Defying the gravity of learning curve: a characteristic of nearest neighbour anomaly detectors

Kai Ming Ting¹ · Takashi Washio² ·
Jonathan R. Wells¹ · Sunil Aryal¹

Received: 20 January 2016 / Accepted: 12 July 2016 / Published online: 29 August 2016
© The Author(s) 2016

Abstract Conventional wisdom in machine learning says that all algorithms are expected to follow the trajectory of a learning curve which is often colloquially referred to as ‘more data the better’. We call this ‘the gravity of learning curve’, and it is assumed that no learning algorithms are ‘gravity-defiant’. Contrary to the conventional wisdom, this paper provides the theoretical analysis and the empirical evidence that nearest neighbour anomaly detectors are gravity-defiant algorithms.

Keywords Learning curve · Anomaly detection · Nearest neighbour · Computational geometry · AUC

1 Introduction

In the machine learning context, learning curve describes the rate of task-specific performance improvement of a learning algorithm as the training set size increases. A typical learning curve of a learning algorithm is provided in Fig. 1. The error, as a measure of the learning algorithm’s performance, decreases at a fast rate when the training sets are small; and the

Editor: Joao Gama.

✉ Jonathan R. Wells
j.wells@federation.edu.au

Kai Ming Ting
kaiming.ting@federation.edu.au

Takashi Washio
washio@ar.sanken.osaka-u.ac.jp

Sunil Aryal
sunil.aryal@federation.edu.au

¹ School of Engineering and Information Technology, Federation University, Churchill, Australia

² The Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Japan

Fig. 1 Examples of typical learning curve and gravity-defiant learning curve

rate of decrease slows gradually until it reaches a plateau as the training sets increase to large sizes.

Conventional wisdom in machine learning says that all algorithms are expected to follow the trajectory of a learning curve, though the actual rate of performance improvement may differ from one algorithm to another. We call this ‘the gravity of learning curve’, and it is assumed that no learning algorithms are ‘gravity-defiant’.

Recent research (Liu et al. 2008; Zhou et al. 2012; Sugiyama and Borgwardt 2013; Wells et al. 2014; Bandaragoda et al. 2014; Pang et al. 2015) has provided an indication that some algorithms may defy the gravity of learning curve, i.e., these algorithms can learn a better performing model using a small training set than that using a large training set. However, no concrete evidence of the ‘gravity-defiant’ behaviour is provided in the literature, let alone the reason why these algorithms behave this way.

‘Gravity-defiant’ algorithms have a key advantage of producing a good performing model using a training set significantly smaller than that required for ‘gravity compliant’ algorithms. They will yield significant saving on time and memory space that the conventional wisdom thought impossible.

This paper focuses on nearest neighbour-based anomaly detectors because they have been shown to be one of the most effective class of anomaly detectors (Breunig et al. 2000; Sugiyama and Borgwardt 2013; Wells et al. 2014; Bandaragoda et al. 2014; Pang et al. 2015).

This paper makes the following contributions:

1. Provide a theoretical analysis of nearest neighbour-based anomaly detection algorithms which reveals that their behaviours defy the gravity of learning curve. This is the first analysis in machine learning research on learning curve behaviour that is based on computational geometry, as far as we know.
2. The theoretical analysis provides an insight into the behaviour of the nearest neighbour anomaly detector. 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. First, increasing sample size increases the likelihood of anomaly contamination in the sample; and any inclusion of anomalies in the sample increases the false negative rate, thus, lowers the AUC. Second, the optimal sample size depends on the data distribution. As long as the data distribution is not sufficiently represented by the current sample, increasing the sample size will improve AUC. The optimal size is the number of instances best represents the geometry of normal instances and anomalies; this gives the optimal separation between normal instances and anomalies, encapsulated

as the average nearest neighbour distance to anomalies. Third, increasing the sample size decreases the average nearest neighbour distance to anomalies. Increasing beyond the optimal sample size reduces the separation between normal instances and anomalies smaller than the optimal. This leads to the decreased AUC and gives rise to the gravity-defiant behaviour.

3. Present empirical evidence of the gravity-defiant behaviour using three nearest neighbour-based anomaly detectors in the unsupervised learning context.

In addition, this paper uncovers two features of nearest neighbour anomaly detectors:

- A. Some nearest neighbour anomaly detector can achieve high detection accuracy with a significantly smaller sample size than others.
- B. Any change in geometrical data characteristics, which affects the detection error, manifests as a change in nearest neighbour distance such that the detection error and anomalies' nearest neighbour distances change in opposite directions. Because nearest neighbour distance can be measured easily and other indicators of detection accuracy are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams. Note that the change in sample size (described in (2) above) does not alter the geometrical data characteristics discussed here.

In the age of big data, the revelation and the knowledge about the gravity-defiant behaviour discovered in this paper have two impacts. First, the capacity provided by big data infrastructures would be overkill because the gravity-defiant algorithms that produce good performing models using small datasets can be executed comfortably in existing computing infrastructures. Second, it opens a whole new direction of research into different types of gravity-defiant algorithms which can achieve high performance with small sample size.

The rest of the paper is organised as follows. We review current anomaly detectors which are reported to perform well using small training sets in the next section. Section 3 provides the theoretical analysis, and Sect. 4 analyses the influence of different factors, including changes in data characteristics. Section 5 discusses the implication of the theoretical analysis on three nearest neighbour-based anomaly detectors. The empirical methodology and evaluation are given in Sects. 6 and 7, respectively. A discussion of related issues is provided in Sect. 8, followed by the conclusion in the last section.

2 Anomaly detectors that perform well using small training sets

This section summarises the anomaly detectors in the literature which are reported to produce a good performing model using small training sets.

Isolation Forest or iForest (Liu et al. 2008) was one of the first to report that a high performing anomaly detector could be trained using a training sample as small as 256 instances to train a model, in an ensemble of 100 models. On a dataset of over half a million instances, iForest could rank almost all anomalies at the top of the ranked list.

An information retrieval system called ReFeat (Zhou et al. 2012), which employed iForest as the ranking model, also exhibits the same behaviour. It requires a subsample of 8 instances only to build a model, in an ensemble of 1000 models to produce a high performing retrieval system. ReFeat is shown to outperform three state-of-the-art information retrieval systems.

Applying the isolation approach to nearest neighbour algorithms, LiNearN (Wells et al. 2014) and iNNE (Bandaragoda et al. 2014) have an explicit training process to build a

model to isolate every instance in a small training sample, where the local region which isolates an instance is defined using the distance to the instance's nearest neighbour.¹ Like iForest, this new nearest neighbour approach is shown to produce high performing models using small training sets in anomaly detection. iNNE could produce competitive anomaly detection accuracy using an ensemble of 1000 models, each trained using 2 instances only in a dataset of over half a million instances.

In an independent study, Sugiyama and Borgwardt (2013) have advocated the use of a nearest neighbour anomaly detector (kNN where $k = 1$) which employs a small sample, and showed that it performed competitively with LOF (Breunig et al. 2000) which employs the entire dataset. They provide a theoretical analysis which yields a lower error bound to explain the reason why a small dataset is sufficient to detect anomalies using a nearest neighbour anomaly detector. It reveals that most instances in a randomly selected subset are likely to be normal instances because the majority of instances in an anomaly detection dataset are normal instances. Finding the nearest neighbour from this subset and using the nearest neighbour distance as the anomaly score lead directly to good anomaly detection accuracy. A recent study shows that ensembles of INN can further improve the detection accuracy of of INN (Pang et al. 2015).

While our analysis and the analysis by Sugiyama and Borgwardt (2013) are based on the same nearest neighbour anomaly detector, the approaches to the theoretical analyses are different which yield different outcomes. Our computational geometry approach reveals that the learning curve behaviour is substantially influenced by the geometry of the areas covering normal instances and anomalies in the data space. In addition, our analysis yields both the lower and upper bounds of the anomaly detector's detection accuracy. In contrast, the analysis based on the probabilistic approach (Sugiyama and Borgwardt 2013) is limited to the lower bound of the probability of the perfect anomaly detection only. Moreover, our resultant bounds are represented by simple closed form expressions, while their result contains a complex probability distribution. Our result enables an interpretation of the learning curve behaviours that are influenced by different components of data characteristics. Most importantly, we show explicitly that the nearest neighbour anomaly detector has the gravity-defiant behaviour, and its detection accuracy is influenced by three factors: the proportion of normal instances (or anomaly contamination rate), the nearest neighbour distances of anomalies in the dataset, and sample size used by the anomaly detector where the geometry of normal instances and anomalies is best represented by the sample with the optimal data size.

3 Theoretical analysis

In this section, we characterise the learning curve of a nearest neighbour anomaly detector and reveal its gravity-defiant behaviour through a theoretical analysis based on computational geometry.

Measuring the detection accuracy of the anomaly detector using area under the receiver operating characteristic curve (AUC), we show that the lower and upper bounds of its performance have simple closed form expressions, and there are three factors which influence the AUC and explain the gravity-defiant behaviour.

¹ Apart from being an eager learner, further distinctions in comparison with the conventional k nearest neighbour learner are provided in Sect. 5.2.

3.1 Preliminary

Let $(\mathcal{M}, m)$ be a metric space, where $\mathcal{M}$ is a d dimensional space and m is a distance measure in $\mathcal{M}$. Let $\mathcal{X}$ be a d dimensional open subset of $\mathcal{M}$. $\mathcal{X}$ is split into a subset of normal instances $\mathcal{X}_N$ and a subset of anomalies $\mathcal{X}_A = \mathcal{X} \setminus \mathcal{X}_N$ by an oracle. Assume that each of $\mathcal{X}$, $\mathcal{X}_N$ and $\mathcal{X}_A$ can be partitioned into a finite number of convex d dimensional subsets. Further, assume that a probability density function $p(x)$ supported on $\mathcal{M}$ is finite and strictly positive in $\mathcal{X}$ and zero outside of $\mathcal{X}$, i.e.,

$$\exists p_\ell, p_u \in \mathcal{R}^+, \forall x \in \mathcal{X}, p_\ell < p(x) < p_u, \text{ and } \forall x \in \mathcal{M} \setminus \mathcal{X}, p(x) = 0. \quad (1)$$

These assumptions do not virtually limit applicability of this analysis, since any area where instances exist in $\mathcal{M}$ is practically approximated by a union of convex d dimensional subsets having $p(x) > 0$.

Let a dataset D be sampled independently and randomly with respect to $p(x)$. All instances in D are located within $\mathcal{X}$ according to Eq. (1). Further let D_N and D_A (in D) be sets of instances belonging to $\mathcal{X}_N$ and $\mathcal{X}_A$, respectively, i.e., $D_N = \{x \in D | x \in \mathcal{X}_N\}$ and $D_A = \{x \in D | x \in \mathcal{X}_A\}$. In anomaly detection, we assume that the size of D_A is substantially smaller than D_N , i.e., $|D_N| \gg |D_A|$.

Let $\mathcal{D}$ be a subsample set consisting of instances independently and randomly sampled from D , and let $\mathcal{D}_N$ and $\mathcal{D}_A$ be sets of normal instances and anomalies, respectively, in $\mathcal{D}$.

Note that the geometrical shapes and sizes of $\mathcal{X}$, $\mathcal{X}_N$ and $\mathcal{X}_A$ are independent of the data size of D , $\mathcal{D}$ and their subsets.

3.2 Definitions of anomaly detector and AUC

We employ a nearest neighbour anomaly detector (1NN) using $\mathcal{D}$ in this analysis; and it uses an anomaly score for any $x \in D$ defined by x 's nearest neighbour distance in $\mathcal{D}$ as follows:

$$q(x; \mathcal{D}) = \min_{y \in \mathcal{D}} m(x, y). \quad (2)$$

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 , where $r > 0$; otherwise x is judged to be a normal instance.

Using this decision rule, $y \in \mathcal{D}_N$ contributes to correctly judge an anomaly x , i.e., a true positive; while $y \in \mathcal{D}_A$ contributes to erroneously judge an anomaly x as normal, i.e., a false negative. In other words, any anomalies in $\mathcal{D}$ have the detrimental effect to reduce the true positive rate by increasing the number of false negatives. However, this effect is not very significant, because $\mathcal{D} \simeq \mathcal{D}_N$ holds from $|D_N| \gg |D_A|$.

Let the AUC (area under the receiver operating characteristic curve) of the anomaly detector using $\mathcal{D}$ be $AUC(\mathcal{D})$. From the description in the last paragraph, $AUC(\mathcal{D})$ is lower than but very close to the AUC of the anomaly detector using $\mathcal{D}_N$, i.e., $AUC(\mathcal{D}_N)$. Accordingly, we investigate $AUC(\mathcal{D}_N)$ in place of $AUC(\mathcal{D})$ for ease of analysis by assuming $AUC(\mathcal{D}) \simeq AUC(\mathcal{D}_N)$. This assumption is true to the extent of the probability of $\mathcal{D} \simeq \mathcal{D}_N$; and this probability is high when $|\mathcal{D}|$ is small as shown in Sugiyama and Borgwardt (2013).

