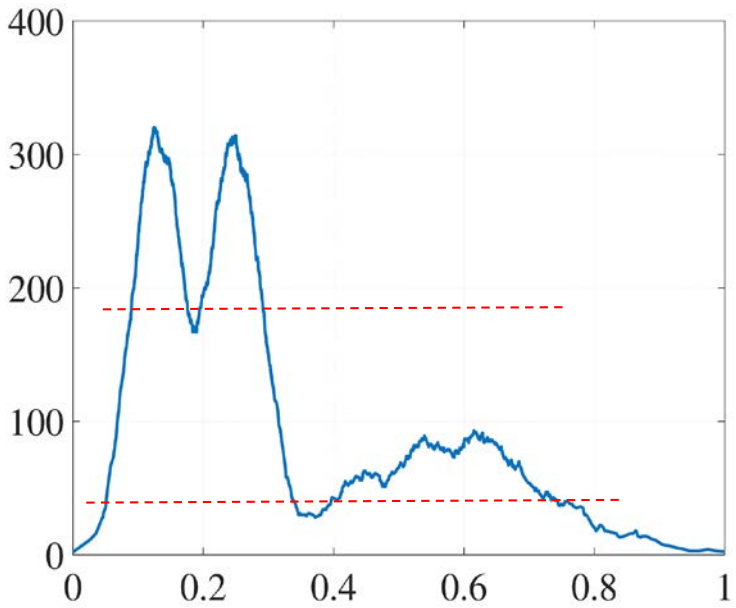
$$N_\epsilon = \#\{y \in D \mid \ell_p(x,y) \leq \epsilon\}$$

Neighborhood mass function

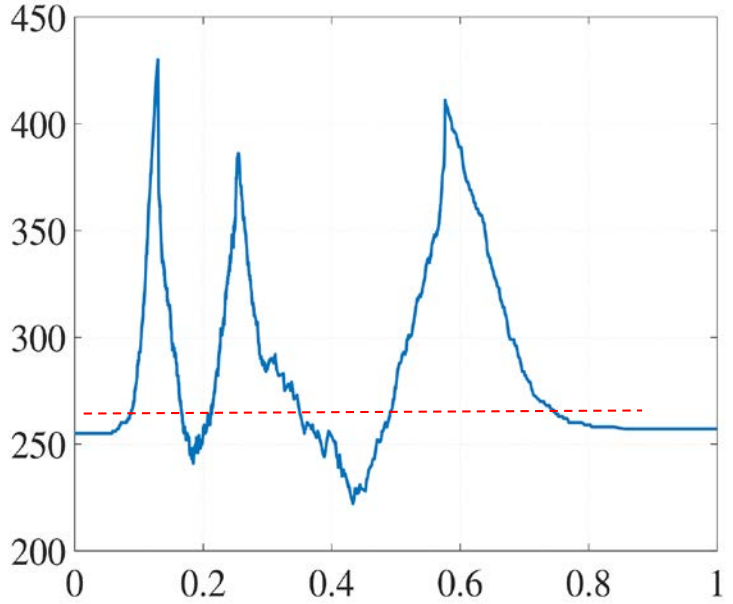
$$M_\alpha = \#\{y \in D \mid \mathfrak{p}_i(x,y) \leq \alpha\}$$

Isolation dissimilarity: $\mathfrak{p}_i(x,y) = 1 - K_\psi(x,y)$

A hard distribution for DBSCAN is not hard for MBSCAN



ϵ-neighborhood density



α-neighborhood mass

Clustering results in high-dimensional datasets

Datasets				DP	DBSCAN			MBSCAN	
Name	#Points	#Dim.	#Clusters	ℓ_2	ℓ_2	ReScale	DScale	iForest	aNNE
High-dimensional data			*(average⇒)*	*0.627*	*0.446*	*0.521*	*0.491*	*0.568*	*0.727*
ALLAML	72	7129	2	0.706	0.484	0.729	0.484	0.747	**0.820**
COIL20	1440	1024	20	0.724	0.842	0.861	0.839	0.865	**0.952**
Human Activity	1492	561	6	**0.595**	0.331	0.352	0.374	0.402	0.502
Isolet	1560	617	26	0.517	0.194	0.234	0.426	0.289	**0.605**
lung	203	3312	5	0.703	0.489	0.544	0.489	0.649	**0.921**
TOX 171	171	5748	4	0.519	0.336	0.403	0.336	0.454	**0.563**

Interesting outcomes

1. iForest and aNNE are successful implementations for Isolation Kernels.
2. Unique kernel characteristic: two points of equal inter-point distance are more similar in sparse region than those in dense region.
3. Improved SVM classifiers by simply replacing commonly used kernels with Isolation Kernel
4. Improved DBSCAN clustering by simply replacing distance measure with Isolation Kernel.
5. The result shows for the first time that it is possible to uplift the clustering performance of the classic DBSCAN, through the use of nearest-neighbour-induced Isolation Kernel, to surpass that of DP, the state-of-the-art density-based clustering algorithm.

