Isolation Distributional Kernel: A New Tool for Kernel based Anomaly Detection

Kai Ming Ting & Bi-Cun Xu
National Key Laboratory for Novel Software Technology, Nanjing University, China
{tingkm,xubc}@lamda.nju.edu.cn

Takashi Washio
The Institute of Scientific and Industrial Research, Osaka University, Japan
washio@ar.sanken.osaka-u.ac.jp

Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University, China
zhouzh@lamda.nju.edu.cn

ABSTRACT

We introduce Isolation Distributional Kernel as a new way to measure the similarity between two distributions. Existing approaches based on kernel mean embedding, which converts a point kernel to a distributional kernel, have two key issues: the point kernel employed has a feature map with intractable dimensionality; and it is *data independent*. This paper shows that Isolation Distributional Kernel (IDK), which is based on a *data dependent* point kernel, addresses both key issues. We demonstrate IDK's efficacy and efficiency as a new tool for kernel based anomaly detection. Without explicit learning, using IDK alone outperforms existing kernel based anomaly detector OCSVM and other kernel mean embedding methods that rely on Gaussian kernel. We reveal for the first time that an effective kernel based anomaly detector based on kernel mean embedding must employ a characteristic kernel which is data dependent.

CCS CONCEPTS

• **Computing methodologies** → **Kernel methods**;

KEYWORDS

Distributional Kernel, Kernel Mean Embedding, Anomaly Detection

ACM Reference format:
Kai Ming Ting & Bi-Cun Xu, Takashi Washio, and Zhi-Hua Zhou. 2020. Isolation Distributional Kernel: A New Tool for Kernel based Anomaly Detection. In *Proceedings of 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, USA, August 23–27, 2020 (KDD '20),* 9 pages.
https://doi.org/10.1145/3394486.3403062

1 INTRODUCTION

In many real-world applications, an object can be represented more effectively and naturally as a set of data points generated from a distribution [4, 9, 19], e.g., a bag of points in multi-instance learning; an image as a collection of patches; and a galaxy cluster of individual galaxies. This treatment demands a way to measure similarity between two distributions.

Kernel mean embedding [10, 16] on distributions is one effective way to build a distributional kernel from a point kernel, enabling similarity between distributions to be measured. The current approach has focused on point kernels which have *a feature map with intractable dimensionality*. This feature map is a known key issue in kernel mean embedding in the literature [10]; and it has led to $O(n^2)$ time complexity, where n is the input data size.

Here, we identify that being *data independent* point kernel is another key issue which compromises the effectiveness of the similarity measurement that impacts on task-specific performance.

These two issues are exemplified in anomaly detection: An existing kernel mean embedding method took more than 9000 seconds and produced a detection accuracy of 0.85 on a dataset having more than half a million data points. In contrast, the proposed method took less than 60 seconds with an accuracy of 0.97. See the experiments section for details.

We propose to employ a *data dependent* point kernel to address the above two key issues directly. As it is implemented using Isolation Kernel [13, 20], we called the proposed kernel mean embedding: Isolation Distributional Kernel or IDK.

Our contributions are:

(1) Proposing a new implementation of Isolation Kernel which has the required data dependent property for anomaly detection. We provide a geometrical interpretation of the new Isolation Kernel in Hilbert Space that explains its superior detection accuracy over existing Isolation Kernel.
(2) Formally proving that the new Isolation Kernel (i) has the following data dependent property: *two points, as measured by Isolation Kernel derived in a sparse region, are more similar than the same two points, as measured by Isolation Kernel derived in a dense region*; and (ii) is a characteristic kernel, i.e., its kernel mean map is injective in Hilbert space.
(3) Introducing Isolation Distributional Kernel (IDK). It is distinguished from existing distributional kernels in two aspects. First, the use of the data dependent Isolation Kernel produces high accuracy in anomaly detection tasks. Second, the Isolation Kernel's exact and finite-dimensional feature map enables IDK to have $O(n)$ time complexity.
(4) Proposing a new kernel based anomaly detector using IDK.
(5) Demonstrating that, without explicit learning, IDK anomaly detector not only significantly outperforms Gaussian kernel based anomaly detectors in detecting point anomalies, but it runs orders of magnitude faster. This is an evidence that an

KDD '20, August 23–27, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403062

Table 1: Key symbols and notations.

$\widehat{\mathcal{K}}_I$	Isolation Distributional Kernel (IDK)
$\widehat{\Phi}$	Exact feature map of IDK
$\widehat{\mathcal{K}}_G$	Gaussian Distributional Kernel (GDK)
$\widehat{\varphi}$	Approximate feature map of GDK
$\widehat{\mathcal{K}}_{NG}$	Nyström accelerated GDK using $\widehat{\varphi}$

effective kernel based anomaly detector constructed from kernel mean embedding must use a point kernel which is *both a characteristic kernel and data dependent.*

(6) Revealing that the proposed anomaly detector is one of the isolation based methods; but it is the only isolation based method which is based on kernel.

In addition, the proposed method is the only detector that makes use of distributional kernel for point anomaly detection.

2 BACKGROUND

In this section, we briefly describe the kernel mean embedding and Isolation Kernel as the background of the proposed method. The key symbols and notations used are shown in Table 1.

2.1 Kernel mean embedding and two key issues

Let S and T be two nonempty datasets where each point x in S and T belongs to a subspace $\mathcal{X} \subseteq \mathbb{R}^d$ and is drawn from probability distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ defined on $\mathbb{R}^d$, respectively. $\mathcal{P}_S$ and $\mathcal{P}_T$ are strictly positive on $\mathcal{X}$ and strictly zero on $\overline{\mathcal{X}} = \mathbb{R}^d \setminus \mathcal{X}$, *i.e.*, $\forall X \subseteq \mathcal{X}$ s.t. $X \neq \emptyset; \mathcal{P}_S(X), \mathcal{P}_T(X) > 0$, and $\forall X \subseteq \overline{\mathcal{X}}$ s.t. $X \neq \emptyset; \mathcal{P}_S(X), \mathcal{P}_T(X) = 0$. We denote the density of $\mathcal{P}_S$ and $\mathcal{P}_T$ as $\mathcal{P}_S(x)$ and $\mathcal{P}_T(x)$, respectively.

Using kernel mean embedding [10, 16], the empirical estimation of the distributional kernel $\widehat{\mathcal{K}}$ on $\mathcal{P}_S$ and $\mathcal{P}_T$, which is based on a point kernel κ on points $x, y \in \mathcal{X}$, is given as:

$$\widehat{\mathcal{K}}_G(\mathcal{P}_S, \mathcal{P}_T) = \frac{1}{|S||T|} \sum_{x \in S} \sum_{y \in T} \kappa(x, y). \quad (1)$$

2.1.1 First issue: Feature map has intractable dimensionality. The distributional kernel $\widehat{\mathcal{K}}_G$ relies on a point kernel κ, e.g., Gaussian kernel, which has a feature map with intractable dimensionality.

The use of a point kernel which has *a feature map with intractable dimensionality is regarded as a fundamental issue of kernel mean embedding* [10]. This can be seen from Eq 1 which has time complexity $O(n^2)$ if each set of S and T has data size n.

It has been recognised that the time complexity can be reduced by utilising the feature map of the point kernel [10].

If the chosen point kernel can be approximated as $\kappa(x, y) \approx \langle \varphi(x), \varphi(y) \rangle$, where φ is a finite-dimensional feature map approximating the feature map of κ. Then, $\widehat{\mathcal{K}}$ can be written as

$$\widehat{\mathcal{K}}_{NG}(\mathcal{P}_S, \mathcal{P}_T) \approx \frac{1}{|S||T|} \sum_{x \in S} \sum_{y \in T} \varphi(x)^\top \varphi(y) \approx \langle \widehat{\varphi}(\mathcal{P}_S), \widehat{\varphi}(\mathcal{P}_T) \rangle \quad (2)$$

where $\widehat{\varphi}(\mathcal{P}_T) = \frac{1}{|T|} \sum_{x \in T} \varphi(x)$ is the empirical estimation of the approximate feature map of $\widehat{\mathcal{K}}_G(\mathcal{P}_T, \cdot)$, or equivalently, the kernel mean map of $\mathcal{P}_T$ in RKHS (Reproducing Kernel Hilbert Space) $\mathscr{H}$ associated with $\widehat{\mathcal{K}}_G$.