The AUC of the anomaly detector using $\mathcal{D}_N$ is provided by the following expression² (Hand and Till 2001).

$$AUC(\mathcal{D}_N) = \int_0^{\infty} G(r; \mathcal{D}_N) f(r; \mathcal{D}_N) dr, \quad (3)$$

where $f(r; \mathcal{D}_N)$ is the probability density of true positive that an anomaly x with its anomaly score $q(x; \mathcal{D}_N) = r$ is correctly judged as anomaly; and $G(r; \mathcal{D}_N) = \int_0^r g(s; \mathcal{D}_N) ds$ is the cumulative distribution of false positive that a normal instance x with anomaly score $q(x; \mathcal{D}_N) = s$ is erroneously judged as anomaly. Because the decision rule $q(x; \mathcal{D}_N) \geq r$ for x is deterministic given $\mathcal{D}_N$ and r , $f(r; \mathcal{D}_N)$ is a summation of $p(x)$ for all $x \in \mathcal{X}_A$ under the condition $q(x; \mathcal{D}_N) = r$ as follows.

$$f(r; \mathcal{D}_N) = \int_{\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}} p(x) dx. \quad (4)$$

Similarly, $g(s; \mathcal{D}_N)$ and its cumulative distribution $G(r; \mathcal{D}_N)$ are represented by $p(x)$, $\mathcal{X}_N$ and $q(x; \mathcal{D}_N)$ as follows.

$$\begin{aligned} g(s; \mathcal{D}_N) &= \int_{\{x \in \mathcal{X}_N | q(x; \mathcal{D}_N) = s\}} p(x) dx, \text{ and} \\ G(r; \mathcal{D}_N) &= \int_0^r g(s; \mathcal{D}_N) ds. \end{aligned} \quad (5)$$

As pointed out by Hand and Till (2001), “the AUC is equivalent to the probability that a randomly chosen anomaly will have a smaller estimated probability of belonging to the normal class than a randomly chosen normal instance.”

3.3 Modeling $\mathcal{X}_N$ and $\mathcal{X}$ based on computational geometry

Here, we model $\mathcal{X}_N$ and $\mathcal{X}$ in $\mathcal{M}$ using computational geometry in relation to the anomaly score $q(x; \mathcal{D}_N)$, and connect this model to the AUC. The idea is to use balls of radius r to cover the geometry occupied by normal instances and anomalies. The AUC can then be computed through integration from 0 to r .

Let the set of all points satisfying $q(x; \mathcal{D}_N) \leq r$ in $\mathcal{M}$ be a union of balls $B_d(y, r)$ for all $y \in \mathcal{D}_N$, where $B_d(y, r)$ is a d dimensional ball centred at y with radius r .

$\mathcal{X}_N$ and $\mathcal{X}$ can now be modelled using these balls which have two critical radii, i.e., the inradius of $\mathcal{X}_N$: $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$; and the covering radius of $\mathcal{X}$: $\rho_u(\mathcal{D}_N, \mathcal{X})$, formally defined as follows:

$$\begin{aligned} \rho_\ell(\mathcal{D}_N, \mathcal{X}_N) &= \sup \arg_r \left[\bigcup_{y \in \mathcal{D}_N} B_d(y, r) \subseteq \mathcal{X}_N \right] \\ &= \sup \arg_r [\{x \in \mathcal{M} | q(x; \mathcal{D}_N) \leq r\} \subseteq \mathcal{X}_N], \text{ and} \\ \rho_u(\mathcal{D}_N, \mathcal{X}) &= \inf \arg_r \left[\bigcup_{y \in \mathcal{D}_N} B_d(y, r) \supseteq \mathcal{X} \right] \\ &= \inf \arg_r [\{x \in \mathcal{M} | q(x; \mathcal{D}_N) \leq r\} \supseteq \mathcal{X}]. \end{aligned}$$

² Note that some convention yields a different expression of AUC, i.e., $AUC(\mathcal{D}_N) = \int_0^1 F(r; \mathcal{D}_N) g(r; \mathcal{D}_N) dr$. This is because the convention has the y-axis and the x-axis reversed for the ROC plot. Both expressions give the same AUC. See Hand and Till (2001) for details.

Fig. 2 Examples of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$, where $\mathcal{X} = \mathcal{X}_N \cup \mathcal{X}_A$, and $\mathcal{D}_N$ is represented by points in $\mathcal{X}_N$. (a) $\mathcal{D}_N$ has one instance. (b) $\mathcal{D}_N$ has four instances

Figure 2 shows two examples of $\mathcal{X}_N$ and $\mathcal{X}$ being modelled using balls having the two radii.

The AUC, governed by $f(r; \mathcal{D}_N)$ and $G(r; \mathcal{D}_N)$ in Eq. (3), can now be modelled as follows. The set of all points satisfying $q(x; \mathcal{D}_N) = r$ in $\mathcal{M}$ can be modelled as a surface of a union of balls $B_d(y, r)$ for all $y \in \mathcal{D}_N$. Thus, $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$, used to determine $f(r; \mathcal{D}_N)$ in Eq. (4), is the intersection of the surface and $\mathcal{X}_A$. A similar modeling applies to $\{x \in \mathcal{X}_N | q(x; \mathcal{D}_N) = s\}$, used to determine $g(s; \mathcal{D}_N)$.

If r is between the two critical radii, the intersection $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$ is not an empty set and some anomalies in $\mathcal{X}_A$ are judged correctly, and thus $f(r; \mathcal{D}_N) > 0$. Otherwise, the intersection is an empty set and $f(r; \mathcal{D}_N) = 0$. This implies that the AUC is solely governed by $f(r; \mathcal{D}_N)$ and $G(r; \mathcal{D}_N)$ with $r \in [\rho_\ell(\mathcal{D}_N, \mathcal{X}_N), \rho_u(\mathcal{D}_N, \mathcal{X})]$. The theoretical results stated in the next subsection show that these two radii play a key role in characterising the AUC of the nearest neighbour anomaly detector.

3.4 Characterisation of the AUC based on computational geometry

Here, we characterise the AUC of the nearest neighbour anomaly detector. The lower and upper bounds of AUC are formulated through the inradius and the covering radius of the balls centred at every instance in the given sample. The lower bound is derived from an extreme case where all normal instances are concentrated in a small area. The upper bound is derived from a more general case where the geometry is not simple and thus requires a large number of balls to cover. The lemmas and theorem are given below.

Let ψ and h be the cardinalities of $\mathcal{D}$ and $\mathcal{D}_N$, respectively. $\psi \geq h \geq 0$ holds because of $\mathcal{D} \supseteq \mathcal{D}_N$. Further let $S_d(x, r)$ be the surface of the ball $B_d(x, r)$, and let B_d and S_d be the volume and the surface area of a d dimensional unit ball. The following lemmas and theorem provide bounds of $f(r; \mathcal{D}_N)$, $G(r; \mathcal{D}_N)$ and AUC with reference to $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$. Their proofs are provided in “Appendix 1”.

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

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

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 $\mathcal{X}_N$.

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.*

$$\begin{aligned} f(r; \mathcal{D}_N) &> p_\ell C_\delta \quad \text{if } \rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta \leq r \leq \rho_u(\mathcal{D}_N, \mathcal{X}) - \delta, \\ f(r; \mathcal{D}_N) &\geq 0 \quad \text{otherwise.} \end{aligned}$$

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.*

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

By combining these lemmas, we derive the following bounds of AUC for a given $\mathcal{D}_N$.

Lemma 5 *AUC($\mathcal{D}_N$) is upper and lower bounded as follows.*

$$\begin{aligned} \text{AUC}(\mathcal{D}_N) &< \frac{p_u^2 |\mathcal{X}_N| S_d}{d} h(\rho_u(\mathcal{D}_N, \mathcal{X})^d - \rho_\ell(\mathcal{D}_N, \mathcal{X}_N)^d), \\ \text{AUC}(\mathcal{D}_N) &> \frac{p_\ell^2 C_\ell C_\delta}{d+1} \{(\rho_u(\mathcal{D}_N, \mathcal{X}) - \delta)^{d+1} - (\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta)^{d+1}\}. \end{aligned}$$

We immediately obtain the following result by taking the expectation of the inequalities in Lemma 5.

Corollary 1 *Let the expectation of the b -th power of radius $\rho_*(\mathcal{D}_N, \mathcal{Y})^b$ over $p(\mathcal{D}_N|h)$ be*

$$\left\langle \rho_*(\mathcal{Y})^b \right\rangle_h = \int \rho_*(\mathcal{D}_N, \mathcal{Y})^b p(\mathcal{D}_N|h) d\mathcal{D}_N.$$

where $*$ = ℓ or u , b is an integer, and $\mathcal{Y} \subset \mathcal{M}$. $p(\mathcal{D}_N|h)$ is the probability distribution of $\mathcal{D}_N$ having its cardinality h .

Then, the expectation of AUC($\mathcal{D}_N$) over $p(\mathcal{D}_N|h)$, i.e., $\langle \text{AUC} \rangle_h$, is upper and lower bounded as

$$\begin{aligned} \langle \text{AUC} \rangle_h &< \frac{p_u^2 |\mathcal{X}_N| S_d}{d} h \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_h - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_h \right), \\ \langle \text{AUC} \rangle_h &> \frac{p_\ell^2 C_\ell C_\delta}{d+1} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_h - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_h \right\}. \end{aligned}$$

Let $P(\mathcal{Y}) = \int_{\mathcal{Y}} p(x) dx$. We define ratio α as

$$\alpha = \frac{P(\mathcal{X}_N)}{P(\mathcal{X}_N) + P(\mathcal{X}_A)} \approx \frac{|D_N|}{|D|} \approx \frac{|\mathcal{D}_N|}{|\mathcal{D}|} = \frac{h}{\psi}.$$

Since $|D_N| \gg |D_A|$ is assumed, α is less than but sufficiently close to 1. Then, we obtain the following theorem.

Theorem 1 *The expectation of $\langle \text{AUC} \rangle_h$ over the distribution of h ($P(h)$), i.e.,*

$\langle \text{AUC} \rangle = \sum_{h=0}^{\psi} P(h) \langle \text{AUC} \rangle_h$, *has the following upper and lower bounds:*

$$\begin{aligned} \langle \text{AUC} \rangle &< \frac{p_u^2 |\mathcal{X}_N| S_d}{d} \psi \alpha^\psi \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_\psi - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_\psi \right) + O((1-\alpha)\psi^2), \text{ and} \\ \langle \text{AUC} \rangle &> \frac{p_\ell^2 C_\ell C_\delta}{d+1} \alpha^\psi \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_\psi - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_\psi \right\} + O((1-\alpha)\psi). \end{aligned}$$

As $(1 - \alpha)$ is sufficiently close to 0, both $O((1 - \alpha)\psi^2)$ and $O((1 - \alpha)\psi)$ can be ignored in the upper and lower bounds, respectively. By denoting $\rho_\delta^b(\psi) = \langle (\rho_u(\mathcal{X}) - \delta)^b \rangle_\psi - \langle (\rho_\ell(\mathcal{X}_N) + \delta)^b \rangle_\psi$, the upper and lower bounds can be expressed in the forms: $C_U \psi \alpha^\psi \rho_\delta^b(\psi)$ and $C_L \alpha^\psi \rho_\delta^b(\psi)$, respectively, where $b = d$ and $\delta = 0$ for the upper bound, and $b = d + 1$ and $0 < \delta < \delta_0$ for the lower bound; C_U and C_L are constants.

In plain language, the three factors can be interpreted as follows: $1 - \alpha^\psi$ reflects the likelihood of anomaly contamination in the subsample set $\mathcal{D}$; ψ represents the number of balls used to represent the geometry of normal instances and anomalies; and $\rho_\delta^b(\psi)$ signifies the separation between anomalies and normal instances, represented by ψ balls.

3.5 Gravity-defiant behaviour

As revealed in the last section, the upper bound of AUC has two critical terms ψ and α^ψ which are monotonic functions changing in opposite directions, i.e., as ψ increases, α^ψ decreases. Therefore, the AUC bounded by $\psi \alpha^\psi$ is expected to reach the optimal at some finite and positive ψ_{opt} ; and the anomaly detector will perform worse if the sample size used is larger than ψ_{opt} , i.e., the gravity-defiant behaviour.

Figure 3 shows two examples of the gravity-defiant behaviour, as a result of the upper bound. They are represented by the $\psi \alpha^\psi$ curves for $\alpha = 0.9$ and 0.99 . This shows that the anomaly detector using anomaly score $q(x; \mathcal{D})$ has the gravity-defiant behaviour, where $\psi_{opt} < \psi_{max}$; and ψ_{max} is either the size of the given dataset or the largest sample size that can be employed. The gravity-defiant behaviour is in contrary to the conventional wisdom.

In addition, $\rho_u(\mathcal{D}_N, \mathcal{X})$ and $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ decrease when ψ increases, and $\rho_u(\mathcal{D}_N, \mathcal{X})$ is usually substantially larger than $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ as depicted in Fig. 2. These behaviours are seen in many other examples shown later. Accordingly, $\langle \rho_u(\mathcal{X})^b \rangle_\psi$ dominates $\langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$ in most cases where $b > 1$. Thus, like α^ψ , the term $\rho_\delta^b(\psi)$ decreases as ψ increases. This fact indicates that $\rho_\delta^b(\psi)$ is positive, smooth and anti-monotonic over the change of ψ .

