Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering

Xiaoyu Qin  
Monash University  
Victoria, Australia 3800  
xiaoyu.qin@ieee.org

Kai Ming Ting  
Federation University  
Victoria, Australia 3842  
kaiming.ting@federation.edu.au

Ye Zhu  
Deakin University  
Victoria, Australia 3125  
ye.zhu@ieee.org

Vincent CS Lee  
Monash University  
Victoria, Australia 3800  
vincent.cs.lee@monash.edu

Abstract

A recent proposal of data dependent similarity called Isolation Kernel/Similarity has enabled SVM to produce better classification accuracy. We identify shortcomings of using a tree method to implement Isolation Similarity; and propose a nearest neighbour method instead. We formally prove the characteristic of Isolation Similarity with the use of the proposed method. The impact of Isolation Similarity on density-based clustering is studied here. We show for the first time that the clustering performance of the classic density-based clustering algorithm DBSCAN can be significantly uplifted to surpass that of the recent density-peak clustering algorithm DP. This is achieved by simply replacing the distance measure with the proposed nearest-neighbour-induced Isolation Similarity in DBSCAN, leaving the rest of the procedure unchanged. A new type of clusters called mass-connected clusters is formally defined. We show that DBSCAN, which detects density-connected clusters, becomes one which detects mass-connected clusters, when the distance measure is replaced with the proposed similarity. We also provide the condition under which mass-connected clusters can be detected, while density-connected clusters cannot.

Introduction

Similarity measure is widely used in various data mining and machine learning tasks. In clustering analysis, its impact to the quality of result is critical (Steinbach, Ertöz, and Kumar 2004). A recent proposal of data dependent similarity called Isolation Kernel has enabled SVM to produce better classification accuracy by simply replacing the commonly used data independent kernel (such as Gaussian kernel) with Isolation Kernel (Ting, Zhu, and Zhou 2018). This is made possible on datasets of varied densities because Isolation Kernel is adaptive to local data distribution such that two points in a sparse region are more similar than two points of equal inter-point distance in a dense region. Despite this success, the kernel characteristic has not been formally proven yet.

This paper extends this line of research by investigating a different implementation of Isolation Similarity. We provide a formal proof of the characteristic of the Isolation Similarity for the first time since its introduction. In addition, we focus on using Isolation Similarity to improve the clustering performance of density-based clustering.

This paper identifies shortcomings of tree-based method currently employed in inducing Isolation Similarities (Ting, Zhu, and Zhou 2018). Instead, we investigate a different method to induce the data dependent Isolation Similarity, and evaluate its clustering performance using DBSCAN (Ester et al. 1996) in comparison with the two existing improvements of density-based clustering, i.e., DScale (Zhu, Ting, and Angelova 2018) and DP (Rodriguez and Laio 2014).

The rest of the paper is organised as follows. We reiterate Isolation Kernel, identify the shortcomings of using the tree-based method to induce Isolation Similarity, provide the proposed alternative that employs a nearest neighbour method, the lemma and proof of the characteristic of Isolation Similarity, and the investigation in using Isolation Similarity in density-based clustering.

The descriptions of existing works are framed in order to clearly differentiate from the contributions we made here.

Isolation Kernel

Isolation Kernel/Similarity is first proposed by (Ting, Zhu, and Zhou 2018) as a new similarity which can adapt to density structure of the given dataset, as opposed to commonly used data independent kernels such as Gaussian and Laplacian kernels.

In the classification context, Isolation Kernel has been shown to be an effective means to improve the accuracy of SVM, especially in datasets which have varied densities in the class overlap regions (Ting, Zhu, and Zhou 2018). This is achieved by simply replacing the commonly used data independent kernel such as Gaussian and Laplacian kernels with the Isolation Kernel.

In the context of SVM classifiers, Isolation Kernel (Ting, Zhu, and Zhou 2018) has been shown to be more effective than existing approaches such as distance metric learning (Zadeh, Hosseini, and Sra 2016; Wang and Sun 2015), multiple kernel learning (Rakotomamonjy et al. 2008; Gönen and Alpaydin 2011) and Random Forest kernel (Breiman 2000; Davies and Ghahramani 2014).