Note that the approximation $\kappa(x, y) \approx \langle \varphi(x), \varphi(y) \rangle$ is essential in order to have a finite-dimensional feature map. **This enables the use of Eq 2 to reduce the time complexity of computing $\widehat{\mathcal{K}}(\mathcal{P}_S, \mathcal{P}_T)$ to $O(n)$** since $\widehat{\varphi}(\mathcal{P})$ can be computed independently in $O(n)$. Otherwise, Eq 1 must be used which costs $O(n^2)$.

A successful approach of finite-dimensional feature map approximation is kernel functional approximation. Representative methods are Nyström method [22] and Random Fourier Features [14, 24].

Existing distributional kernels such as the level-2 kernel used in support measure machine (SMM) [9], Mean map kernel (MMK) [19] and Efficient Match Kernel (EMK) [4] have exactly the same form as shown in Eq 1, where κ can be any of the existing data independent point kernels. Both MMK and EMK employ a kernel functional approximation in order to use Eq 2.

In summary, using a point kernel which has a feature map with intractable dimensionality, the kernel functional approximation is an enabling step to approximate the point kernel with a finite-dimensional feature map. Otherwise, the mapping from T to a point $\widehat{\varphi}(\mathcal{P}_T)$ cannot be performed; and Eq 2 cannot be computed. However, these methods of kernel functional approximation are computationally expensive; and they almost always weaken the final outcome in comparison with that derived from Eq 1.

2.1.2 Second issue: kernel is data independent. In addition to the known key issue mentioned above, we identify that *a data independent point kernel is a key issue* which leads to poor task-specific performance.

The weakness of using a data independent kernel/distance is well recognised in the literature. For example, distance metric learning [21, 23, 25] aims to transform the input space such that points of the same class become closer and points of different classes are lengthened in the transformed space than those in the input space. Distance metric learning has been shown to improve the classification accuracy of k nearest neighbour classifiers [21, 23, 25].

A recent work has shown that data independent kernels such as Laplacian kernel and Gaussian Kernel are the source of weaker predictive SVM classifiers [20]. Unlike distance metric learning, it creates a data dependent kernel directly from data, requiring neither class information nor explicit learning. It also provides a reason why a data dependent kernel is able to improve the predictive accuracy of SVM that uses a data independent kernel.

Here we show that the use of data independent kernel reduces the effectiveness of kernel mean embedding in the context of anomaly detection. The resultant anomaly detectors which employ Gaussian kernel, using either $\widehat{\mathcal{K}}_G$ or $\widehat{\mathcal{K}}_{NG}$, perform poorly (see Section 6.) This is because a data independent kernel is employed.

In a nutshell, **the source of the two key issues is: the point kernel employed has a feature map with intractable dimensionality and is data independent**. We address both issues from its source by using a recently introduced Isolation Kernel [13, 20].

2.2 Isolation Kernel

Let $D \subset \mathcal{X} \subseteq \mathbb{R}^d$ be a dataset sampled from an unknown $\mathcal{P}_D$; and $\mathbb{H}_\psi(D)$ denote the set of all partitionings H that are admissible from $\mathcal{D} \subset D$, where each point $z \in \mathcal{D}$ has the equal probability of being

selected from D; and $|\mathcal{D}| = \psi$. Each $\theta[z] \in H$ isolates a point $z \in \mathcal{D}$ from the rest of the points in $\mathcal{D}$. Let $\mathbb{1}(\cdot)$ be an indicator function.

Definition 2.1. [13, 20] For any two points $x, y \in \mathbb{R}^d$, Isolation Kernel of x and y is defined to be the expectation taken over the probability distribution on all partitionings $H \in \mathbb{H}_\psi(D)$ that both x and y fall into the same isolating partition $\theta[z] \in H$, where $z \in \mathcal{D} \subset D$, $\psi = |\mathcal{D}|$:

$$\begin{aligned}\kappa_I(x, y \mid D) &= \mathbb{E}_{\mathbb{H}_\psi(D)}[\mathbb{1}(x, y \in \theta[z] \mid \theta[z] \in H)] \\ &= \mathbb{E}_{\mathcal{D} \subset D}[\mathbb{1}(x, y \in \theta[z] \mid z \in \mathcal{D})] \\ &= P(x, y \in \theta[z] \mid z \in \mathcal{D} \subset D) \end{aligned} \quad (3)$$

In practice, κ_I is constructed using a finite number of partitionings $H_i, i = 1, \ldots, t$, where each H_i is created using randomly subsampled $\mathcal{D}_i \subset D$; and θ is a shorthand for $\theta[z]$:

$$\begin{aligned}\kappa_I(x, y \mid D) &= \frac{1}{t}\sum_{i=1}^{t} \mathbb{1}(x, y \in \theta \mid \theta \in H_i) \\ &= \frac{1}{t}\sum_{i=1}^{t}\sum_{\theta \in H_i} \mathbb{1}(x \in \theta)\mathbb{1}(y \in \theta) \end{aligned} \quad (4)$$

Isolation Kernel is positive semi-definite as Eq 4 is a quadratic form. Thus, Isolation Kernel defines a RKHS $\mathscr{H}$.

The isolation partitioning mechanisms which have been used previously to implement Isolation Kernel are iForest [20], and Voronoi diagram [13] (they are applied to SVM and clustering.)

3 PROPOSED ISOLATION KERNEL

Here we introduce a new implementation of Isolation Kernel, together with its exact and finite feature map, and its data dependent property in the following three subsections.

3.1 A new implementation of Isolation Kernel

An isolation mechanism for Isolation Kernel which has not been employed previously is given as follows:

Each point $z \in \mathcal{D}$ is isolated from the rest of the points in $\mathcal{D}$ by building a hypersphere that covers z only. The radius of the hypersphere is determined by the distance between z and its nearest neighbor in $\mathcal{D} \setminus \{z\}$. In other words, a partitioning H consists of ψ hyperspheres $\theta[z]$ and the $(\psi+1)$-th partition. The latter is the region in $\mathbb{R}^d$ which is not covered by all ψ hyperspheres. Note that $2 \leq \psi < |D|$. Figure 1 (left) shows an example using $\psi = 3$.

This mechanism has been shown to produce large partitions in a sparse region and small partitions in a dense region [2]. Although it was previously used as a point anomaly detector called iNNE [2], its use in creating a kernel is new.

3.2 Feature map of new Isolation Kernel

Given a partitioning H_i, let $\Phi_i(x)$ be a ψ-dimensional binary column vector representing all hyperspheres $\theta_j \in H_i, j = 1, \ldots, \psi$; where x falls into either only one of the ψ hyperspheres or none. The j-component of the vector is: $\Phi_{ij}(x) = \mathbb{1}(x \in \theta_j \mid \theta_j \in H_i)$. Given t partitionings, $\Phi(x)$ is the concatenation of $\Phi_1(x), \ldots, \Phi_t(x)$.

Definition 3.1. **Feature map of Isolation Kernel.** For point $x \in \mathbb{R}^d$, the feature mapping $\Phi : x \to \{0,1\}^{t\times\psi}$ of κ_I is a vector

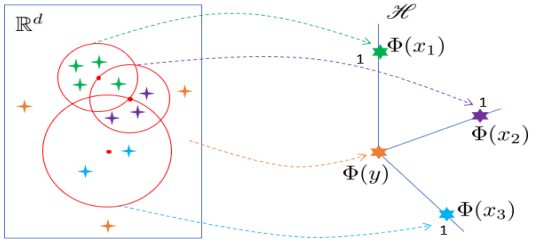

Figure 1: An illustration of feature map Φ of Isolation Kernel with one partitioning ($t = 1$) of three hyperspheres, each centred at a point (as red dot) $z \in \mathcal{D}$ where $|\mathcal{D}| = \psi = 3$ are randomly selected from the given dataset D. When a point x falls into an overlapping region, x is regarded to be in the hypersphere whose centre is closer to x.

that represents the partitions in all the partitioning $H_i \in \mathbb{H}_\psi(D)$, $i = 1, \ldots, t$; where x falls into either only one of the ψ hyperspheres or none in each partitioning H_i.