Thus, both α^ψ and $\rho_\delta^b(\psi)$ decrease as ψ increases and wield the similar influence on the nearest neighbour anomaly detector to exhibit the gravity-defiant behaviour.

The lower bound denotes that AUC always decreases as ψ increases. The actual behaviour of an INN anomaly detector follows this lower bound when $\mathcal{X}_N$ is concentrated on a small area, as in the case of sharp Gaussian distribution, which yields $\psi_{opt} = 1$. This is evident

Fig. 3 The $\psi \alpha^\psi$ curves as a function of ψ with (a) $\alpha = 0.9$ and (b) $\alpha = 0.99$. The left and right y-axes are for α^ψ and ψ curves, respectively. Note that the y-axis scale of the $\psi \alpha^\psi$ curves is not shown. (a) $\alpha = 0.9$: $\psi \alpha^\psi$ has $\psi_{opt} = 10$. (b) $\alpha = 0.99$: $\psi \alpha^\psi$ has $\psi_{opt} = 100$

from the proofs of Lemmas 3-5 in “Appendix 1” where the lower bound comes from the data points in a small part of $\mathcal{X}_N$.

As the lower bound is limited to special cases only, we will discuss in more details about the effect on AUC, in terms of the upper bound, due to changes in different data characteristics in the next section.

4 Analysis of factors which influence the AUC of the nearest neighbour anomaly detector

The theoretical analysis in Sect. 3 reveals that there are three factors which influence the AUC of the nearest neighbour anomaly detector:

- (a) The proportion of normal instances, α : According to the argument in the last section, the nearest neighbour anomaly detector is expected to improve its AUC by the rate α^ψ as the proportion of normal instances increases. The change in α does not affect the other two factors, if the change does not affect the geometry of $\mathcal{X}_N$ and $\mathcal{X}$. In addition to the effect on the magnitude of AUC, Fig. 3 also shows that ψ_{opt} becomes larger as α increases. This is because α^ψ increases as α increases.
- (b) The difference between the covering radius of $\mathcal{X}$ and the inradius of $\mathcal{X}_N$, i.e., $\rho_\delta^b(\psi) = \langle \rho_u(\mathcal{X})^b \rangle_\psi - \langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$: This factor depends on the geometry of normal clusters as well as anomalies, and influences the AUC in the following scenarios:
 - (b1) $\mathcal{X}_A$ becomes bigger. The change in $\mathcal{X}_A$ with fixed $\mathcal{X}_N$ directly affects $\mathcal{X}$. Examples of this change from Fig. 2 are shown in Fig. 4. The enlarged $\mathcal{X}_A$, thus the enlarged $\mathcal{X}$, leads to larger $\rho_\delta^b(\psi)$ and higher AUC because the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ gets larger while the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ is fixed for a given ψ .
 - (b2) $\mathcal{X}_N$ becomes bigger. The change in $\mathcal{X}_N$ with fixed $\mathcal{X}$ affects both the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$. Examples of this change from Fig. 2 are depicted in Fig. 5. The enlarged $\mathcal{X}_N$ leads to smaller $\rho_\delta^b(\psi)$ and thus lower AUC because the difference between the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ gets smaller—a result of the increased $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and the decreased $\rho_u(\mathcal{D}_N, \mathcal{X})$ for a given ψ .
 - (b3) Number of clusters in $\mathcal{X}_N$ increases. If $\mathcal{X}_N$ consists of multiple well-separated clusters as shown in Fig. 6, then $\langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$ is determined by the minimum of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ of single clusters, regardless of the total volume or the number of

Fig. 4 $\mathcal{X}_A$ is larger than that in Fig. 2. Examples of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$. (a) $\mathcal{D}_N$ has one instance. (b) $\mathcal{D}_N$ has four instances

Fig. 5 $\mathcal{X}_N$ and $\mathcal{X}$ have approximately the same size. Examples of $\rho_u(\mathcal{D}_N, \mathcal{X})$ and $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$. (a) $\mathcal{D}_N$ has one instance. (b) $\mathcal{D}_N$ has four instances

Fig. 6 $\mathcal{X}_N$ consists of multiple clusters. Examples of $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$ and $\rho_u(\mathcal{D}_N, \mathcal{X})$. (a) $\mathcal{D}_N$ has one instance. (b) $\mathcal{D}_N$ has seven instances

clusters in $\mathcal{X}_N$. This is despite the fact that the total volume of $\mathcal{X}_N$ has increased from that in Fig. 2. The expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ in Fig. 6 is less than that in Fig. 2 because the clusters are scattered in $\mathcal{X}$. With fixed ψ , the AUC is expected to decrease in Fig. 6 in comparison with that in Fig. 2 because of the decreased $\langle \rho_u(\mathcal{X})^b \rangle_\psi$, which is obvious in the change from Fig. 2b, 3, 4, 5 and 6b.

Anomalies' nearest neighbour distances As indicated in (b1) and (b2), enlarging $\mathcal{X}_N$ has the same effect of shrinking $\mathcal{X}$ in decreasing AUC. It is instructive to note that either of these two changes effectively reduces the anomalies' nearest neighbour distances because the area occupied by $\mathcal{X}_A$ decreases. This can be seen from $\rho_\delta^b(\psi) = \langle \rho_u(\mathcal{X})^b \rangle_\psi - \langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$, where $\rho_\delta^b(\psi)$ changes in the same direction of the expected $\rho_u(\mathcal{D}_N, \mathcal{X})$ or in the opposite direction of the expected $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$. The nearest neighbour distance of anomaly,³ which can be measured easily, is a proxy to $\rho_\delta^b(\psi)$.

In a nutshell, any changes in $\mathcal{X}$ and $\mathcal{X}_N$ that matter—which finally vary AUC—are manifested as changes in the anomalies' nearest neighbour distances (Δ_A). The AUC, $\langle \rho_u(\mathcal{X})^b \rangle_\psi - \langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$ and Δ_A change in the same direction.

³ A more accurate proxy is the distance between anomaly and its nearest normal instance. In the unsupervised learning context, this distance cannot be measured easily. We will see in the experiment section that the nearest neighbour distance of anomaly is a good proxy to $\rho_\delta^b(\psi)$, even in a dataset with clustered anomalies—a factor not considered in the analysis.

Table 1 Changes in AUC and ψ_{opt} as one data characteristic (α , $\mathcal{X}$ or $\mathcal{X}_N$) 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	=	=	=	↑	↑
(b1) $\mathcal{X}_A$ becomes bigger	↑	=	↑	↑	*
(b2) $\mathcal{X}_N$ becomes bigger	↓	↑	↓	↓	*
(b3) Number of clusters in $\mathcal{X}_N$ increases	↓	=	↓	↓	*

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

Note that $\rho_\delta^b(\psi)$ has the same effect of α^ψ to shift ψ_{opt} . But the influence of $\rho_\delta^b(\psi)$ is more difficult to predict because it depends on the rate of decrease between the covering radius of $\mathcal{X}$ and inradius of $\mathcal{X}_N$ which in turn depend on the geometry of $\mathcal{X}$ and $\mathcal{X}_N$; and it is hard to measure in practice too.

- (c) The sample size (ψ) used by the anomaly detector : The optimal sample size is the number of instances best represents the geometry of normal instances and anomalies ($\mathcal{X}_N$ and $\mathcal{X}$). The sample size also affects two other factors, i.e., as ψ increases, both α^ψ and $\rho_\delta^b(\psi)$ decrease. The direction of the change in AUC depends on the interaction between ψ and $\alpha^\psi \rho_\delta^b(\psi)$ which change in opposite directions. In general, as ψ increases from a small value, AUC improves until it reaches the optimal. Further increase from ψ_{opt} degrades AUC which gives rise to the gravity-defiant behaviour. Note that the change in ψ does not alter the data characteristics (i.e., α , $\mathcal{X}_N$ and $\mathcal{X}$).

Changes in the first two factors, which affect the data characteristics, are summarised in Table 1. Each effect shown is a result of an isolated factor. However, a change in one factor often affects one or more other factors. We will examine some interactions between these factors in the empirical evaluation in Sect. 7.

While one can expect the nearest neighbour anomaly detector to exhibit gravity-defiant learning curves, there are two scenarios in which only half of the curve can be observed.

- First half of the curve: For a dataset which requires large ψ_{opt} , the dataset size needs to be very large in order to observe the gravity-defiant behaviour. In the case that the data collected is not large enough, ψ_{opt} may not be achievable in practice.
- Second half of the curve: This is observed on a dataset which requires small ψ_{opt} e.g., $\psi_{opt} = 1$.

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. First, increasing sample size increases the likelihood of anomaly contamination in the sample; and any inclusion of anomalies in the sample increases the false negative rate, thus, lowers the AUC. Second, the optimal sample size depends on the data distribution. As long as the data distribution is not sufficiently represented by the current sample, increasing the sample size will improve AUC. The optimal size is the number of instances best represents the geometry of normal instances and anomalies; this gives the optimal separation between normal instances and anomalies, encapsulated as the average nearest neighbour distance to anomalies. Third, increasing the sample size decreases the average nearest neighbour distance to anomalies. Increasing beyond the optimal sample size reduces the separation between normal instances and anomalies smaller than the optimal. This leads to the decreased AUC and gives rise to the gravity-defiant behaviour.

The above impacts are due to the change in sample size which, by itself alone, does not alter anomaly contamination rate or geometry of normal instances and anomalies in the given dataset [described in (a) and (b) above]. Any change in geometrical data characteristics, which affects the AUC, manifests as a change in nearest neighbour distance such that the AUC and anomalies' nearest neighbour distances change in the same direction. Because nearest neighbour distance can be measured easily and other indicators of detection accuracy are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams.

5 Does the theoretical result apply to other nearest neighbour-based anomaly detectors?

The above theoretical analysis is based on the simplest nearest neighbour (1NN) anomaly detector with a small sample. We believe that this result applies to other nearest neighbour-based anomaly detectors too, although a direct analysis is not straightforward in some cases.

We provide our reasoning as to why the theoretical result can be applied to three nearest neighbour-based anomaly detectors, i.e., an ensemble of nearest neighbours, a recent nearest neighbour-based ensemble method called iNNE (Bandaragoda et al. 2014), and k-nearest neighbour. These are provided in the following three subsections.

5.1 The effect of ensemble

Building an ensemble of nearest neighbour anomaly detectors is straightforward. We name such an ensemble aNNE (a Nearest Neighbour Ensemble). It computes the anomaly score for $x \in D$ by averaging its nearest neighbour distance between x and $\mathcal{D}_i \subset D$, $i = \{1, 2, \dots, t\}$ as follows.

$$\bar{q}(x; D) = \frac{1}{t} \sum_{i=1}^t q(x; \mathcal{D}_i) = \frac{1}{t} \sum_{i=1}^t \min_{y \in \mathcal{D}_i} m(x, y).$$

When the size of the original dataset D is significantly larger than the subsample size ψ , the subsample sets $\mathcal{D}_i$ ($i = 1, \dots, t$) are almost mutually i.i.d. Thus, the anomaly score obtained from each subsample set; $q(x; \mathcal{D}_i)$ is also i.i.d. As well known and shown in LiNearN (Wells et al. 2014), such an ensemble operation reduces the expected variance of the anomaly score $\bar{q}(x; D)$ by the factor $1/t$ while maintaining its expected mean. Approximately the same rate of reduction applies to the variances of $\langle \rho_u(\mathcal{X})^b \rangle_\psi$, $\langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$ and hence $\langle AUC \rangle$.

A significant advantage of this technique is that the effect of the ensemble size t is almost solely limited to the variance of the estimation only. In other words, we can easily control the variance by choosing some appropriate t with almost no other side effects.

Thus, the theoretical analysis also applies to aNNE, and it is expected to have the gravity-defiant behaviour.

5.2 iNNE is a variant of aNNE

Like aNNE, a recent nearest neighbour-based method iNNE (Bandaragoda et al. 2014) employs small samples and aggregates the nearest neighbour distances from all samples to compute the score for each test instance. iNNE is a variant of aNNE because iNNE can be

Table 2 Conceptual comparison of aNNE and iNNE. The first column indicates the nearest neighbour of x , a variant of nearest neighbour $cnn(x)$, and the nearest neighbour distance for a given $\mathcal{D}$

	aNNE	iNNE
η_x	$\arg \min_{y \in \mathcal{D}} m(x, y)$	—
$cnn(x)$	—	$\arg \min_{c \in \mathcal{D}} \{\tau(c) : x \in B(c, \tau(c))\}$
$q(x; \mathcal{D})$	$m(x, \eta_x)$	$\tau(cnn(x))$

interpreted as identifying anomalies as instances which have the furthest nearest neighbours, as in aNNE. The reasoning is given in the following paragraphs.

Conceptually, iNNE isolates every instance c in a random sample $\mathcal{D}$ by building a hypersphere centred at c with radius $\tau(c)$ in the training process. The hypersphere is defined as follows:

Let hypersphere $B(c, \tau(c))$, centred at c with radius $\tau(c)$, be $\{x : m(x, c) < \tau(c)\}$, where $\tau(c) = \min_{y \in \mathcal{D} \setminus \{c\}} m(y, c)$, $x \in \mathcal{M}$ and $c \in \mathcal{D}$.