The characteristic of Isolation Kernel is akin to one aspect of human-judged similarity as discovered by psychologists (Krumhansl 1978; Tversky 1977), i.e., human will judge the two same Caucasians as less similar when compared in Europe (which have many Caucasians) than in Asia.
We restate the definition and kernel characteristic (Ting, Zhu, and Zhou 2018) below.

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

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

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

(1)

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

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

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

(2)

The characteristic of Isolation Kernel is: two points in a sparse region are more similar than two points of equal inter-point distance in a dense region, i.e.,

**Characteristic of $K_{\psi}$:** $\forall x, y \in \mathcal{X}_S$ and $\forall x', y' \in \mathcal{X}_T$ such that $\|x - y\| = \|x' - y'\|$, $K_{\psi}$ satisfies the following condition:

$$K_{\psi}(x, y) > K_{\psi}(x', y')$$

(3)

where $\mathcal{X}_S$ and $\mathcal{X}_T$ are two subsets of points in sparse and dense regions of $\mathbb{R}^d$, respectively; and $\|x - y\|$ is the distance between $x$ and $y$.

To get the above characteristic, the required property of the space partitioning mechanism is to create large partitions in the sparse region and small partitions in the dense region such that two points are more likely to fall into the same partition in a sparse region than two points of equal inter-point distance in a dense region.

## Shortcomings of tree-based isolation partitioning

Isolation Kernel (Ting, Zhu, and Zhou 2018) employs isolation trees or iForest (Liu, Ting, and Zhou 2008) to measure the similarity of two points because its space partitioning mechanism produces the required partitions which have volumes that are monotonically decreasing wrt the density of the local region.

Here we identify two shortcomings in using isolation trees to measure Isolation Similarity, i.e., each isolation tree (i) employs axis-parallel splits; and (ii) is an imbalanced tree.

Figure 1(a) shows an example partitioning due to axis-parallel splits of an isolation tree. The tree-based isolating partitions generally satisfy the requirement of small partitions in dense region and large partitions in sparse region. However, it produced some undesirable effect, i.e., some partitions are always overextended for the first few splits close to the root of an imbalanced tree. These are manifested as elongated rectangles in Figure 1(a).

While using balanced trees can be expected to overcome this problem, the restriction to hyper-rectangles remains due to the use of axis-parallel splits.

To overcome these shortcomings of isolation trees, we propose to use a nearest neighbour partitioning mechanism which creates a Voronoi diagram (Aurenhammer 1991) where each cell is an isolating partition (i.e., isolating one point from the rest of the points in the given sample). An example is provided in Figure 1(b). Note that these partitions also satisfy the requirement of small partitions in the dense region and large partitions in the sparse region. But they do not have the undesirable effect of elongated rectangles. We provide our implementation in the next section.

### Nearest neighbour-induced Isolation Similarity

Instead of using trees in its first implementation (Ting, Zhu, and Zhou 2018), we propose to implement Isolation Similarity using nearest neighbours.

Like the tree method, the nearest neighbour method also produces each $H$ model which consists of $\psi$ isolating partitions $\theta$, given a subsample of $\psi$ points. Rather than representing each isolating partition as a hyper-rectangle, it is represented as a cell in a Voronoi diagram (Aurenhammer 1991), where the boundary between two points is the equal distance from these two points.

While the Voronoi diagram is nothing new, its use in measuring similarity is new.

Using the same notations as used earlier, $H$ is now a Voronoi diagram, built by employing $\psi$ points in $D$, where each isolating partition or Voronoi cell $\theta \in H$ isolates one

---

1Imbalanced are a necessary characteristic of isolation trees (Liu, Ting, and Zhou 2008) for their intended purpose of detecting anomalies, where anomalies are expected to be isolated with few splits; and normal points can only be isolated using a large number of splits.
data point from the rest of the points in \( D \). We call the point which determines a cell as the cell centre.

Given a Voronoi diagram \( H \) constructed from a sample \( D \) of \( \psi \) points, the Voronoi cell centred at \( z \in D \) is:

\[
\theta[z] = \{ x \in \mathbb{R}^d \mid x = \arg\min_{z \in D} \ell_p(x - z) \}
\]

where \( \ell_p(x, y) \) is a distance function and we use \( p = 2 \) as Euclidean distance in this paper.