Non-obvious application #3

ReFeat: Relevance feature mapping for Content-based multimedia information retrieval (CBMIR) [34]

- Uses iForest to generate relevant features for a query
- Uses anomaly score of the query in isolation trees to rank their relevance to the query – assigned weight to the trees

  - trees where query has larger path lengths are considered to be more relevant
- Database instances are ranked according to the weighted average of relevant feature values

[34] Zhou, G. T., Ting, K. M., Liu, F. T., Yin, Y. (2012) Relevance feature mapping for content-based multimedia information retrieval. Pattern Recognition, 45(4), 1707-1720.

Mapping data to relevant features

- Uses t isolation trees to map data from d-dimensional space to t-dimensional space

  $\boldsymbol{x} \in R^d$ is mapped to $\boldsymbol{x}' \in R^t$ as

  $$\boldsymbol{x}' = [\ell_1(\boldsymbol{x}), \ell_2(\boldsymbol{x}), \cdots, \ell_t(\boldsymbol{x})]$$

  where $\ell_t(\boldsymbol{x})$ is the path length of $\boldsymbol{x}$ in i^{th} tree
- Trees where $\boldsymbol{x}$ has longer path lengths are considered to be more relevant for $\boldsymbol{x}$
  - $\boldsymbol{x}$ is significantly different from the rest of data (anomaly) if it has shorter path length and hence the data profile captured by the tree is not very useful in retrieving data relevant to $\boldsymbol{x}$

CBMIR task – Ranking database instances

- Given a query $\boldsymbol{q}$, the relevance score of $\boldsymbol{x}$:

$$score(\boldsymbol{x}|\boldsymbol{q}) = \frac{1}{t}\sum_{i=1}^{t} w_i(\boldsymbol{q}) \times \ell_i(\boldsymbol{x})$$

where $w_i(\boldsymbol{q}) = \ell_i(\boldsymbol{q})/C$ and C is a normalisation term

Query Image

Relevant Image

Irrelevant Image

Source of images: Zhou et al. (2012)

ReFeat: Conclusions

- Isolation trees based relevance features have richer information than the original features
- Ranking of database instances does not require to measure their (dis)similarity with the query
  - can run significantly faster as the mapping of database instances can be learned and stored in the preprocessing
  - online retrieval only requires to find path lengths of $\boldsymbol{q}$ in each itree which can be done in constant time as it is independent of database size and dimensionality.

Contents

Potential future research

- Issues in high dimensional datasets
  - Distance measure issue rather than algorithmic issue
  - Resolving the concentration of measure issue
- The difference between noise and anomalies
  - Noise often outnumber anomalies
  - How to make use of feedback
- Dealing with change of interest or different interests
- Efficient subspace search in subspace anomaly detection & outlying aspect mining
- How to make use of label information/domain knowledge when available

Top recommended anomaly detectors from the three studies: Potentials

- Isolation based methods: Different isolation mechanisms
- Nearest neighbour-based methods: Well explored methods
- Kernel methods: have untapped potentials

Factors to consider when choosing an anomaly Detector

- Few parameters
  - parameter-free the best
  - Easy to tune; not too sensitive to parameter setting
- Fast runtime: can scale up to large datasets and high dimensional datasets
- Low space complexity
- Known behaviours under different data properties
- Can deal with different types of anomalies
- Its ability to deal with high dimensional problems
- Understand the nature of anomalies and the best match algorithm

References (1)

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

Aggarwal, C. C. (2017) Outlier Analysis. Second edition. Springer International Publishing.

Aggarwal, C., Procopiuc, C., Wolf, J., Yu, P. and Park, J. (1999). Fast Algorithms for Projected Clustering. ACM SIGMOD Conference.

Aggarwal, C. C. and Sathe, S. (2015) Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explorations* 17(1):24-47

Aggarwal, C. C. and Sathe, S. (2017) Outlier Ensembles: An Introduction. Springer International Publishing.

Aggarwal, C. C. and Yu, P. S. (2001) Outlier Detection in High Dimensional Data, In *Proceedings of ACM SIGMOD Conference*.