Let $\mathbb{1}$ be a shorthand of $\Phi_i(x)$ such that $\Phi_{ij}(x) = 1$ and $\Phi_{ik}(x) = 0, \forall k \neq j$ for any $j \in [1, \psi]$.

Φ has the following geometrical interpretation:

(a) $\Phi(x) = [\mathbb{1}, \ldots, \mathbb{1}]$: $\|\Phi(x)\| = \sqrt{t}$ and $\kappa_I(x, x|D) = 1$ iff $\Phi_i(x) \neq \mathbf{0}$ for all $i \in [1, t]$.

(b) For point x such that $\exists i \in [1, t], \Phi_i(x) = \mathbf{0}$; then $\|\Phi(x)\| < \sqrt{t}$.

(c) If point $x \in \mathbb{R}^d$ falls outside of all hyperspheres in H_i for all $i \in [1, t]$, then it is mapped to the origin of the feature space $\Phi(x) = [\mathbf{0}, \ldots, \mathbf{0}]$.

Let T be a set of normal points and S a set of anomalies. Let the given dataset $D = T$ which consists of normal points only[1]. In the context of anomaly detection, assuming that $\mathcal{P}_T$ is the distribution of normal points x and the largely different $\mathcal{P}_S$ is the distribution of point anomalies y. The geometrical interpretation gives rise to: (i) point anomalies $y \in S$ are mapped close to the origin of RKHS because they are different from normal points $x \in T$—they largely satisfy condition (c) and sometimes (b); and (ii) Φ of individual normal points $x \in T$ have norm equal or close to $\sqrt{t}$—they largely satisfy condition (a) and sometimes (b). In other words, normal points are mapped to or around $[\mathbb{1}, \ldots, \mathbb{1}]$. Note that $[\mathbb{1}, \ldots, \mathbb{1}]$ is not a single point in RKHS, but points which have $\|\Phi(x)\| = \sqrt{t}$ and $\Phi_i(x) = \mathbb{1}$ for all $i \in [1, t]$.

Figure 1 shows an example mapping Φ of Isolation Kernel using $\psi = 3$ and $t = 1$, where all points falling into a particular hypersphere are mapped to the same point in RKHS. Those points which fall outside of all hyperspheres are mapped to the origin of RKHS.

The previous implementations of Isolation Kernel [13, 20] possess condition (a) only; thus, they have no capability to separate anomalies from normal points by using the norm of Φ only.

[1]This assumption is for clarity of the exposition only; and D is used to derive Isolation Kernel. In practice, D could contain anomalies but has a minimum impact on κ_I. See details later.

Note that Definition 3.1 is an **exact feature map** of Isolation Kernel. Re-express Eq 4 using Φ gives:

$$\kappa_I(x,y) = \frac{1}{t}\langle \Phi(x), \Phi(y)\rangle \tag{5}$$

In contrast, existing finite-dimensional feature map derived from a data independent kernel is an **approximate feature map**, i.e.,

$$\kappa(x,y) \approx \langle \varphi(x), \varphi(y)\rangle \tag{6}$$

This leads to the approximation used in existing kernel mean embedding, showed in Eq 2.

3.3 Data dependent property of new Isolation Kernel

The new partitioning mechanism produces large hyperspheres in a sparse region and small hyperspheres in a dense region. This yields the following property [20]: *two points in a sparse region are more similar than two points of equal inter-point distance in a dense region.*

Here we provide a theorem for the equivalent property: **two points, as measured by Isolation Kernel derived in a sparse region, are more similar than the same two points, as measured by Isolation Kernel derived in a dense region.**

Theorem 1. *Given two probability distributions $\mathcal{P}_D, \mathcal{P}_{D'} \in \mathbb{P}$ from which points in datasets D and D' are drawn, respectively. Let $\mathcal{E} \subset \mathcal{X}$ be a region such that $\forall_{w\in\mathcal{E}}, \mathcal{P}_D(w) < \mathcal{P}_{D'}(w)$, i.e., D is sparser than D' in $\mathcal{E}$. Assume that $\psi = |\mathcal{D}|$ is large such that $\tilde{z}$ is the nearest neighbour of z, where $z, \tilde{z} \in \mathcal{D} \subset D$ in $\mathcal{E}$, under a given metric distance ℓ (the same applies to $z', \tilde{z}' \in \mathcal{D}' \subset D'$ in $\mathcal{E}$.)*

Isolation Kernel κ_I based on hyperspheres $\theta(z) \in H$ has the property that $\kappa_I(x,y \mid D) > \kappa_I(x,y \mid D')$ for any point-pair $x, y \in \mathcal{E}$.

Proof. Let ℓ between x and y be ℓ_{xy}. $x, y \in \theta[z]$ if and only if the nearest neighbour of both x and y is z in $\mathcal{D}$, and $\ell_{z\tilde{z}} > \max(\ell_{xz}, \ell_{yz})$ holds for $\tilde{z}$ the nearest neighbour of z in $\mathcal{D}$. Moreover, the triangular inequality $\ell_{xz} + \ell_{yz} > \ell_{xy}$ holds because ℓ is a metric distance. Accordingly,

$$\begin{aligned}
&P(x,y \in \theta[z] \mid z \in \mathcal{D} \subset D)\\
&\quad = P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{yz}\} \wedge \{\ell_{yz} > \ell_{xy} - \ell_{yz}\}) +\\
&\quad\quad P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{xy} - \ell_{yz}\} \wedge \{\ell_{xy} - \ell_{yz} > \ell_{yz}\}) +\\
&\quad\quad P(\{\ell_{z\tilde{z}} > \ell_{yz} > \ell_{xz}\} \wedge \{\ell_{xz} > \ell_{xy} - \ell_{xz}\}) +\\
&\quad\quad P(\{\ell_{z\tilde{z}} > \ell_{yz} > \ell_{xy} - \ell_{xz}\} \wedge \{\ell_{xy} - \ell_{xz} > \ell_{xz}\})\\
&\quad = 2P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{yz}\} \wedge \{\ell_{yz} > \ell_{xy} - \ell_{yz}\}) +\\
&\quad\quad 2P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{xy} - \ell_{yz}\} \wedge \{\ell_{xy} - \ell_{yz} > \ell_{yz}\})
\end{aligned} \tag{7}$$

subject to the nearest neighbour $z \in \mathcal{D}$ of both x and y. The last equality holds by the symmetry of ℓ_{xz} and ℓ_{yz}.

Given a hypersphere $v(c,\ell_{cz})$ centered at $c \in \mathcal{E}$ and having radius ℓ_{cz} equal to the distance from c to its nearest neighbour $z \in \mathcal{D}$, let $\mathcal{P}(u(c,\ell_{cz}))$ be the probability density of probability mass $u(c,\ell_{cz})$ in $v(c,\ell_{cz})$; $u(c,\ell_{cz}) = \int_{v(c,\ell_{cz})} \mathcal{P}_D(w)dw$. Note that $u(c,\ell_{cz})$ is strictly monotonic to ℓ_{cz} if $v(c,\ell_{cz}) \cap \mathcal{X} \neq \emptyset$, since $\mathcal{P}_D$ is strictly positive in $\mathcal{X}$. Then, the followings are derived.