To score $x \in \mathcal{M}$, the centre of the smallest hypersphere that covers x is defined as:

$$cnn(x) = \arg \min_{c \in \mathcal{D}} \{\tau(c) : x \in B(c, \tau(c))\}$$

In contrast, for aNNE, the nearest neighbour of $x \in \mathcal{M}$ is defined as:

$$\eta_x = \arg \min_{y \in \mathcal{D}} m(x, y)$$

where $x \neq \eta_x$.

$cnn(x)$ can be viewed as a variant of nearest neighbour of x because $cnn(x) = \eta_x$, except in two conditions: (i) $x \in B(cnn(x), \tau(cnn(x)))$, but $x \notin B(\eta_x, \tau(\eta_x))$ when $\tau(cnn(x)) \geq \tau(\eta_x)$; and (ii) $cnn(x)$ could be *nil* or undefined when x is not covered by any hypersphere $\forall c \in \mathcal{D}$.

The anomaly score for iNNE,⁴ $q(x; \mathcal{D})$, is simply defined by $\tau(cnn(x))$.

Anomalies identified by aNNE are the instances in $\mathcal{D}$ which have the longest distance to η_x in $\mathcal{D}$. Similarly, anomalies identified by iNNE are the instances in $\mathcal{D}$ which have the longest distance to $cnn(\cdot)$, i.e., their (variant of) nearest neighbours in $\mathcal{D}$. Viewed from another perspective, anomalies identified by iNNE are those covered by the largest hyperspheres.

The conceptual comparison of aNNE and iNNE is summarised in Table 2.

For ease of reference, the algorithms for aNNE and iNNE are provided in “Appendix 2”. Note that the key difference between aNNE and iNNE in Algorithms 1 and 2 are in steps 1 and 4 only: The construction of hyperspheres in training and the version of nearest neighbour employed in evaluation.

As a consequence of the similarity between aNNE and iNNE, we expect that iNNE also has the gravity-defiant behaviour.

5.3 Extension to kNN

The extension of our analysis to the anomaly detector using the k nearest neighbour (kNN) distance is not straightforward. This is because the geometrical shape, cast from kNN distance, cannot be simply characterised by the inradius and the covering radius of the data space. In

⁴ iNNE’s original score (Bandaragoda et al. 2014) is a relative measure. We employ the base measure of the relative measure to point out that the basic algorithm has a lot in common with aNNE.

addition, the optimal k is a monotonic function of the data size (see the discussion on bias-variance analysis in Sect. 8.2).

Nevertheless, 1NN and kNN have the identical algorithmic procedure, except a small operational difference in the decision function—using one or k nearest neighbours. Thus, we can expect both 1NN and kNN to have the same gravity-defiant behaviour. We show that this is the case when $k = \sqrt{n}$ (the rule suggested by Silverman 1986).

Section summary

aNNE can be expected to behave similarly as 1NN, but with a lower variance; and the size of the variance reduction is proportional to the ensemble size. Thus, the result of the theoretical analysis applies directly to aNNE. We shall refer to aNNE rather than 1NN hereafter, both in our discussion and empirical evaluation, because aNNE has a lower variance than 1NN.

The analyses on kNN and iNNE are not a straightforward extension of the analysis on aNNE, and the optimal k for kNN depends on data size. Given that they all based on the basic operation: nearest neighbour, we can expect iNNE, aNNE and kNN to have the same behaviour in terms of learning curve.

In a nutshell, all three algorithms, aNNE, kNN and iNNE, can be expected to have the gravity-defiant behaviour. However, at what sample size (ψ_{opt}) each will arrive at its optimal detection accuracy is of great importance in choosing the algorithm to use in practice. We will investigate this issue empirically in Sect. 7.

6 Experimental methodology

Algorithms used in the experiments are aNNE, iNNE and kNN (where the anomaly score is computed from the average distance of k nearest neighbours (Bay and Schwabacher 2003)).

The experiments are designed to:

- I. Verify that each of aNNE, iNNE and kNN has the gravity-defiant behaviour.
- II. Compare ψ_{opt} of these algorithms.
- III. Attest the effect of each of the three factors on the detection accuracy, as revealed by the theoretical analyses in Sects. 3 and 4.

The performance measure is anomaly detection error, measured as $1 - \text{AUC}$, where AUC is the area under the receiver operating characteristics curve which measures the ‘goodness’ of the ranking result. Error = 0 if an anomaly detector ranks all anomalies at the top; and error = 1 if all anomalies are ranked at the bottom; a random ranker will produce error = 0.5.

A learning curve is produced for each anomaly detector on each dataset.

aNNE and iNNE have two parameters: sample size ψ and ensemble size t . To produce a learning curve, the training data is constructed using a sample of size $t\psi$ where $t = 100$ and ψ is 1, 2, 5, 10, 20, 35, 50, 75, 100, 150, 200, 500 and 1000 for each point on the curve. The parameter k in kNN was set as $k = \lfloor \sqrt{n} \rfloor$ (Silverman 1986) [chapter 1]⁵ where n is the number of training instances ($n = t\psi$). Note that the minimum ψ setting is 1 for aNNE and kNN; but because each hypersphere iNNE built derives its size from the hypersphere’s centre to its nearest neighbour, it requires a minimum of two instances in each sample.

For an ensemble of t models, the total number of training instances employed is $t\psi$. To train a single model such as kNN, a training set of $t\psi$ instances is used in order to ensure that a fair comparison is made between an ensemble and a single model.

⁵ Recent references to the use of this rule can be found in Zitzler et al. (2004) and Pandya et al. (2013).

Table 3 Datasets used in the experiments

Data	Size n	d	anomaly class
CoverType	286,048	10	class 4 (0.9 %) vs. class 2
Mulcross	262,144	4	1 % anomalies
Smtp	95,156	3	attack (0.03 %)
U2R	60,821	34	attack (0.37 %)
P53Mutant	31,159	5408	active (0.5 %) vs. inactive
Mammography	11,183	6	class1 (2.32 %)
Har	4728	561	sitting, standing & laying (1.2 %)
ALOI	100,000	64	0.553 % anomalies with 900 normal clusters

The Euclidean distance is used in all three algorithms.

A total of eight datasets are used in the experiment. Six datasets are from the UCI Machine Learning Repository (Lichman 2013), one dataset is produced from the Mulcross data generator,⁶ and the ALOI dataset is from the MultiView dataset collection.⁷ They are chosen because they represent different data characteristics of data size, number of dimensions, and proportion of normal instances and anomalies. Each dataset is normalised using the min-max normalisation. The data characteristics of these datasets are given in Table 3.

The ALOI dataset has $C = 900$ normal clusters and 100 anomaly clusters, where each anomaly cluster has between 1 and 10 instances. It is used in Sects. 7.3–7.5 because it allows us to easily change $\mathcal{X}_N$ or $\mathcal{X}$ to examine the resultant effect on AUC as predicted by the theoretical analysis. Sections 7.1, 7.4 and 7.5 employ the ALOI $C = 10$ dataset where the ten normal clusters are randomly selected from the 900 clusters.

In every experiment, each dataset is randomly split into two equal-size stratified subsets, where one is used for training and the other for testing. For example, in each trial, the CoverType dataset is randomly split into two subsets, each has 143,024 instances. The subset, which is used to produce training instances, is sampled without replacement to obtain the required t samples, each having ψ instances. As $t = 100$ and the maximum ψ is 1000, the maximum training instances employed is 100,000. For datasets which have less than 200,000 instances, the sampling without replacement process is restarted with the same subset when the instances have run out.

The result in each dataset is obtained from an average over 20 trials. Each trial employs a training set to train an anomaly detector and its detection performance is measured using a testing set.

As nearest neighbour distance is an important indicator of the detection error of the nearest neighbour-based anomaly detectors, we produce a bar-chart showing the average nearest neighbour distances of two groups: normal and anomaly, on a dataset created for each trial but before the data is split into the training and testing subsets. The reported result is averaged over 20 trials, and it is used in Sects. 7.2–7.5. Note that this is $q(x; D)$, unlike $q(x; \mathcal{D})$ used by aNNE. It allows us to measure the nearest neighbour distance for a given dataset, independent of the anomaly detector used. We use Δ to denote $q(x; D)$ hereafter.

⁶ <http://lib.stat.cmu.edu/jasasoftware/rocke>.

⁷ <http://elki.dbs.ifi.lmu.de/wiki/DataSets/MultiView>. Accessed: 11, November 2014.

Table 4 Changes in error(=1-AUC), ψ_{opt} and Δ_A as data characteristics (α , $\mathcal{X}$, $\mathcal{X}_N$) 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	=	$\downarrow$	$\uparrow$	$\diamond$	$\uparrow$	$\uparrow$
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	$\uparrow$	=	$\cong$	$\cong$	$\downarrow$	$\uparrow$

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

7 Empirical evaluation

The empirical evaluation is divided into five subsections. The first subsection investigates whether the three nearest neighbour anomaly detectors are gravity-defiant algorithms.

The other four subsections investigate the influence of three factors identified in the theoretical analysis. The experiments are designed by making specific changes in data characteristics in order to observe the resultant changes in detection error and ψ_{opt} , as predicted by the theoretical analysis. Table 4 shows the changes made to a dataset in each subsection in terms of α , $\mathcal{X}$, $\mathcal{X}_N$.

Table 4 also summarises the experimental outcomes in terms of changes in error and ψ_{opt} in each subsection. The nearest neighbour distance of anomaly (Δ_A) is included because it is the single most important indicator of change in data characteristics (due to $\mathcal{X}_N$ or $\mathcal{X}$) and error.

Recall that, based on the theoretical analysis in Sect. 3.5, the error and $\rho_\delta^b(\psi)\alpha^\psi$ change in opposite directions. The theoretical analysis correctly predicts the error outcomes of all six cases shown in Table 4.

The details of the experiments are given in the following five subsections.

7.1 Gravity-defiant behaviour

We investigate whether iNNE, aNNE and kNN have the gravity-defiant behaviour using the eight datasets in this section. The learning curves for iNNE, aNNE and kNN on each dataset are shown in Fig. 7.

Table 5 summarises the result in Fig. 7 by showing the optimal ψ_{opt} for iNNE, aNNE and kNN⁸ in each dataset. Recall that $\psi_{opt} < \psi_{max}$ shows the gravity-defiant behaviour. As we have used ψ up to 1000 in the experiment, $\psi_{opt} = \psi_{max} = 1000$ shows the gravity-compliant behaviour. All three anomaly detectors exhibit the gravity-defiant behaviour on all datasets, except the Smtpt dataset.

One interesting result in Table 5 is that ψ_{opt} for iNNE is significantly smaller than those for aNNE and kNN on all datasets. The only exception is the Smtpt dataset. Figure 8 shows that the geometric means of ψ_{opt} and error at ψ_{opt} relative to iNNE over the eight datasets. This result shows that aNNE and kNN require about 5 and 4 t times ψ_{opt} of iNNE, respectively, in order to achieve the optimal detection performance. While all three algorithms have about the same optimal detection performance overall, iNNE has the best on four datasets, equal on two and worse performance than aNNE and kNN on two.

⁸ Note that the actual optimal training set size for kNN is $t\psi_{opt}$.

Fig. 7 Learning curves of iNNE, kNN and aNNE [Sect. 7.1]: Error is defined as 1-AUC. kNN uses training set size of $t\psi$ and $k = \sqrt{t\psi}$; aNNE and iNNE are using $t = 100$. The results of the other four datasets are shown in Fig. 15 in “Appendix 4”. (a) CoverType, (b) Smtsp, (c) P53Mutant, (d) Har

Table 5 ψ_{opt} for iNNE, aNNE and kNN, where $\psi_{max} = 1000$ and $t = 100$

Data	iNNE	aNNE	kNN
CoverType	35	200	$200t$
Mulcross	2	10	$5t$
Smtsp	1000	1000	$1000t$
U2R	20	200	$100t$
P53Mutant	2	20	$75t$
Mammograhpy	200	500	$500t$
Har	2	50	$5t$
ALOI $C = 10$	150	200	$200t$

Another interesting observation in Fig. 7 is that the learning curves of iNNE almost always have steeper gradient than those of aNNE and kNN.

We will investigate in the next section the reason why the Smtsp dataset does not allow the algorithms to exhibit the gravity-defiant behaviour.

7.2 Enlarge $\mathcal{X}$ by increasing anomalies’ distances to normal instances

The Smtsp dataset has $\psi_{opt} = \psi_{max}$ in the previous experiment. By examining the dataset, we have found out that all the anomalies are very close to normal clusters. Thus, it is an ideal

Fig. 8 Geometric mean of ψ_{opt} and error at ψ_{opt} relative to iNNE over eight datasets. (a) ψ_{opt} relative to iNNE. (b) Error at ψ_{opt} relative to iNNE

Fig. 9 Changing distance offset [Sect. 7.2]. (a) Learning curves of iNNE for the different distance offset values on the anomalies of the Smt dataset. The offset values are for the transposition of the anomaly points from the original position on the first two dimensions only (i.e., the third dimension is not altered). Note that the top line (blue) is using the right y-axis, whereas the bottom three lines (red) are using the left y-axis. (b) The average nearest neighbour distance and ψ_{opt} for each of the distance offset value. The histogram is using the left y-axis (Avg NN Dist, Δ), whereas the line graph is using the right y-axis (ψ_{opt}). The learning curves of aNNE and kNN can be found in Fig. 16 in “Appendix 4”. (a) iNNE, (b) summarised results (Color figure online)

dataset to examine the effect of an enlarged $\mathcal{X}$ by increasing the distance between anomalies and normal clusters. We offset all anomalies by a fixed distance diagonally in the first two dimensions (but the third dimension is unchanged.) The offsets used are 0.0, 0.03, 0.075, 0.2. An example offset is shown in “Appendix 3”.