Azmandian, F., Yilmazer, A., Dy, J. G., Aslam, J. and Kaeli, D. R. (2012) GPU-accelerated feature selection for outlier detection using the local kernel density ratio. In *Proceedings of IEEE ICDM*, 51–60.

Bandaragoda, T. R., Ting, K. M., Albrecht, D., Liu F. T. and Wells, J. R. (2018). Isolation-based Anomaly Detection using Nearest Neighbour Ensembles. *Computational Intelligence*. Doi:10.1111/coin.12156.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. 26-33.

Bay, S. D. and Schwabacher, M. (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *KDD*, 29–38. ACM Press.

References (2)

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

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenkova, B., Schubert, E., Assent, I. and Houle, M. E. (2016) On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. *Data Mining and Knowledge Discovery*, 30(4), 891–927.

Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly Detection: A Survey, *ACM Computing Surveys.* 41, 3, Article 15 (July 2009), 58 pages

Chen, J., Sathe, S., Aggarwal, C. and Turaga, D. (2017) Outlier Detection with Autoencoder Ensembles. In *Proceedings of SIAM Conference on Data Mining.*

Cover, T. M. and Hart, P. E. (1967) Nearest Neighbour Pattern Classification. *IEEE Transactions on Information Theory*, vol IT-13, 21-27.

Dang, X., Misenkova, B., Assent, I. and Ng, R. (2013) Outlier detection with space transformation and spectral analysis. In *Proceedings of* SIAM *SDM conference*

De Vries, T., Chawla, S. and Houle, M. E. (2010). Finding Local Anomalies in Very High Dimensional Space, In *Proceedings of IEEE International Conference on Data Mining*, 128-137.

De Vries, T., Chawla, S. and Houle, M. E. (2012). Density-preserving projections for large-scale local anomaly detection, Knowledge and Information System 32(1): 25–52.

Duan L., Tang, G., Pei, J., Bailey, J., Campbell, A. and Tang, C. (2015). Mining outlying aspects on numeric data. Data Mining and Knowledge Discovery, 29(5):1116–1151.

References (3)

Emmott, A. F., Das, S., Dietterich, T. G., Fern, A. and Wong, W.-K. (2013). Systematic construction of anomaly detection benchmarks from real data. *ODD '13 Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. 16-21.

Emmott, A. F., Das, S., Dietterich, T. G., Fern, A. and Wong, W.-K. (2016). A Meta-Analysis of the Anomaly Detection Problem. *arXiv*:1503.01158

Evans, D., Jones, A. J. and Schmidt, W. M. (2002) Asymptotic moments of near-neighbour distance distributions. *Proceedings: Mathematical, Physical and Engineering Sciences* 458(2028):2839-2849

Goldstein, M. and Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PloS One*, 11(4), e0152173.

Guha, S., Mishra, N., Roy, G., and Schrijver, O. (2016) Robust Random Cut Forest Based Anomaly Detection On Streams. In *Proceedings of ICML Conference*, 2712–2721.

Halevy, A., Norvig, P. and Pereira, F. (2009). The Unreasonable Effectiveness of Data, in *IEEE Intelligent Systems*, vol. 24(2), pp. 8-12

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2), 171-186.

Hawkins, D. (1980) Identification of outliers. *Monographs on Applied Probability and Statistics*.

He, Z., Deng, S. and Xu, X. (2005) A Unified Subspace Outlier Ensemble Framework for Outlier Detection, *Advances in Web Age Information Management*.

Hoffmann, H. (2007). Kernel PCA for Novelty Detection, *Pattern Recognition*, 40(3), 863–874.

Keller, F., Muller, E. and Bohm, K. (2012) HiCS: High-Contrast Subspaces for Density-based Outlier Ranking, In *Proceedings of IEEE ICDE Conference*, 1037-1048.

Reference (4)

Kriegel, H-P., Schubert, M. and Zimek, A. (2008) Angle-based outlier detection in high-dimensional data. In *Proceedings of KDD*, 444-452.

Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection, KDD 157–166

Liu, F. T., Ting, K. M. and Zhou, Z .H. (2008) Isolation Forest. In *Proceedings of IEEE ICDM*, 413-422.

Mu, X., Ting, K. M. and Zhou, Z. H. (2017). Classification under Streaming Emerging New Classes: A Solution using Completely-random Trees. *IEEE Transactions on Knowledge and Data Engineering*, Vol.29, Issue.8, 1605-1618.

Muller, E., Assent, I., Iglesias, P., Mulle, Y. and Bohm, K. (2012) Outlier Ranking via Subspace Analysis in Multiple Views of the Data, In *Proceedings of ICDM Conference*.