$$\begin{aligned}
&P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{yz}\} \wedge \{\ell_{yz} > \ell_{xy} - \ell_{yz}\})\\
&\quad = P(\ell_{z\tilde{z}} > \ell_{xz} > \ell_{yz} > \ell_{xy}/2)\\
&\quad = \int_{u(z,\ell_{xy}/2)}^{1} \mathcal{P}(u(z,\ell_{z\tilde{z}})) \int_{u(x,\ell_{xy}/2)}^{u(x,\ell_{z\tilde{z}})} \mathcal{P}(u(x,\ell_{xz})) \times\\
&\quad\quad \int_{u(y,\ell_{xy}/2)}^{u(y,\ell_{xz})} \mathcal{P}(u(y,\ell_{yz}))du(y,\ell_{yz})du(x,\ell_{xz})du(z,\ell_{z\tilde{z}})\\
&\quad \approx \int_{u(z,\ell_{xy}/2)}^{u(z,\hat{\ell}_{z\mathcal{E}})} \mathcal{P}(u(z,\ell_{z\tilde{z}})) \int_{u(x,\ell_{xy}/2)}^{u(x,\ell_{z\tilde{z}})} \mathcal{P}(u(x,\ell_{xz})) \times\\
&\quad\quad \int_{u(y,\ell_{xy}/2)}^{u(y,\ell_{xz})} \mathcal{P}(u(y,\ell_{yz}))du(y,\ell_{yz}du(x,\ell_{xz})du(z,\ell_{z\tilde{z}}),
\end{aligned} \tag{8}$$

$$\begin{aligned}
&P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{xy} - \ell_{yz}\} \wedge \{\ell_{xy} - \ell_{yz} > \ell_{yz}\})\\
&\quad = P(\{\ell_{z\tilde{z}} > \ell_{xz} > \ell_{xy} - \ell_{yz}\} \wedge \{\ell_{xy}/2) > \ell_{yz}\})\\
&\quad = \int_{u(z,\ell_{xy}/2)}^{1} \mathcal{P}(u(z,\ell_{z\tilde{z}})) \int_{0}^{u(y,\ell_{xy}/2)} \mathcal{P}(u(y,\ell_{yz})) \times\\
&\quad\quad \int_{u(x,\ell_{xy}-\ell_{yz})}^{u(x,\ell_{z\tilde{z}})} \mathcal{P}(u(x,\ell_{xz}))du(x,\ell_{xz})du(y,\ell_{yz})du(z,\ell_{z\tilde{z}})\\
&\quad \approx \int_{u(z,\ell_{xy}/2)}^{u(z,\hat{\ell}_{z\mathcal{E}})} \mathcal{P}(u(z,\ell_{z\tilde{z}})) \int_{0}^{u(y,\ell_{xy}/2)} \mathcal{P}(u(y,\ell_{yz})) \times\\
&\quad\quad \int_{u(x,\ell_{xy}-\ell_{yz})}^{u(x,\ell_{z\tilde{z}})} \mathcal{P}(u(x,\ell_{xz}))du(x,\ell_{xz})du(y,\ell_{yz})du(z,\ell_{z\tilde{z}}),
\end{aligned} \tag{9}$$

where $\hat{\ell}_{z\mathcal{E}} = \sup_{v(z,\ell_{z\tilde{z}})\subseteq\mathcal{E}} \ell_{z\tilde{z}}$. The approximate equality holds by the assumption in the theorem which implies that the integral from $u(z,\hat{\ell}_{z\mathcal{E}})$ to 1 for $u(z,\ell_{z\tilde{z}})$ is negligible. The same argument applied to $P(x,y \in \theta[z'] \mid z' \in \mathcal{D}' \subset D')$ which derives the identical result.

[7] provided the expressions of $\mathcal{P}(u(c,\ell_{cz}))$ and $\mathcal{P}(u(z,\ell_{z\tilde{z}}))$ as

$$\begin{aligned}
\mathcal{P}(u(c,\ell_{cz})) &= \psi(1-u(c,\ell_{cz}))^{\psi-1},\\
\mathcal{P}(u(z,\ell_{z\tilde{z}})) &= (\psi-1)(1-u(z,\ell_{z\tilde{z}}))^{\psi-2}.
\end{aligned}$$

With these expressions and the definition of $u(c,\ell_{cz})$, both $\mathcal{P}(u(c,\ell_{cz}))$ and $\mathcal{P}(u(z,\ell_{z\tilde{z}}))$ are lower if $\mathcal{P}_D(z)$ becomes higher. Accordingly,

$$P(x,y \in \theta[z] \mid z \in \mathcal{D} \subset D) > P(x,y \in \theta[z'] \mid z' \in \mathcal{D}' \subset D')$$

holds by the fact $\forall_{w\in\mathcal{E}}, \mathcal{P}_D(w) < \mathcal{P}_{D'}(w)$, Eq. 7, Eq. 8 and Eq. 9.

This result and Definition 2.1 prove the theorem. □

Theorem 1 is further evidence that the data dependent property of Isolation Kernel only requires that the isolation mechanism produces large partitions in sparse region and small partitions in dense region, regardless of the actual space partitioning mechanism. We use hyperspheres to partition the space here; where the previous works use Voronoi diagram [13] and axis-parallel partitions in a tree structure [20].

4 ISOLATION DISTRIBUTIONAL KERNEL

4.1 Definition of IDK

Given the feature map Φ (defined in Definition 3.1) and Eq 5, the empirical estimation of kernel mean embedding can be expressed based on the feature map of Isolation Kernel $\kappa_I(x,y)$.

Definition 4.1. Isolation Distributional Kernel of two distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ is given as:

$$\begin{aligned}\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T) &= \frac{1}{t|S||T|}\sum_{x\in S}\sum_{y\in T}\Phi(x)^\top\Phi(y)\\ &= \frac{1}{t}\left\langle\widehat{\Phi}(\mathcal{P}_S), \widehat{\Phi}(\mathcal{P}_T)\right\rangle \end{aligned} \quad (10)$$

where $\widehat{\Phi}(\mathcal{P}_S) = \frac{1}{|S|}\sum_{x\in S}\Phi(x)$ is the empirical feature map of the kernel mean embedding.

Condition (a) wrt $\|\Phi(x)\|$ in Section 3.2 leads to $\|\widehat{\Phi}(x)\| = \sqrt{t}$; and similarly $0 \le \|\widehat{\Phi}(x)\| < \sqrt{t}$ holds under conditions (b) and (c) in Section 3.2. Thus, $\left\langle\widehat{\Phi}(\mathcal{P}_S), \widehat{\Phi}(\mathcal{P}_T)\right\rangle \in [0, t]$ *i.e.*, $\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T) \in [0, 1]$.

We call this particular implementation of kernel mean embedding, Isolation Distributional Kernel or IDK.

The key advantages of IDK over existing kernel mean embedding [4, 9, 19] are: (i) Φ is an exact and finite-dimensional feature map of a data dependent point kernel; whereas φ in Eq 2 is an approximate feature map of a data independent point kernel. (ii) The distributional characterisation of $\widehat{\Phi}(\mathcal{P}_T)$ is derived from Φ's adaptability to local density in T; whereas the distributional characterisation of $\widehat{\varphi}(\mathcal{P}_T)$ lacks such adaptability because φ of Gaussian kernel is data independent. This is despite the fact that both Isolation Kernel and Gaussian kernel are a characteristic kernel (see the next section.)

4.2 Theoretical Analysis: Is Isolation Kernel a characteristic kernel?

As defined in subsection 2.1, we consider $\mathcal{P}_S, \mathcal{P}_T \in \mathbb{P}$, where $\mathbb{P}$ is a set of probability distributions on $\mathbb{R}^d$ which are admissible but strictly positive on $\mathcal{X}$ and strictly zero on $\overline{\mathcal{X}}$. This implies that no data points exist outside of $\mathcal{X}$. Thus we limit our analysis to the property of the kernel on $\mathcal{X}$. A positive definite kernel κ is a characteristic kernel if its kernel mean map $\widehat{\Phi} : \mathbb{P} \to \mathscr{H}$ is injective, *i.e.*, $\|\widehat{\Phi}(\mathcal{P}_S) - \widehat{\Phi}(\mathcal{P}_T)\|_{\mathscr{H}} = 0$ if and only if $\mathcal{P}_S = \mathcal{P}_T$ [10]. If the kernel κ is non-characteristic, two different distributions $\mathcal{P}_S \neq \mathcal{P}_T$ may be mapped to the same $\widehat{\Phi}(\mathcal{P}_S) = \widehat{\Phi}(\mathcal{P}_T)$.

Isolation Kernel derived from the partitioning H_i can be interpreted that $\mathcal{X}$ is packed by hyperspheres having random sizes. This is called random-close packing [3]. Previous studies revealed that the upper bound of the rate of the packed space in a 3-dimensional space is almost 64% for any random-close packing [3, 17]. The packing rates for the higher dimensions are known to be far less than 100% for any distribution of sizes of hyperspheres. This implies that the $(\psi+1)$-th partition, which is not covered by any hyperspheres, always has nonzero volume, even when ψ is very large.