**Definition 2** For any two points \( x, y \in \mathbb{R}^d \), the nearest neighbour-induced Isolation Similarity of \( x \) and \( y \) wrt \( D \) is defined to be the expectation taken over the probability distribution on all Voronoi diagrams \( H \in \mathcal{H}_\psi(D) \) that both \( x \) and \( y \) fall into the same Voronoi cell \( \theta \in H \):

\[
K_\psi(x, y \mid D) = \mathbb{E}_{H \sim \mathcal{H}_\psi(D)}[\mathbb{I}(x, y \in \theta[z] \mid \theta[z] \in H)] = \mathbb{E}_{z \sim \mathcal{D}_\psi}[\mathbb{I}(x, y \in \theta[z] \mid z \in \mathcal{D})] = P(x, y \in \theta[z] \mid z \in D \subseteq \mathcal{D}) \tag{4}
\]

where \( P \) denotes the probability.

The Voronoi diagram has the required property of the space-partitioning mechanism to produce large partitions in a sparse region and small partitions in a dense region. This yields the characteristic of Isolation Similarity: **two points in a sparse region are more similar than two points of equal inter-point distance in a dense region**.

The use of nearest neighbours facilitates a proof of the above characteristic that was previously hampered by the use of trees. We provide the proof in the next section.

**Lemma and Proof of the characteristic of Isolation Similarity**

Let \( \rho(x) \) denote the density at point \( x \), a lemma based on definition 4 is given below:

**Lemma 1** \( \forall x, y \in \mathcal{X}_S \) (sparse region) and \( \forall x', y' \in \mathcal{X}_T \) (dense region) such that \( \forall z \in \mathbb{R}^d, \rho(z) < \rho(z') \), the nearest neighbour-induced Isolation Similarity \( K_\psi \) has the characteristic that for \( \ell_p(x - y) = \ell_p(x' - y') \) implies

\[
P(x, y \in \theta[z]) > P(x', y' \in \theta[z']) \iff K_\psi(x, y \mid D) > K_\psi(x', y' \mid D)
\]

Sketch of the proof: (i) If two points fall into the same Voronoi cell, then the distances of individual points to this cell centre must be shorter than those to every other cell centre (or at most equal to that to one other cell centre) in a Voronoi diagram formed by all these cell centres. (ii) In a subset of \( \psi \) points, sampled from \( D \), used to form a Voronoi diagram, the probability of two points falling into the same Voronoi cell can then be estimated based on the condition stated in (i). (iii) The probability of two points of equal inter-point distance falling into the same Voronoi cell is a monotonically decreasing function wrt the density of the cell.

**Proof 1** Let a local region \( V(x, y) \) covering both \( x \) and \( y \) as a ball centred at the middle between \( x \) and \( y \) having \( \ell_p(x, y) \) as the diameter of the ball. Assume that the density in \( V(x, y) \) is uniform and denoted as \( \rho(V(x, y)) \).

Let \( N_\epsilon(x) \) be the \( \epsilon \)-neighbourhood of \( x \), i.e., \( N_\epsilon(x) = \{ y \in \mathcal{D} \mid \ell_p(x, y) \leq \epsilon \} \). The probability of both \( x \) and \( y \) in the same Voronoi cell \( \theta[z] \) is equivalent to the probability of a point \( z \in \mathcal{D} \) being the nearest neighbour of both \( x \) and \( y \) wrt all other points in \( \mathcal{D} \), i.e., the probability of selecting \( \psi - 1 \) points which are all located outside the region \( U(x, y, z) \), where \( U(x, y, z) = N_\epsilon(x) \cup N_\epsilon(y) \).

To simplify notation, \( z \in \mathcal{D} \) is omitted. Then the probability of \( x, y \in \theta[z] \) can be expressed as follows:

\[
P(x, y \in \theta[z] \mid z \in V(x, y)) = P(z_1, z_2, \ldots, z_{\psi - 1} \notin U(x, y, z)) \propto (1 - \mathbb{E}_{z \sim V(x, y)}[|U(x, y, z)|]/|D|^{(\psi - 1)})
\]

where \( |W| \) denotes the cardinality of \( W \).

Assume that \( U(x, y, z) \) is also uniformly distributed, having the same density \( \rho(V(x, y)) \), the expected value of \( |U(x, y, z)| \) can be estimated as:

\[
\mathbb{E}_{z \sim V(x, y)}[|U(x, y, z)|] = \mathbb{E}_{z \sim V(x, y)}[v(U(x, y, z)) \times \rho(V(x, y))] = \mathbb{E}_{z \sim V(x, y)}[v(U(x, y, z))] \times \rho(V(x, y))
\]

where \( v(W) \) denotes the volume of \( W \).