As everything else stays the same, the offset enlarges $\mathcal{X}$ without changing $\mathcal{X}_N$. The theoretical analysis suggests that this will lower the error.

The results are shown in Fig. 9. It is interesting to note that in all three algorithms, the error decreases as the offset increases. This is manifested as the increase in anomalies’ nearest neighbour distances (Δ_A) as shown in Fig. 9b.

In addition, ψ_{opt} of iNNE decreases (from 1000 to 150)—the gravity-defiant behaviour now prevails. This is despite the fact that both aNNE and kNN still have $\psi_{opt} = \psi_{max}$ as the offset increases.⁹ This phenomenon is consistent with the result in the previous section that iNNE has significantly smaller ψ_{opt} than those of aNNE and kNN.

⁹ Note that the trend of the learning curves for both aNNE and kNN (shown in “Appendix 4”) is similar to that for iNNE—error decreases as the distance offset increases. Because ψ_{opt} of iNNE decreases, ψ_{opt} ’s of aNNE and kNN are expected to follow the same trend although they cannot be determined because aNNE and kNN require significantly larger data size in order to reach their optimal performances.

Fig. 10 Changing $\mathcal{X}_N$ [Sect. 7.3(i)] different complexities of ten normal clusters. **(a)** Learning curves of iNNE for the different complexities of the ALOI $C = 10$ dataset. **(b)** The average nearest neighbour distance and ψ_{opt} . The learning curves of aNNE and kNN can be found in Fig. 17 in “Appendix 4”. **(a)** iNNE, **(b)** summarised results

This experiment verifies the analysis in Sect. 4 that a dataset which exhibits the first half of the learning curve has large ψ_{opt} ; and enlarging $\mathcal{X}_A$ increases Δ_A , as predicted in row b1 in Table 1. The offset apparently decreases ψ_{opt} —enables the gravity-defiant learning curve to be observed.

7.3 Changing $\mathcal{X}_N$ by changing normal clusters only

In this section, we conduct two experiments to examine the effect of changing $\mathcal{X}_N$. In the first experiment, we use a fixed number of normal clusters of increasing complexities which increases the volume of $\mathcal{X}_N$. In the second experiment, we increase the number of normal clusters where the increased diffusion of clusters that makes up $\mathcal{X}_N$ has a more important influence than the total volume. In both experiments, the number of anomalies is kept unchanged.

The ALOI dataset, which has 900 (C) normal clusters, is employed in these two experiments because it has the required characteristics to make the two changes; and they are done as follows:

- (i) ALOI(10): We use three categories of ten normal clusters ($C = 10$) from the ALOI dataset. The low, medium and high categories indicate the three different complexities of single normal clusters based on ψ_{opt} of aNNE.¹⁰ The results are shown in Fig. 10.
- (ii) ALOI(C): We increase the number of normal clusters, i.e., $C = 10, 100, 500, 900$, which indicate that high C has high diffusion of clusters. The resultant subsets of the ALOI dataset used in the experiments are shown in Table 6. The results are shown in Fig. 11.

The results, shown in Figs. 10a and 11a, reveal that increasing either the complexity of normal clusters or the number of normal clusters increases the error for all three algorithms. Yet, the sources that lead to this apparently same outcome are different which are manifested as the decrease in anomalies’ nearest neighbour distances (Δ_A) in case (ii); and the increase in normal instances’ nearest neighbour distances (Δ_N), though Δ_A also decreases to a lesser degree, in case (i). These results are shown in Figs. 10b and 11b.

¹⁰ 900 ALOI(1) single normal cluster datasets are formed by combining each of the 900 normal clusters with all anomalies. aNNE is applied to each dataset to find its ψ_{opt} . The datasets of ALOI(1) are then grouped based on aNNE’s ψ_{opt} ($=1, 2, 5, 10, 20, 35, 50, 75, 100, 200, 500, 1000$), i.e., which have different data characteristics/complexities manifested as learning curves with different ψ_{opt} values. Each dataset of ALOI(10), used in the experiment, has 10 normal clusters which are randomly selected from one of the following three categories: low has $\psi_{opt} = 1, 2, 5, 10$; medium has $\psi_{opt} = 75$; and high has $\psi_{opt} = 200, 500, 1000$.

Table 6 Subsets of the ALOI dataset with $C = 10, 100, 500, 900$

C	Number of instances			α
	Normal	Anomaly	Total	
10	1104	553	1657	0.67
100	11,047	553	11,600	0.95
500	55,241	553	55,794	0.990
900	99,447	553	100,000	0.994

Fig. 11 Changing $\mathcal{X}_N$ [Sect. 7.3(ii)] changing the number of normal clusters. **(a)** Learning curves of iNNE for the ALOI dataset ($C = 10, 100, 500, 900$). **(b)** The average nearest neighbour distance and ψ_{opt} . The learning curves of aNNE and kNN can be found in Fig. 18 in “Appendix 4”. Note that the line graphs in **(b)** stop at low C values because ψ_{opt} is beyond $\psi_{max} = 10000$ for high C values. **(a)** iNNE, **(b)** summarised results

Note that α is unchanged in case (i). But in case (ii), α of ALOI(C) increases as C increases which has the effect of decreasing error, if nothing else changes according to the analysis in Sect. 4. However, the increased diffusion of clusters shrinks $\rho_u(\mathcal{X})$ drastically which outweighs the effect of increasing α . Though the increased number of clusters increases the total volume of $\mathcal{X}_N$, but this may not increase $\rho_\ell(\mathcal{X}_N)$ [as discussed in item (b3) in Sect. 4.] Thus, the increased diffusion of clusters is the most dominant factor in case (ii).

We observe increased ψ_{opt} for all three algorithms in both cases. It is due to the increased α in case (ii); but it is due to the change in $\rho_\delta^b(\psi)$ in case (i).

The net effect in both cases is that the error increases for all three algorithms, as predicted in rows b2 and b3 in Table 1.

We investigate the effect of changing α that is due to the changing number of anomalies only in the next section.

7.4 Changing α by changing the number of anomalies only

Here we decrease α by increasing the number of anomalies through random selection using the ALOI $C = 10$ dataset: The number of anomalies is changed from 25, 270 to 553 (i.e., $\alpha = 0.98, 0.80, 0.67$)

The results are shown in Fig. 12. In all three algorithms, the error increases as the number of anomalies increases. Here ψ_{opt} decreases as α decreases. The changes in error and ψ_{opt} are as predicted in row a in Table 1.

However, the decrease in anomalies’ nearest neighbour distances (Δ_A), as shown in Fig. 12b, is not predicted by the theoretical analysis because there is a minimum impact on $\mathcal{X}$. The

Fig. 12 Changing α by changing the number of anomalies only [Sect. 7.4]. **(a)** Learning curves of iNNE for different numbers of randomly selected anomalies of the ALOI $C = 10$ dataset. **(b)** The average nearest neighbour distance and ψ_{opt} . The histogram is using the *left* y-axis (Avg NN Dist, Δ), whereas the line graph is using the *right* y-axis (ψ_{opt}). The learning curves of aNNE and kNN can be found in Fig. 19 in “Appendix 4”. **(a)** iNNE, **(b)** summarised results

decrease in Δ_A in this case is a direct result of the characteristics of anomalies in this dataset, i.e., many anomalies are in clusters. When the number of anomalies is small, anomalies’ nearest neighbours are normal instances. As the number of anomalies increases, members of the same clusters become their nearest neighbours, resulting in the reduction of Δ_A we have observed. It is interesting to note that even in this case, the change in Δ_A has correctly predicted the movement of error.

7.5 Changing α by changing the number of normal instances only

This experiment examines the impact of changing α by changing the number of normal instances through random selection, which has a minimum impact on $\mathcal{X}_N$ or $\mathcal{X}$.

We employ the same ALOI $C = 10$ dataset, as used in the previous subsection, but vary the percentage of normal instances from 50, 70 to 100 %, which are equivalent to $\alpha = 0.5, 0.58, 0.67$.

Figure 13 shows that increasing the number of normal instances reduces the errors. This change has a minimum impact on $\mathcal{X}$, which is reflected on the minor change in nearest neighbour distances (Δ) of both normal instances and anomalies. The increase in α also leads to the increase in ψ_{opt} . The extend of the increase is small because α is much smaller than 1.00 which does not cause a significant change to ψ_{opt} .¹¹ These observed changes are as predicted in row a in Table 1.

Section summary

Table 4 summarises the result in each subsection. The overall summary is given as follows:

1. All three anomaly detectors, aNNE, iNNE and kNN exhibit the gravity-defiant behaviours in seven out of the eight datasets, having $\psi_{opt} < \psi_{max}$. Even for the only dataset exhibiting the first half of the learning curves, which appears to be gravity-compliant, we have shown that the gravity-defiant behaviour will prevail if some variants of the dataset are employed, as predicted by the theoretical analysis.

¹¹ The derivative $d(\psi\alpha^\psi)/d\psi = \alpha^\psi + \psi \log(\alpha)\alpha^\psi = 0$ gives $\psi_{opt} = -1/\log(\alpha)$. When α is close to 1, ψ_{opt} changes drastically; otherwise the change to ψ_{opt} is small.

Fig. 13 Changing α by changing the number of normal instances only [Sect. 7.5]. (a) Learning curves of iNNE for different percentages of normal points of the ALOI $C = 10$ dataset. (b) The average nearest neighbour distance and ψ_{opt} for each of the different percentages. The histogram is using the *left* y-axis (Avg NN Dist, Δ), whereas the line graph is using the *right* y-axis (ψ_{opt}). The learning curves of aNNE and kNN can be found in Fig. 20 in “Appendix 4”. (a) iNNE, (b) summarised results

- The error is influenced by components of data characteristics, i.e., α , $\mathcal{X}$ or $\mathcal{X}_N$. The error changes in the opposite direction of $\rho_\delta^b(\psi)\alpha^\psi$.
- All changes in $\mathcal{X}$ or $\mathcal{X}_N$ in the experiments result in changes in anomalies' nearest neighbour distances (Δ_A), a direct consequence of $\rho_\delta^b(\psi) = \langle \rho_u(\mathcal{X})^b \rangle_\psi - \langle \rho_\ell(\mathcal{X}_N)^b \rangle_\psi$.
- The average nearest neighbour distance of anomalies (Δ_A) and error change in opposite directions.
- α and $\rho_\delta^b(\psi)$, independently, shift ψ_{opt} in the same direction. When both α and $\rho_\delta^b(\psi)$ impart their influence, the net effect on ψ_{opt} is hard to predict. Also, because $\rho_\delta^b(\psi)$ is the most difficult indicator to measure in practice, the direction of change in ψ_{opt} can be difficult to predict when $\rho_\delta^b(\psi)$ plays a significant part.
- The experiments verify the theoretical analysis that the change in Δ_A is able to predict the movement of error accurately, even in the case of clustered anomalies which is a factor not considered in the analysis.
- The empirical evaluation suggests that limiting the predictions only for instances covered by hyperspheres centred at $cnn(x)$ with radius $\tau(cnn(x))$ in iNNE (rather than always making a prediction based on $m(x, \eta_x)$ regardless of the distance between η_x and x as used in aNNE and kNN) has led to two positive outcomes: the learning curves of iNNE have a markedly steeper gradient and significantly smaller ψ_{opt} than those of aNNE and kNN.

8 Discussion

8.1 Other factors that influence the learning curve

The geometrical shape of $\mathcal{X}$ or $\mathcal{X}_N$ is an important factor which influences the AUC. Our descriptions in Sects. 4 and 7 have focused on the volume; but many of the changes would have affected the shape as well. For examples, the changes from Figs. 2, 3, 4 and 5, and the changes made to ALOI in Sect. 7.3 have affected both the volume and the shape. While the former is obvious, the latter can be assumed because the normal instances are from different normal clusters, even though the multi-dimensional ALOI dataset cannot be visualised.