Let $R \subset \mathcal{X}$ and $\overline{R} = \mathcal{X} \setminus R$ be regions such that $\forall x \in R, \mathcal{P}_S(x) \neq \mathcal{P}_T(x)$ and $\forall x \in \overline{R}, \mathcal{P}_S(x) = \mathcal{P}_T(x)$.

From the fact that $\int_{\overline{R}}(\mathcal{P}_S(x) - \mathcal{P}_T(x))dx = 0$ and $\int_{\mathcal{X}}\mathcal{P}_S(x)dx = \int_{\mathcal{X}}\mathcal{P}_T(x)dx = 1$,

$$\int_R (\mathcal{P}_S(x) - \mathcal{P}_T(x))dx = 0 \quad (11)$$

is deduced. In conjunction with this relation and the fact $\forall x \in R$, $\mathcal{P}_S(x) \neq \mathcal{P}_T(x)$, there exists at least one $R' \subset R$ such that

$$\int_{R'} (\mathcal{P}_S(x) - \mathcal{P}_T(x))dx \neq 0. \quad (12)$$

This requires R to contain at least two distinct points in $\mathcal{X}$.

Accordingly, a partitioning H_i of Isolation Kernel satisfies one of the following two mutually exclusive cases:

- Case 1: $\exists\theta \in H_i, \theta \supseteq R$. From Eq 11, the probability of $x \sim \mathcal{P}_S$ falling into θ and that of $x \sim \mathcal{P}_T$ falling into θ are identical. Thus, the difference between $\mathcal{P}_S$ and $\mathcal{P}_T$ does not produce any difference between $\Phi(x \sim \mathcal{P}_S)$ and $\Phi(x \sim \mathcal{P}_T)$ in expectation.
- Case 2: $\exists\theta \in H_i, \theta \cap R \neq \emptyset$ and $\theta \not\supseteq R$. If $\theta \cap R$ is one of R' satisfying Eq 12, then $\Phi(x \sim \mathcal{P}_S)$ and $\Phi(x \sim \mathcal{P}_T)$ are different in expectation.

These observations give rise to the following theorem:

Theorem 2. *The kernel mean map of the new Isolation Kernel (generated from D) $\widehat{\Phi} : \mathbb{P} \to \mathscr{H}$ is characteristic in $\mathcal{X}$ with probability 1 in the limit of $t \to \infty$, for $\psi, t \ll |D|$.*

Proof. To define H, points in $\mathcal{D} \subset D$ are drawn from $\mathcal{P}_D(x)$ which is strictly positive on $\mathcal{X}$, *i.e.*, $\forall X \subseteq \mathcal{X}$ s.t. $X \neq \emptyset, \mathcal{P}_D(X) > 0$. This implies that any partitioning H of $\mathcal{X}$, created by $\mathcal{D}$, has non-zero probability, since the points in $\mathcal{D}$ can be anywhere in $\mathcal{X}$ with non-zero probability. Also recall that $\psi = |\mathcal{D}| = 2$ is the minimum sample size required to construct the hyperspheres in H (see Section 3.1.) Let's call this: the property of H.

Due to the property of H and $\psi \ge 2$, there exists $H_i \in \{H_1, \dots, H_t\}$ and $\theta_j \in H_i$ with probability 1 such that $x \in \theta_j$ and $y \notin \theta_j$ for any mutually distinct points x and y in $\mathcal{X}$, as $t \to \infty$. This implies that, as $t \to \infty$, there exists $\Phi_{ij}(x)$ for any $x \in D$ with probability 1 such that $\Phi_{ij}(x) = 1$ and $\Phi_{ij}(y) = 0, \forall y \in D, y \neq x$. Then, the Gram matrix of Isolation Kernel is full rank, because the feature maps $\Phi(x)$ for all points $x \in D$ are mutually independent. Accordingly, Isolation Kernel is a positive definite kernel with probability 1 in the limit of $t \to \infty$.

Because of the property of H and the fact that all $\theta \in H_i$ including the $\psi+1$-th partition have non-zero volumes for any $\psi \ge 2$, the probability of Case 1 is not zero. In addition, because R contains at least two distinct points in $\mathcal{X}$, the probability of Case 2 is not zero for any $\psi \ge 2$. These facts yield $0 < p < 1$, where p is the probability of an event that H_i satisfies Case 1 but not Case 2. Since $\psi, t \ll |D|$, H_i are almost independently sampled over $i = 1, \dots, t$, and the probability of the event occurring over all $i = 1, \dots, t$ is p^t. If $t \to \infty$, then $p^t \to 0$. This implies that Isolation Kernel is injective with probability 1 in the limit of $t \to \infty$.

Both the positive definiteness and the injectivity imply that Isolation Kernel is a characteristic kernel. □

Some data independent kernels such as Gaussian kernel are characteristic [10]. Because an empirical estimation uses a finite dataset, their kernel mean maps that ensure injectivity (see section 4 in [18]) are as good as that using Isolation Kernel with large t.

To use the kernel mean map for anomaly detection, a point kernel deriving the kernel mean map must be characteristic. Otherwise, anomalies of $\mathcal{P}_S$ may not be properly separated from normal points of $\mathcal{P}_T$ because some anomalies and normal points may be mapped to an identical point $\widehat{\Phi}(\mathcal{P}_T) = \widehat{\Phi}(\mathcal{P}_S)$. As we will show in the experiment section, Isolation Kernel using $t = 100$ is sufficient to produce better result than Gaussian kernel (which is characteristic) in kernel mean embedding for anomaly detection.

4.3 Data dependent property of IDK

Following Definitions 4.1, IDK of two distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ can be redefined as the expected probability over the probability distribution on all partitioning $H \in \mathbb{H}_\psi(D)$ that a randomly selected point-pair $x \in S$ and $y \in T$ falls into the same isolation partition $\theta[z] \in H$, where $z \in \mathcal{D} \subset D$, $\psi = |\mathcal{D}|$:

$$\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D) = \mathbb{E}_{\mathbb{H}_\psi(D)}\left[\mathbb{I}(x, y \in \theta[z] \mid \theta[z] \in H; x \in S, y \in T)\right]$$

If the supports of both $\mathcal{P}_S$ and $\mathcal{P}_T$ are included in $\mathcal{E}$ and $\forall_{w \in \mathcal{E}}$, $\mathcal{P}_D(w) < \mathcal{P}_{D'}(w)$ holds, the above expression leads to the following proposition, since every point-pair from S and T follows Theorem 1.

Proposition 1. *Under the conditions on $\mathcal{P}_D$, $\mathcal{P}_{D'}$, D, D' and $\mathcal{E}$ of Theorem 1, given two distribution-pairs $\mathcal{P}_S, \mathcal{P}_T \in \mathbb{P}$ where the supports of both $\mathcal{P}_S$ and $\mathcal{P}_T$ are in $\mathcal{E}$, the IDK $\widehat{\mathcal{K}}_I$ based on hyperspheres $\theta(z) \in H$ has the property that $\widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D) > \widehat{\mathcal{K}}_I(\mathcal{P}_S, \mathcal{P}_T|D')$.*

In other words, the data dependent property of Isolation Kernel leads directly to the data dependent property of IDK:
Two distributions, as measured by IDK derived in sparse region, are more similar than the same two distributions, as measured by IDK derived in dense region.

5 PROPOSED METHOD FOR KERNEL BASED ANOMALY DETECTION

Let $\widehat{\Phi}(\mathcal{P}_D)$ be a kernel mean mapped point of a distribution $\mathcal{P}_D$ of a dataset D. The kernel mean embedding $\widehat{\mathcal{K}}$, as a result of the mapping $\widehat{\Phi}(\mathcal{P}_D)$, shall produce the following:

- If $x \sim \mathcal{P}_D$, $\widehat{\mathcal{K}}(\delta(x), \mathcal{P}_D)$ is large, which can be interpreted as x is likely to be part of $\mathcal{P}_D$.
- If $y \nsim \mathcal{P}_D$, $\widehat{\mathcal{K}}(\delta(y), \mathcal{P}_D)$ is small, which can be interpreted as y is not likely to be part of $\mathcal{P}_D$.