Thus, we have

\[
P(x, y \in \theta[z] \mid z \in V(x, y)) \propto \left(1 - \mathbb{E}_{z \sim V(x, y)}[v(U(x, y, z))] \times \frac{\rho(V(x, y))}{|D|^{(\psi - 1)}}\right)
\]

In other words, the higher the density in the area around \( x \) and \( y \), the smaller \( P(x, y \in \theta[z] \mid z \in V(x, y)) \) is, as the volume of \( V(x, y) \) is constant given \( x \) and \( y \).

Given two pairs of points from two different regions but of equal interpoint distance as follows: \( \forall x, y \in \mathcal{X}_S \) (sparse region) and \( \forall x', y' \in \mathcal{X}_T \) (dense region) such that \( \ell_p(x, y) = \ell_p(x', y') \).

Assume that data are uniformly distributed in both regions, and we sample \( z, z' \in D \) from \( D \) such that \( z \in V(x, y) \) and \( z' \in V(x', y') \). We have

\[
\mathbb{E}_{z \sim V(x, y)}[v(U(x, y, z))] = \mathbb{E}_{z' \sim V(x', y')}[v(U(x', y', z'))]
\]

because the volume of \( V(x, y) \) is equal to that of \( V(x', y') \) for \( \ell_p(x, y) = \ell_p(x', y') \), independent of the density of the region.

Supposing that we choose a sufficient large sample size \( \psi \) of \( D \) which contains points from both \( V(x, y) \) and \( V(x', y') \). When the data are uniformly distributed in \( U(x, y, z) \in \mathcal{X}_S \) and \( U(x', y', z') \in \mathcal{X}_T \), based on Equation 5, we have

\[
P(x, y \in \theta[z] \mid z \in V(x, y)) > P(x', y' \in \theta[z'] \mid z' \in V(x', y')) = K_\psi(x, y \mid D) > K_\psi(x, y \mid D)
\]

This means that \( x' \) and \( y' \) (in a dense region) are more like to be in different cells than \( x \) and \( y \) in \( V(x, y) \) (in a sparse region), as shown in Figure 1. \( \square \)
The rate of decrease of $P$

1. To be consistent with the concept of distance as a kind of similarity density-based clustering and its counterpart called neighbourhood density function commonly used in the density-based clustering in the rest of the paper. We describe change slower in area far from the centre point and faster in $\ell$ of tour maps of $p$

Figure 2: (a) Reference points used in the simulations, where inter-point distance $\| x - y \| = \| x' - y' \|$ increases. Simulation results as $\psi$ increases (b); and as inter-point distance increases (c). $t = 10000$ is used.

A simulation validating the above analysis is given in Figure 2. It compares $P(x, y \in \theta_S)$ and $P(x', y' \in \theta_T)$ when $x, y$ from a sparse region and $x', y'$ from a dense region with equal inter-point distance. Given a fixed $\psi < |D|$ or a fixed inter-point distance, properties observed from Figure 2 are given as follows:

1. $P(x, y \in \theta_S) > P(x', y' \in \theta_T)$.
2. The rate of decrease of $P(x', y' \in \theta_T)$ is faster than that of $P(x, y \in \theta_S)$. Thus $P(x', y' \in \theta_T)$ reaches 0 earlier.

Isolation Dissimilarity and contour maps

To be consistent with the concept of distance as a kind of dissimilarity, we use Isolation Dissimilarity hereafter:

Isolation dissimilarity: $p_i(x, y) = 1 - K_\psi(x, y)$.

Like $\ell_p$ norm, $\forall x, p_i(x, x) = 0$ and $p_i(x, y) = p_i(y, x)$. However, $\forall x \neq y$, $p_i(x, y)$ depends on the data distribution and how $p_i$ is implemented, not the geometric positions only.

We denote the nearest-neighbour-induced Isolation Dissimilarity $p_i$-aNNE; and the tree-induced version $p_i$-iForest. An example comparison of the contour maps produced the two dissimilarities are given in Figure 3. Note that the contour maps of $p_i$ depend on the data distribution, whereas that of $\ell_2$ is not. Also, comparing to $\ell_2$, the other dissimilarity change slower in area far from the centre point and faster in area close to the centre point.