It is possible that the change in shape alone has an impact on AUC. For example, using a hypothetical variant of the ALOI(C) dataset which allows us to increase the number of

Table 7 Squared bias and variance of kNN (for large k) and LiNearN (for $d > 1$), and their bias-variance trade-off parameters. The analytical results are extracted from [Fukunaga \(1990\)](#) and [Wells et al. \(2014\)](#)

	kNN	LiNearN
Squared bias	$O((k/n)^{\frac{4}{d}})$	$O(\psi^{-2/d})$
Variance	$O(k^{-1})$	$O(t^{-1} \psi^{1-2/d+\epsilon} \Psi^{-1})$
Bias-variance trade-off parameter	k	ψ

where ψ is the sample size used to build the hyperspheres, Ψ is the sample size used to estimate the density in each hypersphere t is the ensemble size, d is the number of dimensions, ϵ is a constant between 0 and 1; n is the size of the given dataset

normal clusters without changing the volume of $\mathcal{X}_N$. In this case, the density of each cluster increases and the volume of each cluster decreases as C increases such that the total volume and α remain unchanged. Assume that each cluster is well separated from others and for a fixed $\mathcal{D}$, this change in shape has the effect to reduce $\rho_u(\mathcal{X})$ significantly because instances in $\mathcal{D}$ are more spread out in $\mathcal{X}$ when there is a large number of normal clusters (i.e., a high C). Though $\rho_\ell(\mathcal{X}_N)$ may also be reduced in the process, $\rho_u(\mathcal{X})$ is reduced at a higher rate than $\rho_\ell(\mathcal{X}_N)$, leading to higher error. In comparison to Sect. 7.3, the source of the change is different though both changes lead to the same outcome, i.e., higher error as C increases.

Although our analysis in Sect. 3.4 assumes that α is close to 1, our experimental results using small α (e.g., in Sects. 7.3–7.5) suggest that the predictions from the analysis can still be applied successfully even though this assumption is violated.

8.2 Bias-variance analyses: density estimation versus anomaly detection

The current discussion on the bias and variance of kNN-based anomaly detectors ([Aggarwal and Sathe 2015](#)) assumes that the result of the bias-variance analysis on kNN density estimator carries over to kNN-based anomaly detectors. This assumption is not true because the accuracy of an anomaly detector depends on not only the accuracy of the density estimator employed, but also the distribution and the rate of anomaly contamination. The latter is not taken into consideration in the bias-variance analysis of density estimator.

For example, the bias-variance analysis on kNN density estimator ([Fukunaga 1990](#)), whose result is shown in Table 7, reveals that the optimal k is a monotonic function of data size; and the error becomes zero as $k \rightarrow \infty$ and $k/n \rightarrow 0$, i.e., gravity-compliance behaviour in terms of density estimation error. However, this does not imply that an anomaly detector based on this density estimator will have the same behaviour in terms of anomaly detection error. Our experiment using kNN anomaly detector shows that it has the gravity-defiant behaviour, i.e., global optimal n exists if $k = \sqrt{n}$.

The above-mentioned assumption, in addition to the dependency between k and n for kNN, creates complication and confusion in terms of learning curve for kNN-based anomaly detectors, as exemplified by the investigations conducted by [Zimek et al. \(2013\)](#) and [Aggarwal and Sathe \(2015\)](#). The former provides a case of gravity-defiant behaviour for kNN-based anomaly detector using a fixed k ; the latter points out the former's wrong reasoning, and shows that both gravity-compliance and gravity-defiant behaviours are possible using a fixed k . [Aggarwal and Sathe \(2015\)](#) posit that only gravity-compliance behaviour is observed if k is set to some fixed proportion of the data size. However, this conclusion is not derived from an analysis of detection error, but by 'stretching' the bias-variance analysis of density estimation error to detection error.

It is important to point out that the bias-variance analysis on kNN density estimator assumes that $1 \ll k \ll n$ (Fukunaga 1990). In other words, the analysis on kNN (Fukunaga 1990; Aggarwal and Sathe 2015) does not apply to 1NN or ensembles of 1NN.

The bias-variance analysis on LiNearN (Wells et al. 2014) is the only analysis available on ensemble of 1NN density estimator. The result of the analysis is shown in Table 7. But this result also cannot be used to explain the behaviour of anomaly detector based on it, for the same reason mentioned earlier.

In summary, to explain the behaviour of anomaly detectors, the analysis must be based on anomaly detection error. The existing bias-variance analysis on density estimator is not an appropriate tool for this purpose.

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

There is no magic to the gravity-defiant algorithms such as aNNE and iNNE which manifest that small data size yields the best performing model. Our 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!

We provide an intuitive example as follows. Consider a simple example that all normal instances are generated from a Gaussian distribution. Assume an oracle which provides the representative exemplar(s) of the given dataset for an 1NN anomaly detector. In this case, the only exemplar required is the instance which locates at the centre of the Gaussian distribution. Using the decision rule in Eq (2), where the oracle-picked exemplar is the only instance in $\mathcal{D}$, anomalies are those instances which have the longest distances from the centre, i.e., at the outer fringes of the Gaussian distribution. In this albeit ideal example, $\psi = 1$ for 1NN (as a single model) is sufficient to produce accurate detection. In fact, $\psi > 1$ can yield worse detection accuracy because $\mathcal{D}$ may now contain anomalies when they exist in the given dataset. Both the lower and upper bounds in our theoretical analysis also yield $\psi_{opt} = 1$ for the case of sharp Gaussian distribution.

In practice, we can obtain a result close to this oracle-induced result by random subsampling, as long as the data distribution admits that instances close to the centre has a higher probability of being selected than instances far from the centre, which is the case for sharp Gaussian distribution. Then, an average of an ensemble of 1NN derived from multiple samples $\mathcal{D}_i$ of $\psi = 1$ (randomly selected one instance) will approximate the result achieved by the oracle-picked exemplar. Pang et al. (2015) report that an ensemble of 1NN (which is the same as aNNE) achieves the best or close to the best result on many datasets using $\mathcal{D}_i$ of $\psi = 1$!

In a complex distribution (e.g., multiple peaks and asymmetrical shape), the oracle will need to produce more than one exemplar to represent fully the structure of the data distribution in order to yield good detection accuracy. For those distributions with moderate complexity, this number can still be significantly smaller than the size of the given dataset. Pang et al. (2015) report that 13 out of the 15 real-world datasets used (having data sizes up to 5 million instances) require $\psi \leq 16$ in their experiments. Note that the dataset size is irrelevant in terms of the number of exemplars required in both the intuitive example and complex distribution scenarios, as long as the dataset contains sufficient exemplars which are likely to be selected to represent the data distribution.

Sugiyama and Borgwardt (2013) have previously advocated the use of 1NN (as a single model) with a small sample size and provided a probabilistic explanation which can be paraphrased as follows: a small sample size ensures that the randomly selected instances are

likely to come from normal instances only; increasing the sample size increases the chance of including anomalies in the sample which leads to an increased number of false negatives (of predicting anomalies as normal instances).

The above intuitive example and our analysis based on computational geometry further reveal that 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.

8.4 Which nearest neighbour anomaly detector to use?

The investigation by Aggarwal and Sathe (2015) highlights the difficulty of using kNN because the accuracy of kNN-based anomaly detectors depends on not only the bias-variance trade-off parameter k , but also the data size. Furthermore, the bias-variance trade-off is delicate in kNN because a change in k alters both bias and variance in opposite directions (see Table 7). Our theoretical analysis points to an additional issue, i.e., kNN which insists on using all the available data (as dictated by the conventional wisdom) has no means to reduce the risk of anomaly contamination in the training dataset.

In contrast, our theoretical analysis reveals that, by using 1NN, the risk of anomaly contamination in the training sample can be controlled by selecting an appropriate sample size (ψ). The previous analysis on ensemble of 1NN¹² density estimator (Wells et al. 2014) shows that the ensemble size (t) can be increased independently to reduce the variance without affecting the bias (see the result shown in Table 7).

In addition, our empirical results show that both aNNE and kNN have approximately the same detection accuracy, but kNN requires approximately t times ψ_{opt} of aNNE in order to achieve its optimal detection accuracy.¹³ Moreover, searching for ψ (which is usually significantly less than k and does not depend on the data size) is a much easier task than searching for k which is a monotonic function of data size (Fukunaga 1990). All in all, we recommend ensembles of 1NN over kNN.

Between the two ensembles of 1NN, we recommend iNNE over aNNE because it reaches its optimal detection accuracy with a significantly smaller sample size.

Comparisons with other state-of-the-art anomaly detectors, which is outside the scope of this paper, can be found in Bandaragoda et al. (2014) and Pang et al. (2015).

8.5 Implications and potential future work

Both of our theoretical analysis and empirical result reveal that any changes in $\mathcal{X}$ or $\mathcal{X}_N$ lead to changes in nearest neighbour distances. In an unsupervised learning setting, changes in $\mathcal{X}$ or $\mathcal{X}_N$ are usually unknown and difficult to measure in practice. Yet any change that leads to the change in detection error can be measured in terms of nearest neighbour distance: If anomalies' nearest neighbour distances become shorter (or normal instances' nearest neighbour distances become longer), then we know that the detection error has increased, and vice versa. This is despite the fact that the source(s) of the change or the prediction error cannot be measured directly in an unsupervised learning task where labels for instances are not available at all times. However, Δ_A , the average nearest neighbour distance of anomalies, can be easily obtained in practice by examining a small proportion of instances which have the longest distances to their nearest neighbours in the given dataset (so can Δ_N by examining

¹² Note that iNNE is a simplified version of LiNearN (Wells et al. 2014) which does not need a second sample, i.e., Ψ shown in Table 7 is not relevant to iNNE or aNNE.

¹³ Similar result applies to ensemble of LOF ($k = 1$) versus LOF. See the result in 'Appendix 5'.

a portion of instances which have the shortest distances to their nearest neighbours), even though the labels are unknown.

This knowledge has a practical impact. In a data stream context, for example, timely model updates are crucial in maintaining the model's detection accuracy along the stream; and the updates rely on the ability to detect changes and the type of change in the stream (for example, whether they are due to changes in anomalies or normal clusters or both.) We are not aware of any good guidance with respect to change detection under different change scenarios in the unsupervised learning context. The majority of current works in data streams (Bifet et al. 2010; Masud et al. 2011; Duarte and Gama 2014) focus on supervised learning.

Our finding suggests that the net effect of any of these changes can be measured in terms of nearest neighbour distance, if a nearest neighbour anomaly detector such as iNNE or aNNE is used. This significantly reduces the type and the number of measurements required for change detection. The end result is a simple, adaptive and effective anomaly detector for data streams. A thorough investigation into the application of this finding in data streams will be conducted in the future.

The revelation of the gravity-defiant behaviour of nearest neighbour anomaly detectors invites broader investigation. Do other types of anomaly detectors, or more generally, learning algorithms for other data mining tasks also exhibit the gravity-defiant behaviour? 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). Is this the domain for which algorithms always comply with the learning curve? Or there is always a limit of domain complexity, over which the gravity-defiant behaviour will prevail. These are open questions that need to be answered.

9 Concluding remarks

As far as we know, this is the first work which investigates algorithms that defy the gravity of learning curve. It provides concrete evidence that there are gravity-defiant algorithms which produce good performing models with small training sets; and models trained with large data sizes perform worse.

Nearest neighbour-based anomaly detectors have been shown to be one of the most effective classes of anomaly detectors. Our analysis focuses on this class of anomaly detectors and provides a deeper understanding of its behaviour that has a practical impact in the age of big data.

The theoretical analysis based on computational geometry gives us an insight into the behaviour of the nearest neighbour anomaly detector. It shows that the AUC changes according to $\psi \alpha^\psi \langle \rho \rangle_\psi$, influenced by three factors: the proportion of normal instances (α), the radii (ρ) of $\mathcal{X}$ and $\mathcal{X}_N$, and the sample size (ψ) employed by the nearest neighbour-based anomaly detector. Because ψ and $\alpha^\psi \langle \rho \rangle_\psi$ are monotonic functions changing in opposite directions, an overly large sample size amplifies the negative impact of $\alpha^\psi \langle \rho \rangle_\psi$, leading to higher error at the tail end of the learning curve—the gravity-defiant behaviour.

We also discover that any change in $\mathcal{X}$ or $\mathcal{X}_N$, which varies the detection error, manifests as a change in nearest neighbour distance such that the detection error and anomalies' nearest neighbour distances change in opposite directions. Because nearest neighbour distance can be measured easily and other indicators of detection error are difficult to measure, it provides a unique useful practical tool to detect change in domains where any such changes are critical in their change detection operations, e.g., in data streams.

The knowledge that some algorithms can achieve high performance with a significantly small sample size is highly valuable in the age of big data because these algorithms consume significantly less computing resources (memory space and time) to achieve the same outcome as those require a large sample size.

We argue that existing bias-variance analyses on kNN-based density estimators are not an appropriate tool to be used to explain the behaviour of kNN-based anomaly detectors; and the analysis on kNN does not apply to 1NN or ensemble of 1NN on which our analysis targets. In addition, we further uncover that 1NN is not a poor cousin of kNN, rather an ensemble of 1NN has an operational advantage over kNN or an ensemble of kNN: It has only one parameter, i.e., sample size, rather than two parameters, k and data size, that influence the bias—this enables it to have a simpler parameter search. In the age of big data, the most important feature of ensemble of 1NN is that it has significantly smaller optimal sample size than kNN. Unless a compelling reason can be found, we recommend the use of ensemble of 1NN instead of kNN or ensemble of kNN.

Interesting future works include analysing the behaviour of (1) other types of anomaly detector, especially those which have shown good performance using small samples, such as iForest; and (2) gravity-defiant algorithms for other tasks such as classification and clustering.