With this interpretation, $y \nsim \mathcal{P}_D$ can be naturally used to identify anomalies in D. The definition of anomaly is thus:

'Given a similarity measure $\widehat{\mathcal{K}}$ of two distributions, an anomaly is an observation whose Dirac measure δ has a low similarity with the distribution from which a reference dataset is generated.'

This is an operational definition, based on distributional kernel $\widehat{\mathcal{K}}$, of the one given by Hawkins (1980):

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

Let $\mathcal{P}_T$ and $\mathcal{P}_S$ be the distributions of normal points and point anomalies, respectively. We assume that $\forall x \in T, x \sim \mathcal{P}_T$; and S consists of subsets of anomalies S_i from mutually distinct distributions, i.e., $S = \cup_{i=1}^{m} \mathrm{S}i$, where $\mathcal{P}_{\mathrm{S}_i} = \delta(x_i)$ is the distribution of an anomaly x_i represented as a Dirac measure δ, and $\mathrm{S}_i = \{x_i\}$.

Note that in the unsupervised learning context, only the dataset $D = S \cup T$ is given without information about S and T, where the majority of the points in D are from $\mathcal{P}_T$; but D also contains few points from $\mathcal{P}_S$. Because $|S| \ll |T|$, $\mathcal{P}_D \approx \mathcal{P}_T$. The kernel mean map of $\mathcal{P}_T$ is empirically estimated from D, i.e., $\widehat{\Phi}(\mathcal{P}_T) \approx \widehat{\Phi}(\mathcal{P}_D) = \frac{1}{|D|}\sum_{x \in D} \Phi(x)$ is thus robust against influences from $\mathcal{P}_S$ because the few points in S create significantly lower weights than the many points in T. Therefore, $\widehat{\Phi}(\mathcal{P}_T) \approx \widehat{\Phi}(\mathcal{P}_D)$ is in a region distant from those of $\widehat{\Phi}(\mathcal{P}_{\mathrm{S}_i})$, $\forall i$. In essence, *IDK derived from D is robust against contamination of anomalies in D.*

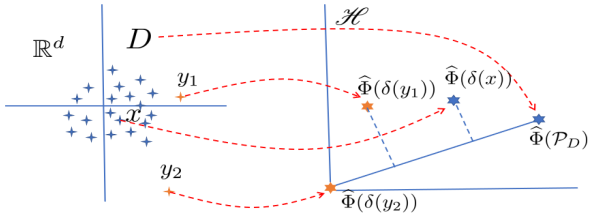

Figure 2: An illustration of kernel mean mapping $\widehat{\Phi}$ used in $\widehat{\mathcal{K}}_I$ for a dataset D, containing anomalies y_1 & y_2. $\widehat{\Phi}$ maps from $\mathbb{R}^d$ of D & individual points to points in $\mathscr{H}$.

We compute IDK by using Isolation Kernel constructed with D, and its kernel mean map projects $\mathcal{P}_D$ and $\mathcal{P}_{\mathrm{S}_i}$ to two distinct points: $\widehat{\Phi}(\mathcal{P}_D)$ and $\widehat{\Phi}(\mathcal{P}_{\mathrm{S}_i})$ in $\mathscr{H}$.

Point anomaly detector $\widehat{\mathcal{K}}$: The proposed detector is one which summarizes the entire dataset D into one mapped point $\widehat{\Phi}(\mathcal{P}_D)$. To detect point anomalies, map each $x \in D$ to $\widehat{\Phi}(\delta(x))$; and compute its similarity w.r.t. $\widehat{\Phi}(\mathcal{P}_D)$, i.e., $\left\langle \widehat{\Phi}(\delta(x)), \widehat{\Phi}(\mathcal{P}_D) \right\rangle$. Then sort all computed similarities to rank all points in D. Anomalies are those points which are least similar to $\widehat{\Phi}(\mathcal{P}_D)$. The anomaly detector due to IDK is computed using Eq 10, denoted as $\widehat{\mathcal{K}}_I$; and the detectors using Gaussian kernel and its Nyström approximation are computed using Eq 1 and Eq 2, respectively, denoted as $\widehat{\mathcal{K}}_G$ and $\widehat{\mathcal{K}}_{NG}$.

Figure 2 shows an example mapping $\widehat{\Phi}$ of a normal point x, two anomalies y_1 and y_2 as well as D from $\mathbb{R}^d$ to $\mathscr{H}$. The global anomaly y_2 is mapped to the origin of $\mathscr{H}$; and y_1 which is just outside the fringe of a normal cluster is mapped to a position where $\widehat{\mathcal{K}}_I(\delta(y_1), \mathcal{P}_D)$ is closer to the origin than that of normal points.

To be an effective anomaly detector using kernel mean embedding, we show in the next section that the kernel employed being characteristic is not sufficient without being data dependent; and **the power of $\widehat{\mathcal{K}}_I$ enables it to be the only kernel-based anomaly detector that does not need explicit learning.**

6 EXPERIMENTS

The detection accuracy of an anomaly detector is measured in terms of AUC (Area under ROC curve). As all the anomaly detectors are unsupervised learners, all models are trained with unlabelled training sets. Only after the models have made predictions, ground truth labels are used to compute the AUC for each dataset. We report the runtime of each detector in terms of CPU seconds.

Parameter settings used in the experiments: $\widehat{\mathcal{K}}_I$ uses $t = 100$; and ψ is searched over $\psi \in \{2^m \mid m = 1, 2 \ldots 12\}$. (These settings are the same for iForest [8] and iNNE [2], to be described in Section 7.) For all methods using Gaussian kernel, the bandwidth is searched over $\{2^m \mid m = -5, \ldots 5\}$. The sample size of the Nyström method is set as $\sqrt{n}$ which is also equal to the number of features. All datasets are normalized to $[0, 1]$ in the preprocessing.

We compare kernel based anomaly detectors $\widehat{\mathcal{K}}_I$, $\widehat{\mathcal{K}}_G$ and $\widehat{\mathcal{K}}_{NG}$ (using the Nyström method [12] to produce an approximate feature map as preprocessing) with an existing kernel based anomaly detector: OCSVM [15] which employs Gaussian kernel.

Table 2: Results of kernel based anomaly detectors (AUC).

Dataset	#Inst	#Ano	#Attr	$\widehat{\mathcal{K}}_I$	$\widehat{\mathcal{K}}_G$	$\widehat{\mathcal{K}}_{NG}$	OCSVM
speech	3686	61	400	0.76	0.46	0.47	0.65
EEG_eye	8422	165	14	0.88	0.55	0.47	0.54
PenDigits	9868	20	17	0.98	0.98	0.95	0.98
MNIST_230	12117	10	784	0.98	0.97	0.97	0.96
MNIST_479	12139	50	784	0.86	0.69	0.60	0.59
mammograg	11183	260	6	0.88	0.85	0.86	0.84
electron	37199	700	50	0.80	0.65	0.57	0.64
shuttle	49097	3511	9	0.98	0.98	0.99	0.98
ALOI	50000	1508	27	0.82	0.60	0.54	0.53
muon	94066	500	50	0.82	0.63	0.55	0.75
smtp	95156	30	3	0.95	0.81	0.77	0.91
IoT_botnet	213814	1000	115	0.99	0.83	0.68	0.93
ForestCover	286048	2747	10	0.97	0.85	0.86	0.96
http	567497	2211	3	0.99	0.99	0.98	0.97
Average rank				1.25	2.64	3.18	2.92

Table 3: Runtime comparison (in CPU seconds) on http.

	$\widehat{\mathcal{K}}_I$	$\widehat{\mathcal{K}}_G$	$\widehat{\mathcal{K}}_{NG}$	OCSVM	iNNE	iForest
train	31	0	106	13738	15	6
test	2	9689	1	1964	1	18

6.1 Evaluation of kernel based detectors

Table 2 shows that $\widehat{\mathcal{K}}_I$ is the best anomaly detector among the four detectors. The huge difference in AUC between $\widehat{\mathcal{K}}_I$ and $\widehat{\mathcal{K}}_G$ (e.g., speech, MNIST_479 & ALOI) shows the superiority of Isolation Kernel over data independent Gaussian kernel. As expected, $\widehat{\mathcal{K}}_{NG}$ performed worse than $\widehat{\mathcal{K}}_G$ in general because it uses an approximate feature map.