We examine the impact of Isolation Dissimilarity on density-based clustering in the rest of the paper. We describe the neighbourhood density function commonly used in the density-based clustering and its counterpart called neighbourhood mass function in the next section.

Neighbourhood density and mass functions

Neighbourhood mass function (Ting et al. 2016) was first proposed as a way to model data in terms of mass distribution, analogous to the neighbourhood density function used in modelling data as a density distribution. The key difference is the measure $\partial$ used in the following function of $x$: $\# \{ y \in D \mid \partial(x, y) \leq \text{cutoff} \}$, where the boundary of the region within which the points are counted is set by a user-specified constant cutoff. When a distance measure is used, it becomes a neighbourhood density function as the ball has a fixed volume when the cutoff is fixed. It is a neighbourhood mass function when a data dependent dissimilarity is used as the volume of the ‘ball’ varies depending on the local distribution even if the cutoff is fixed.

Mass-connected clusters

Here we define mass-connected clusters defined in terms of neighbourhood mass function:

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

Definition 3 Using an $\alpha$-neighbourhood mass estimator $M_\alpha(x) = \# \{ y \in D \mid p_i(x, y) \leq \alpha \}$, mass-connectivity with threshold $\tau$ between $x_1$ and $x_p$ via a sequence of $p$ unique points from $D$, i.e., \{ $x_1, x_2, x_3, \ldots, x_p$ \} is denoted as $MConnect_\alpha^p(x_1, x_p)$, and it is defined as:

$$MConnect_\alpha^p(x_1, x_p) \leftrightarrow$$

$$[(p_i(x_1, x_2) \leq \alpha) \land (M_\alpha(x_1) \geq \tau) \lor (M_\alpha(x_2) \geq \tau))]$$

$$\lor \exists \{x_1, x_2, \ldots, x_p\} \{ \forall i \in \{2, \ldots, p\} p_i(x_{i-1}, x_i) \leq \alpha \}$$

$$\land \forall i \in \{2, \ldots, p-1\} M_\alpha(x_i) \geq \tau)]$$

(6)

The second line denotes direct connectivity between two neighbouring points when $p = 2$. The last two lines denote transitive connectivity when $p > 2$.

Definition 4 A mass-connected cluster $\tilde{C}$, which has a mode $c = \arg \max_{x \in \tilde{C}} M_\alpha(x)$, is a maximal set of points that are mass-connected with its mode, i.e., \$ \tilde{C} = \{ x \in D \mid MConnect_\alpha^p(x, c) \} \$. 

4758
Note that density-connectivity and density-connected clusters are similarly defined in DBSCAN (Ester et al. 1996) when \( M_\alpha = \# \{ y \in D \mid p_\alpha(x, y) \leq \alpha \} \) is replaced with \( N_\epsilon = \# \{ y \in D \mid \ell_\epsilon(x, y) \leq \epsilon \} \) in the above two definitions. In other words, DBSCAN (Ester et al. 1996) which uses \( N_\epsilon \) detects density-connected clusters; whereas DBSCAN which uses \( M_\alpha \) detects mass-connected clusters.

The only difference between a density-connected cluster and a mass-connected cluster is the dissimilarity measure used in Equation 6. We called the DBSCAN procedure which employs \( M_\alpha \): MBSCAN, since it detects mass-connected clusters rather than density-connected clusters.

**Condition under which MBSCAN detects all mass-connected clusters**

Let a valley between two cluster modes be the points having the minimum estimated \( M_\alpha(\cdot) \), i.e., \( \hat{g}_{ij} \), along any path linking cluster modes \( c_i \) and \( c_j \). A path between two points \( x \) and \( y \) is non-cyclic linking a sequence of unique points starting with \( x \) and ending with \( y \) where adjacent points lie in each other’s \( \alpha \)-neighbourhood: \( p_\alpha(\cdot, \cdot) \leq \alpha \).

Because \( p_\alpha \) (unlike \( \ell_\epsilon \) used in \( N_\epsilon \)) is adaptive to the density of local data distribution, it is possible to adjust \( \psi \) and \( \alpha \) to yield an \( M_\alpha \) distribution such that all valley-points have close enough small values, if there exist such \( \psi \) and \( \alpha \).