Acknowledgements We would like to express our gratitude to Dr. Mahito Sugiyama in The Institute of Scientific and Industrial Research, Osaka University for his informative discussion with us. This material is based upon work supported by the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AOARD) under Award Numbers: 15IOA009-154006 (Kai Ming Ting), and 15IOA008-154005 (Takashi Washio). This work is also supported by JSPS KAKENHI Grant Number 2524003, awarded to Takashi Washio.

Appendix 1: Proofs

Proof of Lemma 1

Proof If $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) \leq r \leq \rho_u(\mathcal{D}_N, \mathcal{X})$, $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\} \neq \emptyset$. From $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\} \subseteq \bigcup_{y \in \mathcal{D}_N} S_d(y, r)$, Eqs. (1) and (4),

$$\begin{aligned} f(r; \mathcal{D}_N) &= \int_{\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}} p(x) dx \leq \int_{\bigcup_{y \in \mathcal{D}_N} S_d(y, r)} p(x) dx \\ &< \int_{\bigcup_{y \in \mathcal{D}_N} S_d(y, r)} p_u dx = \sum_{y \in \mathcal{D}_N} \int_{S_d(y, r)} p_u dx = p_u S_d h r^{d-1}. \end{aligned}$$

Otherwise, if $0 < r < \rho_\ell(\mathcal{D}_N, \mathcal{X}_N)$, $\{x | q(x; \mathcal{D}_N) = r\} \subseteq \mathcal{X}_N$ holds. Else $\rho_u(\mathcal{D}_N, \mathcal{X}) \leq r$ and thus $\{x | q(x; \mathcal{D}_N) = r\} \subseteq \mathcal{M} \setminus \mathcal{X}$ hold. In both cases, $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\} = \emptyset$. Hence,

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

□

Proof of Lemma 2

Proof From Eqs. (1) and (5),

$$G(r; \mathcal{D}_N) = \int_0^r g(s; \mathcal{D}_N) ds = \int_0^r \int_{\{x \in \mathcal{X}_N | q(x; \mathcal{D}_N) = s\}} p(x) dx ds$$

$$\begin{aligned}
&= \int_{\mathcal{X}_N \cap \{\bigcup_{y \in \mathcal{D}_N} B_d(y, r)\}} p(x) dx < \int_{\mathcal{X}_N \cap \{\bigcup_{y \in \mathcal{D}_N} B_d(y, r)\}} p_u dx \\
&\leq \min \left[\int_{\mathcal{X}_N} p_u dx, \int_{\{\bigcup_{y \in \mathcal{D}_N} B_d(y, r)\}} p_u dx \right] \\
&\leq \min[p_u |\mathcal{X}_N|, p_u B_d h r^d] \leq p_u |\mathcal{X}_N|.
\end{aligned}$$

□

Proof of Lemma 3

Proof From the assumption on $\mathcal{X}_A$, a finite partitioning of $\mathcal{X}_A$: $PT(\mathcal{X}_A)$ exists, where every subset $\mathcal{X}_a \in PT(\mathcal{X}_A)$ is convex and d dimensional. Moreover, some $\mathcal{X}_a \in PT(\mathcal{X}_A)$ intersects with $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$ at any r satisfying $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) < r < \rho_u(\mathcal{D}_N, \mathcal{X})$ by the definitions of these radii.

Let $r_\ell(\mathcal{X}_a)$ and $r_u(\mathcal{X}_a)$ be lower and upper bounds of r where $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$ intersects with $\mathcal{X}_a$. Because of $\mathcal{X}_a \subset \mathcal{X}_A$, $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) \leq r_\ell(\mathcal{X}_a) < r_u(\mathcal{X}_a) \leq \rho_u(\mathcal{D}_N, \mathcal{X})$ holds. According to the convexity of $\mathcal{X}_a$, the area that $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$ intersects with $\mathcal{X}_a$, i.e., $|\mathcal{X}_a \cap \{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}|$, is strictly positive at all $r \in [r_\ell(\mathcal{X}_a) + \delta, r_u(\mathcal{X}_a) - \delta]$ for any given $\delta \in (0, (r_u(\mathcal{X}_a) - r_\ell(\mathcal{X}_a))/2)$. Thus, there exists a positive constant $c_\delta(\mathcal{X}_a)$ satisfying

$$|\mathcal{X}_a \cap \{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}| \geq c_\delta(\mathcal{X}_a) \text{ for all } r \in [r_\ell(\mathcal{X}_a) + \delta, r_u(\mathcal{X}_a) - \delta]. \quad (6)$$

By the nature of $\mathcal{X}_a$ described in the first paragraph, there exist some non-empty $\mathcal{X}_a \cap \{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$ holding Eq. (6) over all r satisfying $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta \leq r \leq \rho_u(\mathcal{D}_N, \mathcal{X}) - \delta$ for every $\delta \in (0, \delta_0)$ under a given $\mathcal{D}_N$ where

$$\delta_0 = \min_{\mathcal{X}_a \in PT(\mathcal{X}_A)} (r_u(\mathcal{X}_a) - r_\ell(\mathcal{X}_a))/2.$$

From this fact, Eq. (1), Eq. (4) and that $\mathcal{X}_a \in PT(\mathcal{X}_A)$ are mutually disjoint, there exists

$$C_\delta \geq \min_{\mathcal{X}_a \in PT(\mathcal{X}_A)} c_\delta(\mathcal{X}_a)$$

holding

$$\begin{aligned}
f(r; \mathcal{D}_N) &= \int_{\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}} p(x) dx \\
&> \sum_{\mathcal{X}_a \in PT(\mathcal{X}_A), \mathcal{X}_a \cap \{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\} \neq \emptyset} p_\ell c_\delta(\mathcal{X}_a) \geq p_\ell C_\delta,
\end{aligned}$$

for any r satisfying $\delta \in (0, \delta_0)$ and $\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta \leq r \leq \rho_u(\mathcal{D}_N, \mathcal{X}) - \delta$.

At some r outside of this interval, any $\mathcal{X}_a \in PT(\mathcal{X}_A)$ does not intersect with $\{x \in \mathcal{X}_A | q(x; \mathcal{D}_N) = r\}$. Thus,

$$f(r; \mathcal{D}_N) \geq 0$$

holds. □

Proof of Lemma 4

Proof From the assumption on $\mathcal{X}_N$, a finite partitioning of $\mathcal{X}_N$: $PT(\mathcal{X}_N)$ exists where every subset $\mathcal{X}_n \in PT(\mathcal{X}_N)$ is convex and d dimensional. Let $\mathcal{D}_N(\mathcal{X}_n) = \{y \in \mathcal{D}_N | y \in \mathcal{X}_n\}$.

Given $\mathcal{X}_n \in PT(\mathcal{X}_N)$ where $\mathcal{D}_N(\mathcal{X}_n) \neq \emptyset$, from Proposition 2.1 in Evans et al. (2002), for every $y \in \mathcal{D}_N(\mathcal{X}_n)$, there exists some constant $c_\ell(y) \in \mathcal{R}^+$ at all $r \in (0, \rho_u(\{y\}, \mathcal{X}_n)]$ such that the volume of the intersection between $\mathcal{X}_n$ and $B_d(y, r)$ is lower bounded as

$$|\mathcal{X}_n \cap B_d(y, r)| > c_\ell(y)r^d.$$

Because $\mathcal{X}_n \cap B_d(y, r) \subseteq \mathcal{X}_n \cap \bigcup_{z \in \mathcal{D}_N(\mathcal{X}_n)} B_d(z, r)$ for all $y \in \mathcal{D}_N(\mathcal{X}_n)$,

$$\left| \mathcal{X}_n \cap \left\{ \bigcup_{z \in \mathcal{D}_N(\mathcal{X}_n)} B_d(z, r) \right\} \right| > c_\ell(\mathcal{X}_n)r^d$$

holds for $0 < r \leq \rho_u(\mathcal{X}_n)$ where $y_{\max} = \arg \max_{y \in \mathcal{D}_N(\mathcal{X}_n)} c_\ell(y)$, $c_\ell(\mathcal{X}_n) = c_\ell(y_{\max})$ and $\rho_u(\mathcal{X}_n) = \rho_u(\{y_{\max}\}, \mathcal{X}_n)$.

From this result, Eq.(1), Eq.(5) and that $\mathcal{X}_n \in PT(\mathcal{X}_N)$ are mutually disjoint,

$$\begin{aligned} G(r; \mathcal{D}_N) &= \int_0^r g(s; \mathcal{D}_N) ds = \int_0^r \int_{\{x \in \mathcal{X}_N | q(x; \mathcal{D}_N) = s\}} p(x) dx ds = \int_{\mathcal{X}_N \cap \{\bigcup_{y \in \mathcal{D}_N} B_d(y, r)\}} p(x) dx \\ &> \sum_{\forall \mathcal{X}_n \in PT(\mathcal{X}_N), \mathcal{D}_N(\mathcal{X}_n) \neq \emptyset} p_\ell c_\ell(\mathcal{X}_n) r^d \geq p_\ell c_\ell(\mathcal{X}_n^{\max}) r^d \end{aligned}$$

holds for $0 < r \leq \rho_u(\mathcal{X}_n^{\max})$ where

$$\mathcal{X}_n^{\max} = \arg \max_{\forall \mathcal{X}_n \in PT(\mathcal{X}_N), \mathcal{D}_N(\mathcal{X}_n) \neq \emptyset} c_\ell(\mathcal{X}_n).$$

By rescaling $c_\ell(\mathcal{X}_n^{\max})$ as $C'_\ell = c_\ell(\mathcal{X}_n^{\max}) \rho_u(\mathcal{X}_n^{\max})^d / \rho_u(\mathcal{D}_N, \mathcal{X}_N)^d$, we have

$$G(r; \mathcal{D}_N) > p_\ell C'_\ell r^d \quad (7)$$

for $0 < r \leq \rho_u(\mathcal{D}_N, \mathcal{X}_N)$.

Similarly, if $\rho_u(\mathcal{D}_N, \mathcal{X}_N) < r < \rho_u(\mathcal{D}_N, \mathcal{X})$, we obtain the following, because $\{x \in \mathcal{X}_N | q(x; \mathcal{D}_N) = r\} = \emptyset$ for any $r > \rho_u(\mathcal{D}_N, \mathcal{X}_N)$.

$$\begin{aligned} G(r; \mathcal{D}_N) &= \int_0^{\rho_u(\mathcal{D}_N, \mathcal{X}_N)} g(q; \mathcal{D}_N) dq + \int_{\rho_u(\mathcal{D}_N, \mathcal{X}_N)}^r g(q; \mathcal{D}_N) dq \\ &= \int_{\mathcal{X}_N \cap \{\bigcup_{y \in \mathcal{D}_N} B_d(y, \rho_u(\mathcal{D}_N, \mathcal{X}_N))\}} p(x) dx \\ &> p_\ell C'_\ell \rho_u(\mathcal{D}_N, \mathcal{X}_N)^d > p_\ell C'_\ell \frac{\rho_u(\mathcal{D}_N, \mathcal{X}_N)^d}{\rho_u(\mathcal{D}_N, \mathcal{X})^d} r^d. \end{aligned} \quad (8)$$

By combining Eqs. (7) and (8), we obtain a lower bound of $G(r; \mathcal{D}_N)$ as

$$G(r; \mathcal{D}_N) > p_\ell C_\ell r^d$$

for any r satisfying $0 < r < \rho_u(\mathcal{D}_N, \mathcal{X})$, where $C_\ell = C'_\ell \rho_u(\mathcal{D}_N, \mathcal{X}_N)^d / \rho_u(\mathcal{D}_N, \mathcal{X})^d < C'_\ell$.