A Friedman-Nemenyi test [5] in Figure 3 shows that $\widehat{\mathcal{K}}_I$ is significantly better than the other algorithms.

Table 3 shows that $\widehat{\mathcal{K}}_I$ has short testing and training times. In contrast, $\widehat{\mathcal{K}}_G$ has the longest testing time though it has the shortest (zero) training time; and OCSVM has the longest training time. This is because its operations are based on points without feature map ($\widehat{\mathcal{K}}_G$ uses Eq 1). While $\widehat{\mathcal{K}}_{NG}$ (uses Eq 2) has significantly reduced the testing time of $\widehat{\mathcal{K}}_G$, it still has longer training time than $\widehat{\mathcal{K}}_I$ because of the Nyström process.

Time complexities: To compute the similarity of two sets, each having n points, $\widehat{\mathcal{K}}_G$ takes $O(n^2)$. For $\widehat{\mathcal{K}}_I$, the preprocessing $\widehat{\Phi}(\mathcal{P}_S)$ of a set S of n points takes $O(nt\psi)$ and needs to be completed once only. $\widehat{\mathcal{K}}_I$ takes $O(t\psi)$ only to compute the similarity between two sets. Thus, the overall time complexity is $O(nt\psi)$. For large datasets, $t\psi \ll n$, this accounts for the huge difference in testing times between the two methods we observed in Table 3. The time complexity of OCSVM in LibSVM is $O(n^3)$.

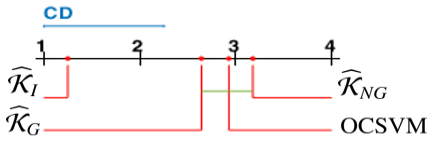

Figure 3: Friedman-Nemenyi test for anomaly detectors at significance level 0.05. If two algorithms are connected by a CD (critical difference) line, then there is no significant difference between them.

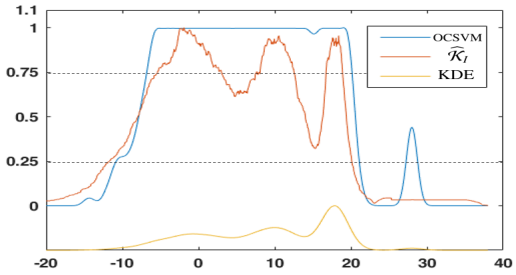

Figure 4: An one-dimensional dataset having three normal clusters of Gaussian distributions and one anomaly cluster (the small cluster on the right). They have a total of 1500+20 points. The bottom line shows the density distribution as estimated by a kernel density estimator (KDE) using Gaussian kernel (its scale is not shown on the y-axis.) The distributions of scores/similarities of OCSVM and $\widehat{\mathcal{K}}_I$ (scales in y-axis) are shown. The scores of OCSVM have been inversed and rescaled to $[0, 1]$ to be comparable to similarity.

6.2 OCSVM fails to detect local anomalies

Here we examine the abilities of $\widehat{\mathcal{K}}_I$ and OCSVM to detect local anomalies and clustered anomalies. The former is the type of anomalies located in close proximity to normal clusters; and the latter is the type of anomalies which formed a small cluster located far from all normal clusters.

Figure 4 shows the distributions of similarities of $\widehat{\mathcal{K}}_I$ and OCSVM on an one-dimensional dataset having three normal clusters of Gaussian distributions (of different variances) and a small group of clustered anomalies on the right. Note that OCSVM is unable to detect all anomalies located in close proximity to all three clusters, as all of these points have the same (or almost the same) similarity to points at the centers of these clusters. This outcome is similar to that using the density distribution (estimated by KDE) to detect anomalies because the densities of these points are not significantly lower than those of the peaks of low density clusters.

In addition, the clustered anomalies have higher similarities, as measured by OCSVM, than many anomalies at either fringes of the normal clusters. This means that the clustered anomalies are not included in the top-ranked anomalies.

If local anomalies are defined as points having similarities between 0.25 and 0.75, then the distribution of similarity of $\widehat{\mathcal{K}}_I$ shows that it detects all local anomalies surrounding all three normal clusters; and regards the clustered anomalies as global anomalies (having similarity < 0.25). In contrast, OCSVM fails to detect many

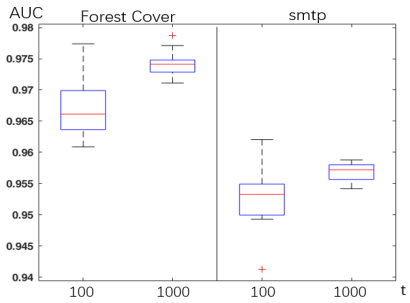

Figure 5: Boxplot of 10 runs of $\widehat{\mathcal{K}}_I$ on ForestCover and smtp. This result shows the increased stability of $\widehat{\mathcal{K}}_I$ predictions as t increases from 100 to 1000.

of the local anomalies detected by $\widehat{\mathcal{K}}_I$; and the clustered anomalies are regarded as local anomalies by OCSVM.

6.3 Stability analysis

This section provides an analysis to examine the stability of the scores of $\widehat{\mathcal{K}}_I$. Because $\widehat{\mathcal{K}}_I$ relies on random partitionings, it is important to determine how stable $\widehat{\mathcal{K}}_I$ is in different trials using different random seeds.

Figure 5 shows a boxplot of the scores produced by $\widehat{\mathcal{K}}_I$ over 10 trials. It shows that $\widehat{\mathcal{K}}_I$ becomes more stable as t increases.

6.4 Robust against contamination of anomalies in the training set

This section examines how robust an anomaly detector is against contamination anomalies in the training set. We use a 'cut-down' version of $\widehat{\mathcal{K}}_I$ which does not employ IDK, but Isolation Kernel only, i.e., $\|\Phi(x)\|$. $\|\Phi(x)\|$ is the base line in examining the robustness of $\widehat{\mathcal{K}}_I$ because $\Phi(x)$ is the basis in computing $\widehat{\mathcal{K}}_I$ (see Equation 10).

Figures 6 and 7 show two example comparisons between $\|\Phi(x)\|$ and $\widehat{\mathcal{K}}_I$ on the ForestCover and http datasets, where exactly the same hyperspheres are used in both detectors. The results show that $\|\Phi(x)\|$ is very unstable in a dataset with high contamination of anomalies (high γ). Despite using exactly the same $\|\Phi(x)\|$, $\widehat{\mathcal{K}}_I$ is very stable for all γ's. This is a result of the distributional characterisation of the entire dataset, described in Section 5.

7 RELATION TO ISOLATION-BASED ANOMALY DETECTORS

Both $\widehat{\mathcal{K}}_I$ and the existing isolation-based detector iNNE [2] employ the same partitioning mechanism. But the former is a distributional kernel and the latter is not.

iNNE employs a score which is a ratio of radii of two hyperspheres, designed to detect local anomalies [2]. The norm $\|\Phi(x)\|$ of Isolation Kernel, which is similar to the score of iNNE, simply counts the number of times x falls outside of a set of all hyperspheres, out of t sets of hyperspheres. This explains why both iNNE and $\|\Phi(x)\|$ have similar AUCs for all the datasets we used for point anomaly detection.

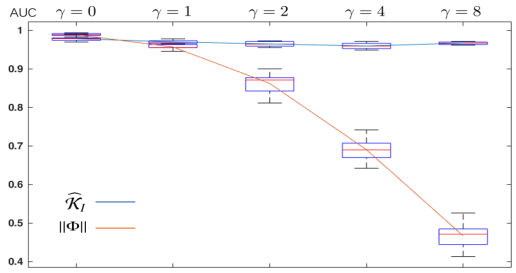

Figure 6: Boxplot of 10 runs of each of $\widehat{\mathcal{K}}_I$ and $\|\Phi(x)\|$ on ForestCover. The original anomaly ratio r is the ratio of the number of anomalies and the number of normal points in the given dataset. $\gamma \times r$ is used in the experiment to increase/decrease the anomalies in the given dataset. $\gamma = 1$ when the given dataset is used without modification; $\gamma = 0$ when no anomalies are used in the training process. $\gamma > 1$ has an increasingly higher chance of including anomalies in the training process.