In other words, for some data distributions \( F \), there exist some \( \psi \) and \( \alpha \) such that the distribution of \( M_\alpha(\cdot) \) satisfies the following condition:

\[
\min_{k \in \{1, \ldots, N\}} M_\alpha(c_k) > \max_{i \neq j \in \{1, \ldots, N\}} \hat{g}_{ij} \tag{7}
\]

where \( \hat{g}_{ij} \) is the largest of the minimum estimated \( M_\alpha(\cdot) \) along any path linking cluster modes \( c_i \) and \( c_j \).

In data distributions \( F \), MBSCAN is able to detect all mass-connected clusters because a threshold \( \tau \) can be used to breaks all paths between the modes by assigning regions with estimated \( M_\alpha(\cdot) \) less than \( \tau \) to noise, i.e.,

\[
\exists \tau \forall k, i \neq j \in \{1, \ldots, N\} M_\alpha(c_k) \geq \tau > \hat{g}_{ij}
\]

An example that \( F \) subsumes \( G \), derived from the same dataset, is shown in Figure 4, where a hard distribution \( G \) in which DBSCAN fails to detect all clusters is shown in Figure 4(a); but MBSCAN succeeds\(^2\).

In other words, the mass distribution afforded by \( M_\alpha \) is more flexible than the density distribution generated by \( N_\epsilon \), which leads directly to MBSCAN’s enhanced cluster detection capability in comparison with DBSCAN, though both are using exactly the same algorithm, except the dissimilarity.

Figure 5 shows the change of neighbourhood function values wrt the change in their parameter for \( N_\epsilon \) using \( \ell_2 \) and \( M_\alpha \) using \( p_\alpha \)-aNNE. This example shows that no \( \epsilon \) exists which enables DBSCAN to detect all three clusters. This is because the line for Peak#3 (which is the mode of the sparse

\[^2\text{Note that the above condition was first described in the context of using mass-based dissimilarity (Ting et al. 2018); but not in relation to mass-connected clusters. We have made the relation to mass-connected clusters more explicitly here.}

\[^3\text{The artificial datasets are from http://cs.uef.fi/sipu/datasets/}

![Figure 4](https://example.com/figure4.png)

**Figure 4:** (a) A hard distribution for DBSCAN as estimated by \( N_\epsilon \), where DBSCAN (which uses \( N_\epsilon \)) fails to detect all clusters using a threshold. (b) The distribution estimated by \( M_\alpha \) from the same dataset, where MBSCAN (which uses \( M_\alpha \)) succeeds in detecting all clusters using a threshold.

![Figure 5](https://example.com/figure5.png)