If $\rho_u(\mathcal{D}_N, \mathcal{X}) \leq r$, $\bigcup_{x \in \mathcal{D}_N} B_d(x, r)$ includes $\mathcal{X}$. From $\mathcal{X}_N \subset \mathcal{X}$, Eq.(1) and Eq.(5),

$$G(r; \mathcal{D}_N) > \int_{\mathcal{X}_N \cap \{\bigcup_{x \in \mathcal{D}_N} B_d(x, r)\}} p_\ell dx = \int_{\mathcal{X}_N} p_\ell dx = p_\ell |\mathcal{X}_N|.$$

□

Proof of Lemma 5

Proof By substituting the results of Lemma 1 and Lemma 2 to Eq.(3),

$$\begin{aligned} AUC(\mathcal{D}_N) &< \int_{\rho_\ell(\mathcal{D}_N, \mathcal{X}_N)}^{\rho_u(\mathcal{D}_N, \mathcal{X})} p_u^2 | \mathcal{X}_N | S_d h r^{d-1} dr \\ &= \frac{p_u^2 | \mathcal{X}_N | S_d}{d} h (\rho_u(\mathcal{D}_N, \mathcal{X})^d - \rho_\ell(\mathcal{D}_N, \mathcal{X}_N)^d). \end{aligned}$$

Furthermore, by substituting the results of Lemmas 3 and 4 to Eq. (3),

$$\begin{aligned} AUC(\mathcal{D}_N) &> \int_{\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta}^{\rho_u(\mathcal{D}_N, \mathcal{X}) - \delta} p_\ell^2 C_\ell C_\delta r^d dr \\ &= \frac{p_\ell^2 C_\ell C_\delta}{d+1} \{ (\rho_\ell(\mathcal{D}_N, \mathcal{X}) - \delta)^{d+1} - (\rho_\ell(\mathcal{D}_N, \mathcal{X}_N) + \delta)^{d+1} \} \end{aligned}$$

□

Proof of Theorem 1

Proof Because α is the probability to draw normal instances from D , and $h = |\mathcal{D}_N|$ is the number of normal instances drawn into $\mathcal{D}$ which has size ψ , h follows a binomial distribution $\mathcal{B}(\psi, \alpha)$, and its probability distribution is represented as

$$P(h) = \binom{\psi}{h} \alpha^h (1 - \alpha)^{\psi-h}.$$

Thus, from Corollary 1, the expectation of AUC is upper and lower bounded by

$$\begin{aligned} \langle AUC \rangle &< \frac{p_u^2 | \mathcal{X}_N | S_d}{d} \sum_{h=0}^{\psi} \binom{\psi}{h} \alpha^h (1 - \alpha)^{\psi-h} h \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_h - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_h \right) \\ &= \frac{p_u^2 | \mathcal{X}_N | S_d}{d} \alpha^\psi \psi \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_\psi - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_\psi \right) \\ &\quad + \frac{p_u^2 | \mathcal{X}_N | S_d}{d} \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1 - \alpha)^{\psi-h} h \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_h - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_h \right), \text{ and} \end{aligned} \quad (9)$$

$$\begin{aligned} \langle AUC \rangle &> \frac{p_\ell^2 C_\ell C_\delta}{d+1} \sum_{h=0}^{\psi} \binom{\psi}{h} \alpha^h (1 - \alpha)^{\psi-h} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_h - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_h \right\} \\ &= \frac{p_\ell^2 C_\ell C_\delta}{d+1} \alpha^\psi \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_\psi - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_\psi \right\} \\ &\quad + \frac{p_\ell^2 C_\ell C_\delta}{d+1} \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1 - \alpha)^{\psi-h} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_h - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_h \right\}. \end{aligned} \quad (10)$$

Since the difference between the covering radius of $\mathcal{X}$ and the inradius of $\mathcal{X}_N$ is finite, there exists some constant M_u to upper bound the last term in Eq. (9) as

$$\begin{aligned} & \frac{p_u^2 |\mathcal{X}_N| S_d}{d} \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1-\alpha)^{\psi-h} h \left(\left\langle \rho_u(\mathcal{X})^d \right\rangle_h - \left\langle \rho_\ell(\mathcal{X}_N)^d \right\rangle_h \right) \\ & \leq M_u \psi \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1-\alpha)^{\psi-h} = M_u \{1 - \alpha^\psi\} \psi = M_u \psi (1-\alpha) \sum_{h=0}^{\psi-1} \alpha^h \\ & < M_u (1-\alpha) \psi^2 = O((1-\alpha) \psi^2). \end{aligned}$$

Similarly, some constant M_ℓ exists to upper bound the last term in Eq. (10) as

$$\begin{aligned} & \frac{p_\ell^2 C_\ell C_\delta}{d+1} \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1-\alpha)^{\psi-h} \left\{ \left\langle (\rho_u(\mathcal{X}) - \delta)^{d+1} \right\rangle_h - \left\langle (\rho_\ell(\mathcal{X}_N) + \delta)^{d+1} \right\rangle_h \right\} \\ & \leq M_\ell \sum_{h=0}^{\psi-1} \binom{\psi}{h} \alpha^h (1-\alpha)^{\psi-h} = M_\ell \{1 - \alpha^\psi\} = M_\ell (1-\alpha) \sum_{h=0}^{\psi-1} \alpha^h \\ & < M_\ell (1-\alpha) \psi = O((1-\alpha) \psi). \end{aligned}$$

These prove the theorem. $\square$

Appendix 2: Algorithms for aNNE and iNNE

The procedure used by aNNE to rank instances is provided in Algorithm 1.

Algorithm 1: aNNE

input : D - Given dataset; t - number of samples; ψ - sample size

output: A ranked list of D

- 1 Generate $\{\mathcal{D}_i \mid i = 1, \dots, t\}$ from D using sampling without replacement (with restart), where $|\mathcal{D}_i| = \psi$
 - 2 **for** each $x \in D$ **do**
 - 3 **for** $i = 1, \dots, t$ **do**
 - 4 $q(x; \mathcal{D}_i) = \min_{y \in \mathcal{D}_i} m(x, y)$
 - 5 **end**
 - 6 $\bar{q}(x; D) = \frac{1}{t} \sum_{i=1}^t q(x; \mathcal{D}_i)$
 - 7 **end**
 - 8 Sort D based on $\bar{q}(x; D)$ in descending order
-

iNNE, shown in Algorithm 2, employs inverse distance rather than distance to avoid the case that $\tau(cnn(x)) = \infty$ if $cnn(x)$ is undefined (i.e., when x is not covered by any hyperspheres.) Thus, the key differences between iNNE and aNNE are in steps 1 and 4: The construction of hyperspheres in training and the nearest neighbour employed in evaluation.

Algorithm 2: iNNE

input : D - Given dataset; t - number of samples; ψ - sample size
output: A ranked list of D

- 1 $\{ \bigcup_{c \in \mathcal{D}_i} B(c, \tau(c)) \mid i = 1, \dots, t \} \leftarrow \text{Generation-of-Hyperspheres}(D, t, \psi)$
- 2 **for** each $x \in D$ **do**
- 3 **for** $i = 1, \dots, t$ **do**
- 4 $inverse_q_i = \begin{cases} \frac{1}{\tau(cnn(x))}, & \text{if } x \in \bigcup_{c \in \mathcal{D}_i} B(c, \tau(c)) \\ 0, & \text{otherwise} \end{cases}$
- 5 **end**
- 6 $\bar{q}(x; D) = [\frac{1}{t} \sum_{i=1}^t inverse_q_i]^{-1}$
- 7 **end**
- 8 Sort D based on $\bar{q}(x)$ in descending order

Algorithm 3: Generation-of-Hyperspheres in iNNE

input : D - Given dataset; t - number of samples; ψ - sample size
output: $\mathbb{H}$ - a set of t sets of ψ hyperspheres

- 1 $\mathbb{H} \leftarrow \emptyset$
- 2 **for** $i = 1, \dots, t$ **do**
- 3 Generate $\mathcal{D}_i$ by randomly selecting ψ points from D without replacement
- 4 $\mathbb{B}_i \leftarrow \emptyset$
- 5 **for** each $c \in \mathcal{D}_i$ **do**
- 6 $B(c, \tau(c)) \leftarrow \text{A hypersphere centred at } c \text{ with radius } \tau(c) = \min_{y \in \mathcal{D}_i \setminus \{c\}} m(y, c).$
- 7 $\mathbb{B}_i \leftarrow \mathbb{B}_i \cup \{B(c, \tau(c))\}$
- 8 **end**
- 9 $\mathbb{H} \leftarrow \mathbb{H} \cup \{\mathbb{B}_i\}$
- 10 **end**

Appendix 3: An example offset on the Smtp dataset

Figure 14 shows an example of offsetting all anomalies on Smtp by transposing along the diagonal by 0.03.

Because of the distribution of the 32 anomalies, only 11 anomalies' nearest neighbours are normal instances. As the offset increases to 0.03, only 1 anomaly's nearest neighbour is a normal instance; and none are normal instances when the offset is 0.2.

Despite this fact, all three anomaly detectors can still take advantage of the enlarged $\mathcal{X}_A$ to reduce the error, as predicted from the theoretical analysis.

Appendix 4: Detailed empirical results

This appendix provides the additional results of the following sections:

- Section 7.1: Results on Mulcross, Mammography, U2R and ALOI C=10 are shown in Fig. 15.

Fig. 14 An example of offsetting the anomalies on Smtip. (a) Data distribution of Smtip on the first two dimensions. (b) All anomalies are transposed along the diagonal by 0.03

Fig. 15 Learning curves of iNNE, kNN and aNNE: Error is defined as 1-AUC. kNN uses training set size of $t\psi$ and $k = \sqrt{t\psi}$; aNNE and iNNE are using $t = 100$. (a) Mulcross, (b) mammography, (c) U2R, (d) ALOI $C = 10$

- Section 7.2: Learning curves of aNNE and kNN for the different offset values of the Smtip dataset are shown in Fig. 16.
- Section 7.3(i): Learning curves of aNNE and kNN for the different complexities of the ALOI $C = 10$ dataset are shown in Fig. 17.
- Section 7.3(ii): Learning curves of aNNE and kNN for the ALOI dataset ($C = 10, 100, 500, 900$) are shown in Fig. 18.

Fig. 16 Learning curves of aNNE and kNN for the different offset values of the Smt dataset. (a) aNNE, (b) kNN

Fig. 17 Learning curves of aNNE and kNN for the different complexities of the ALOI $C = 10$ dataset. (a) aNNE, (b) kNN

Fig. 18 Learning curves of aNNE and kNN for the ALOI dataset ($C = 10, 100, 500, 900$). (a) aNNE, (b) kNN

- Section 7.4: Changing α by changing the number of anomalies only. Learning curves of aNNE and kNN for the different numbers of randomly selected anomalies of the ALOI $C = 10$ dataset are shown in Fig. 19.
- Section 7.5: Changing α by changing the number of normal instances only. Learning curves of aNNE and kNN for different percentages of normal instances of the ALOI $C = 10$ dataset are shown in Fig. 20.

Fig. 19 Changing α by changing the number of anomalies only. Learning curves of aNNE and kNN for the different numbers of randomly selected anomalies of the ALOI $C = 10$ dataset. (a) aNNE, (b) kNN

Fig. 20 Changing α by changing the number of normal instances only. Learning curves of aNNE and kNN for different percentages of normal instances of the ALOI $C = 10$ dataset. (a) aNNE, (b) kNN

Fig. 21 Learning curves of LOF and an 1NN ensemble version of LOF ($k = 1$): Error is defined as $1 - \text{AUC}$. LOF used training set size of $t\psi$ and $k = \sqrt{t\psi}$; the 1NN ensemble version of LOF used $t = 100$. (a) ALOI $C = 10$, (b) CoverType

Appendix 5: Example gravity-defiant behaviours of LOF

This appendix shows the gravity-defiant behaviour of LOF (Breunig et al. 2000), which employs kNN, and the 1NN ensemble version of LOF on two datasets in Fig. 21. The WEKA (Hall et al. 2009) implementation of LOF was used in this experiment.

References

- Aggarwal, C. C., & Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explorations*, 17(1), 24–47.
- Bandaragoda, T., Ting, K. M., Albrecht, D., Liu, F., & Wells, J. (2014). Efficient anomaly detection by isolation using nearest neighbour ensemble. In *Proceedings of the 2014 IEEE international conference on data mining, workshop on incremental classification, concept drift and novelty detection* (pp. 698–705).
- Banko, M., & 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, association for computational linguistics, ACL '01* (pp. 26–33).
- Bay, S., & Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 29–38).
- Bifet, A., Frank, E., Holmes, G., & Pfahringer, B. (2010). Accurate ensembles for data streams: Combining restricted hoeffding trees using stacking. In *JMLR workshop and conference proceedings. The 2nd Asian conference on Machine learning* (Vol. 13, pp. 225–240).
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data, ACM* (pp. 93–104).
- Duarte, J., & Gama, J. (2014). Ensembles of adaptive model rules from high-speed data streams. *JMLR workshop and conference proceedings. The 3rd international workshop on big data* (Vol. 36, pp. 198–213). Algorithms, Systems, Programming Models and Applications: Streams and Heterogeneous Source Mining.
- Evans, D., Jones, A. J., & Schmidt, W. M. (2002). Asymptotic moments of near-neighbour distance distributions. *Proceedings: Mathematical, Physical and Engineering Sciences*, 458(2028), 2839–2849.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). San Diego: Academic Press.
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 10–18.
- Hand, D. J., & 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.
- Lichman, M. (2013). UCI machine learning repository. archive.ics.uci.edu/ml
- Liu, F., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *Proceedings of the eighth IEEE international conference on data mining* (pp. 413–422).
- Masud, M., Gao, J., Khan, L., Han, J., & Thuraishingham, B. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6), 859–874.
- Pandya, D., Upadhyay, S., & Harsha, S. (2013). Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using apf-knn. *Expert Systems with Applications*, 40(10), 4137–4145.
- Pang, G., Ting, K. M., & Albrecht, D. (2015). LeSiNN: Detecting anomalies by identifying least similar nearest neighbours. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 623–630).
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Sugiyama, M., & Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems*, 26, 467–475.
- Wells, J. R., Ting, K. M., & Washio, T. (2014). LiNearN: A new approach to nearest neighbour density estimator. *Pattern Recognition*, 47(8), 2702–2720.
- 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.
- Zimek, A., Gaudet, M., Campello, R. J., & 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* (pp. 428–436).
- Zitzler, E., Laumanns, M., & Bleuler, S. (2004). A tutorial on evolutionary multiobjective optimization. In X. Gandibleux, M. Sevaux, K. Sörensen, & V. T'Kindt (Eds.), *Metaheuristics for multiobjective optimisation* (pp. 3–37). Berlin, Heidelberg: Springer.