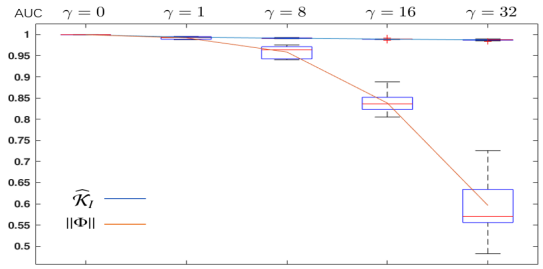

Figure 7: Boxplot of 10 runs of each of $\widehat{\mathcal{K}}_I$ and $\|\Phi(x)\|$ on http.

Table 4: Results of isolation-based anomaly detectors (AUC).

Dataset	$\widehat{\mathcal{K}}_I$	$\|\Phi(x)\|$	iNNE	iForest
speech	0.76	0.75	0.75	0.46
EEG_eye	0.88	0.87	0.87	0.58
PenDigits	0.98	0.96	0.96	0.93
MNIST_230	0.98	0.97	0.97	0.88
MNIST_479	0.86	0.60	0.86	0.45
mammograg	0.88	0.86	0.84	0.87
electron	0.80	0.78	0.79	0.80
shuttle	0.98	0.98	0.98	0.99
ALOI	0.82	0.82	0.82	0.55
muon	0.82	0.81	0.82	0.74
smtp	0.95	0.92	0.94	0.92
IoT_botnet	0.99	0.99	0.99	0.94
ForestCover	0.97	0.96	0.96	0.93
http	0.99	0.99	0.99	0.99
Average rank	1.50	2.75	2.43	3.32

In comparison with $\|\Phi(x)\|$ and iNNE, $\widehat{\mathcal{K}}_I$ has an additional distributional characterisation of the entire dataset. Table 4 shows that $\widehat{\mathcal{K}}_I$'s score based on this characterization leads to equal or

better accuracy than the scores used by both $\|\Phi(x)\|$ and iNNE. This is because the characterisation provides an effective reference for point anomaly detection, robust to contamination of anomalies in a dataset. This robustness is important when using points in a dataset which contains anomalies to build a model (that consists of hyperspheres used in $\widehat{\mathcal{K}}_I$, $\|\Phi(x)\|$ and iNNE.)

Since $\widehat{\mathcal{K}}_I$, iNNE [2] and iForest [8] have same time complexity $O(nt\psi)$, they have approximately the same runtime on the largest dataset http, as shown in Table 3.

In a nutshell, existing isolation-based anomaly detectors, i.e., iNNE and iForest, employ a score similar to the norm $\|\Phi(x)\|$. This is an interesting revelation because isolation-based anomaly detectors were never considered to be related to a kernel-based method before the current work. The power of isolation-based anomaly detectors can now be directly attributed to the norm of the feature map $\|\Phi(x)\|$ of Isolation Kernel.

From another perspective, $\widehat{\mathcal{K}}_I$ can be regarded as a member of the family of Isolation-based anomaly detectors [2, 8]; and iForest [8] has been regarded as one of the state-of-the-art point anomaly detectors [1, 6]. iNNE [2] is recently proposed to be an improvement of iForest. **$\widehat{\mathcal{K}}_I$ is the only isolation-based anomaly detector that is also a kernel based anomaly detector.**

The improvement of iNNE over iForest is mainly due to the use of a better isolating mechanism—overcoming four weaknesses of iForest [2]. The improvement of $\widehat{\mathcal{K}}_I$ over iNNE is mainly due to the distributional characterisation—it also contributes to $\widehat{\mathcal{K}}_I$ being the most robust isolation-based anomaly detector against contamination of anomalies in the training set.

In addition, the proposed method is the only detector that makes use of distributional kernel for point anomaly detection. While existing detector OCSMM [11] employs a distributional kernel, it is a group anomaly detector, not a point anomaly detector.

8 CONCLUSIONS

We show that Isolation Distributional Kernel addresses two key issues of kernel mean embedding, where the kernel employed has: (i) a feature map with intractable dimensionality which leads to high computational cost; and (ii) data independency which leads to poor detection accuracy in anomaly detection.

Our theoretical analyses reveal that a point kernel must be both data dependent and characteristic in order to produce an effective anomaly detector. Gaussian kernel, being characteristic but data independent, does not have the sufficient conditions.

We introduce a new Isolation Kernel and establish the geometrical interpretation of its feature map. This interpretation provides the insight that point anomalies and normal points are mapped into distinct regions in the feature space. This is the source of the power of $\widehat{\mathcal{K}}_I$. We also reveal that the distributional characterisation makes $\widehat{\mathcal{K}}_I$ robust to contamination of anomalies in a dataset.

Our evaluation shows that $\widehat{\mathcal{K}}_I$, without explicit learning, produces better detection accuracy than existing key kernel-based methods in detecting point anomalies, while achieving short testing and training times. In contrast, $\widehat{\mathcal{K}}_G$ and OCSVM, which employ Gaussian kernel, have lower detection accuracy in most datasets, and run up to three orders of magnitude slower in large datasets.

9 ACKNOWLEDGEMENTS

This research was supported by the National Key R&D Program of China (2018YFB1004300), NSFC (61921006), 111 Program (B14020) and JST CREST (JPMJCR1666).

REFERENCES

[1] Charu C. Aggarwal and Saket Sathe. 2017. *Outlier Ensembles: An Introduction.* Springer International Publishing.
[2] Tharindu R. Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, Ye Zhu, and Jonathan R. Wells. 2018. Isolation-based Anomaly Detection using nearest neighbour ensembles. *Computational Intelligence* 34, 4 (2018), 968–998.
[3] Vasili Baranau and Ulrich Tallarek. 2014. Random-close packing limits for monodisperse and polydisperse hard spheres. *Soft Matter* 10 (2014), 3826–3841.
[4] Liefeng Bo and Cristian Sminchisescu. 2009. Efficient Match Kernels Between Sets of Features for Visual Recognition. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*. 135–143.
[5] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* (2006), 1–30.
[6] Andrew Emmott, Shubhomoy Das, Thomas G. Dietterich, Alan Fern, and Weng-Keen Wong. 2016. A Meta-Analysis of the Anomaly Detection Problem. *CoRR* abs/1503.01158 (2016).
[7] Fukunaga Keinosuke. 1990. *Introduction to Statistical Pattern Recognition.* Academic Press, Chapter 6, 268–270.
[8] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*. 413–422.
[9] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Scholkopf. 2012. Learning from Distributions via Support Measure Machines. In *Advances in Neural Information Processing Systems*. 10–18.
[10] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard SchÃűlkopf. 2017. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning* 10 (1âĂŞ2) (2017), 1–141.
[11] Krikamol Muandet and Bernhard Schölkopf. 2013. One-class Support Measure Machines for Group Anomaly Detection. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. 449–458.
[12] Cameron Musco and Christopher Musco. 2017. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*. 3833–3845.
[13] Xiaoyu Qin, Kai Ming Ting, Ye Zhu, and Vincent Cheng Siong Lee. 2019. Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence*.
[14] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-scale Kernel Machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. 1177–1184.
[15] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Computing* 13, 7 (2001), 1443–1471.
[16] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. 2007. A Hilbert Space Embedding for Distributions. In *Algorithmic Learning Theory*, Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto (Eds.). Springer, 13–31.
[17] Chaoming Song, Ping Wang, and Hernan A. Makse. 2008. A phase diagram for jammed matter. *Nature* 453, 7195 (2008), 629–632.
[18] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. 2010. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research* 11 (2010), 1517–1561.
[19] Dougal J. Sutherland. 2016. *Scalable, Flexible and Active Learning on Distributions.* PhD Thesis, School of Computer Science, Carnegie Mellon University.
[20] Kai Ming Ting, Yue Zhu, and Zhi-Hua Zhou. 2018. Isolation Kernel and its effect on SVM. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2329–2337.
[21] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 2 (2009), 207–244.
[22] Christopher K. I. Williams and Matthias Seeger. 2001. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.). 682–688.
[23] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. Distance Metric Learning, with Application to Clustering with Side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*. 521–528.
[24] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. 2012. Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 476–484.
[25] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. 2016. Geometric mean metric learning. In *International Conference on Machine Learning*. 2464–2471.