**Figure 5:** Change of neighbourhood density/mass wrt its parameter. \( N_\epsilon \) uses \( \ell_2 \); and \( M_\alpha \) uses \( p_\alpha \)-aNNE. The peak numbers and valley numbers refer to those shown in Figure 4(a).
Table 1: Clustering results in $F_1$ scores. The best performer is boldfaced; the second best is underlined.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DP</th>
<th>DBSCAN</th>
<th>MBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\ell_2$</td>
<td>$\ell_2$</td>
<td>ReScale</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DScale</td>
</tr>
<tr>
<td>Artificial data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aggregation</td>
<td>0.997</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>compound</td>
<td>0.867</td>
<td>0.791</td>
<td>0.942</td>
</tr>
<tr>
<td>Jain</td>
<td>0.976</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>pathbased</td>
<td>0.943</td>
<td>0.828</td>
<td>0.987</td>
</tr>
<tr>
<td>hard distribution</td>
<td>0.994</td>
<td>0.667</td>
<td>0.995</td>
</tr>
<tr>
<td>High-dimensional data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALLAML</td>
<td>0.706</td>
<td>0.484</td>
<td>0.484</td>
</tr>
<tr>
<td>COIL20</td>
<td>0.724</td>
<td>0.842</td>
<td>0.861</td>
</tr>
<tr>
<td>Human Activity</td>
<td>0.595</td>
<td>0.331</td>
<td>0.352</td>
</tr>
<tr>
<td>Isolet</td>
<td>0.517</td>
<td>0.194</td>
<td>0.234</td>
</tr>
<tr>
<td>lung</td>
<td>0.703</td>
<td>0.489</td>
<td>0.544</td>
</tr>
<tr>
<td>TOX 171</td>
<td>0.519</td>
<td>0.336</td>
<td>0.403</td>
</tr>
<tr>
<td>General data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast</td>
<td>0.970</td>
<td>0.824</td>
<td>0.951</td>
</tr>
<tr>
<td>control</td>
<td>0.736</td>
<td>0.531</td>
<td>0.663</td>
</tr>
<tr>
<td>gsp</td>
<td>0.811</td>
<td>0.753</td>
<td>0.811</td>
</tr>
<tr>
<td>iris</td>
<td>0.967</td>
<td>0.848</td>
<td>0.905</td>
</tr>
<tr>
<td>seeds</td>
<td>0.909</td>
<td>0.750</td>
<td>0.885</td>
</tr>
<tr>
<td>shape</td>
<td>0.761</td>
<td>0.581</td>
<td>0.680</td>
</tr>
<tr>
<td>thyroid</td>
<td>0.868</td>
<td>0.584</td>
<td>0.850</td>
</tr>
<tr>
<td>WDBC</td>
<td>0.933</td>
<td>0.600</td>
<td>0.765</td>
</tr>
<tr>
<td>wine</td>
<td>0.933</td>
<td>0.645</td>
<td>0.866</td>
</tr>
<tr>
<td>Grand Average</td>
<td>0.823</td>
<td>0.653</td>
<td>0.760</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>$k \in [2...40]$</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>$\text{MinPts} \in [2...40]$</td>
</tr>
<tr>
<td>MBSCAN</td>
<td>$\text{MinPts} \in [2...40]$</td>
</tr>
<tr>
<td>ReScale</td>
<td>$\text{Precision factor} \in [0.05...0.95]$</td>
</tr>
<tr>
<td>DScale</td>
<td>$\text{Neighbourhood size} \in [0.001...0.999]$</td>
</tr>
<tr>
<td>iForest</td>
<td>$\text{Ensemble size} \in [2...[n/2]]$</td>
</tr>
</tbody>
</table>

They are from three categories: 5 artificial datasets, 6 high-dimensional datasets, and 9 general datasets. They are selected because they represent diverse datasets in terms of data size, number of dimensions and number of clusters. The data characteristics of these datasets are shown in the first four columns of Table 1. All datasets are normalised using the min-max normalisation so that each attribute is in $[0,1]$ before the experiments begin.

We compared all clustering results in terms of the best $F_1$ score (Rijsbergen 1979) that is obtained from a search of the algorithm’s parameter. We search each parameter within a reasonable range. The ranges used for all algorithms/dissimilarities are provided in Table 2. Because $p_i$ used in MBSCAN is based on randomised methods, we report the mean $F_1$ score over 10 trials for each dataset.

Table 2: Search ranges of parameters used.

<table>
<thead>
<tr>
<th>Description</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>$k \in [2...40]$</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>$\text{MinPts} \in [2...40]$</td>
</tr>
<tr>
<td>MBSCAN</td>
<td>$\text{MinPts} \in [2...40]$</td>
</tr>
<tr>
<td>ReScale</td>
<td>$\text{Precision factor} \in [0.05...0.95]$</td>
</tr>
<tr>
<td>DScale</td>
<td>$\text{Neighbourhood size} \in [0.001...0.999]$</td>
</tr>
<tr>
<td>iForest</td>
<td>$\text{Ensemble size} \in [2...[n/2]]$</td>
</tr>
</tbody>
</table>

* $f$ parameter is required for ReScale only.
* A search of 10 values with equal interval in the range.

Clustering results

Table 1 shows that MBSCAN using $p_i$-aNNE has the best performance overall. Its $F_1$ scores are the best on 14 out of 20 datasets. The closest contender DP has the best $F_1$ scores on 5 datasets only. In two other performance measures, MBSCAN using $p_i$-aNNE has 13 wins, 2 draws and 5 losses.
against DP; and has higher average $F_1$ score too (0.867 versus 0.823).

One notable standout is on the high-dimensional datasets: MBSCAN using $p_r$-aNNE has the largest gap in average $F_1$ score in comparison with other contenders among the three categories of datasets. With reference to DBSCAN, the gap is close to 0.3 $F_1$ score; even compare with the closest contender DP, the gap is 0.1 $F_1$ score. The superiority of $p_r$-aNNE over $p_r$-iForest is also highlighted on these high-dimensional datasets.