Muller, E., Schiffer, M. and Seidl, T. (2011) Statistical Selection of Relevant Subspace Projections for Outlier Ranking. In *Proceedings of ICDE Conference*, 434–445.

Nguyen, X. V., Chan, J., Simone R., Bailey, J., Leckie, C., Kotagiri R. and Pei, J. (2016). Discovering outlying aspects in large datasets. *Data Mining and Knowledge Discovery*. Volume 30, Issue 6, pp 1520–1555.

Orair, G., Teixeira, C., Meira Jr, W., Wang, Y. and Parthasarathy, S. (2010). Distance-Based Outlier Detection: Consolidation and Renewed Bearing. *Proceedings of the VLDB Endowment*, 3(1–2), pp. 1469–1480.

Pang, G., Ting, K. M., Albrecht, D. and Jin, H., (2016). ZERO++: Harnessing the power of zero appearances to detect anomalies. *Journal of Artificial Intelligence Research*. Vol 57, 593-620.

Pang, G., Cao, L., Chen, L., Lian, D. and Liu, H. (2018) Sparse Modeling-based Sequential Ensemble Learning for Effective Outlier Detection in High-dimensional Numeric Data, AAAI

Reference (5)

Pang, G., Cao, L., Chen, L. and Liu, H. (2018). Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection, KDD. 2041-2050.

Pevny, T. (2016) Loda: Lightweight On-line Detector of Anomalies. *Machine Learning*, 102(2), 275–304.

Pimentel, M. A. F., Clifton, D. A., Clifton and L., Tarassenko, L. (2014) A review of novelty detection, *Signal Processing*, Vol. 99, 215-249.

Qin, X., Ting, K. M., Zhu, Y. and Lee, C. S. V. (2019). Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. To appear in AAAI 2019.

Radovanovic, M., Nanopoulos, A. and Ivanovic, M. (2010). On the Existence of Obstinate Results in Vector Space Models. *ACM SIGIR Conference*.

Ramaswamy, S., Rastogi, R. and Shim, K. (2000) Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD international conference on Management of data*, 427–438. ACM Press.

Rayana, S., Zhong, W. and Akoglu, L. (2016) Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective. In *Proceedings of IEEE ICDM Conference*.

Rayana, S. and Akoglu, L. (2016) Less is More: Building Selective Anomaly Ensembles. *ACM Transactions on Knowledge Discovery and Data Mining*, 10(4), 42.

Sathe S. and Aggarwal C. (2013) LODES: local density meets spectral outlier detection. *SIAM SDM conference*.

Sathe, S. and Aggarwal, C. (2016) Subspace Outlier Detection in Linear Time with Randomized Hashing. In *Proceedings of ICDM Conference*. Also appear in *Knowledge Information Systems* (2018) 56:691-715.

Reference (6)

Scholkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. and Platt, J. C. (2000) Support-vector Method for Novelty Detection, *Advances in Neural Information Processing Systems*.

Sugiyama M. and 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. and Aryal, S. (2017). Defying the Gravity of Learning Curve: A Characteristic of Nearest Neighbour Anomaly Detectors. *Machine Learning*. Vol 106, Issue 1, 55-91.

Ting, K. M., Zhu, Y., Carman, M. and Zhu, Y. (2016) Overcoming Key Weaknesses of Distance-Based Neighbourhood Methods using a Data Dependent Dissimilarity Measure. *KDD*, 1205-1214.

Ting, K. M., Zhu, Y. and Zhou, Z. H. (2018). Isolation Kernel and Its Effect on SVM. *KDD*. 2329-2337.

Wells, J. R. and Ting, K. M. (2017). A simple efficient density estimator that enables fast systematic search. CoRR, abs/1707.00783.

Zhang, J., Lou, M., Ling, T. W., and Wang, H. (2004). HOS-Miner: A system for detecting outlying subspaces of high-dimensional data. In Proceedings of the Thirtieth International Conference on VLDB, Toronto, Canada. 1265–1268.

Zhou, C. and Paffenroth, R. C. (2017) Anomaly detection with robust deep autoencoders. In *Proceedings of SIGKDD*. ACM, 665–674.

Zhou, G. T., Ting, K. M., Liu, F. T. and Yin, Y. (2012) Relevance feature mapping for content-based multimedia information retrieval. *Pattern Recognition*, 45(4), 1707-1720.

Reference (7)

Zimek, A., Gaudet, M., Campello, R. J. and Sander, J. (2013) Subsampling for efficient and effective unsupervised outlier detection ensembles. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 428-436

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