MBSCAN using $p_r$-iForest wins over the original DBSCAN on all datasets except one (aggregation). This version of MBSCAN uplifted the clustering performance of DBSCAN significantly to almost the same level of DP.

A significance test is conducted over MBSCAN with $p_r$-aNNE, MBSCAN with $p_r$-iForest and DP. Figure 6 shows the result of the test—MBSCAN using $p_r$-aNNE performs the best and is significantly better than DP and MBSCAN using $p_r$-iForest; and there is no significant difference between DP and MBSCAN using $p_r$-iForest.

Figure 6: Critical difference (CD) diagram of the post-hoc Nemenyi test ($\alpha = 0.10$). A line is showed between two algorithms if their rank gap is smaller than CD; otherwise, the difference is significant.

DBSCAN is known to be sensitive to its parameter settings. As MBSCAN is using exactly the same algorithm, it has same sensitivity.

A caveat is in order. Although Isolation Similarity consistently outperforms distance measure in DBSCAN, our preliminary experiment using DP shows that the result is mixed. An analysis of DP, similar to that provided in this paper, is required in order to ascertain the condition(s) under which DP performs well and Isolation Similarity can help.

Complexity and runtime

MBSCAN with $p_r$-aNNE is much faster than MBSCAN with $p_r$-iForest because the time complexity to build a maximum-size isolation tree and testing one point is $O(\psi^2)$; and aNNE takes $O(\psi)$. The space complexity to store trained aNNE model is $O(t \cdot \psi)$; and that of iForest is $O(t \cdot \psi \cdot \log \psi)$.

Table 3 shows the GPU runtime results on the four largest datasets. In contrast to DBSCAN and DP, MBSCAN needs to pre-compute the dissimilarity matrix in the pre-processing step, and this takes $O(n^2)$ time. This pre-processing constitutes the most of the time of MBSCAN reported in Table 3. $p_r$-aNNE is still faster than ReScale in high-dimensional datasets, though it is one order of magnitude slower than DBSCAN and DP.

### Table 3: Runtime in GPU seconds

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DP</th>
<th>DBSCAN</th>
<th>MBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>ReScale</td>
<td>iForest</td>
</tr>
<tr>
<td>Hard dist.</td>
<td>0.11</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>COIL20</td>
<td>0.03</td>
<td>0.02</td>
<td>3.74</td>
</tr>
<tr>
<td>Human Act.</td>
<td>0.10</td>
<td>0.03</td>
<td>2.12</td>
</tr>
<tr>
<td>Isolet</td>
<td>0.08</td>
<td>0.03</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Summary: aNNE versus iForest implementations of Isolation Similarity

We find that the aNNE implementation of Isolation Similarity is better than the iForest implementation because the aNNE implementation is more:

- Effective on datasets with varied densities and high-dimensional datasets.
- Amenable to GPU acceleration because aNNE can be implemented in almost pure matrix manipulations. Thus, aNNE runs many orders of magnitude faster than iForest if a large $\psi$ is required because aNNE in the GPU implementation has almost constant runtime wrt $\psi$.
- Stable because aNNE’s randomisation is a result of sampling data subsets only.

Conclusions

We make four contributions in this paper:

1) Identifying shortcomings of tree-induced Isolation Similarity; proposing a nearest neighbour-induced Isolation Similarity to overcome these shortcomings; and establishing three advantages of the nearest neighbour-induced Isolation Similarity over the tree-induced one.

2) Formally proving the characteristic of the nearest neighbour-induced Isolation Similarity. This is the first proof since the introduction of Isolation Kernel (Ting, Zhu, and Zhou 2018).

3) Providing a formal definition of mass-connected clusters and an explanation why detecting mass-connected clusters is a better approach in overcoming the shortcoming of DBSCAN (which detects density-connected clusters) in datasets with varied densities. This differs fundamentally from the existing density-based approaches of the original DBSCAN, DP and ReScale which all employ a distance measure to compute density.

4) Conducting an empirical evaluation to validate the advantages of (i) nearest-neighbour-induced Isolation Similarity over tree-induced Isolation Similarity; and (ii) mass-based clustering using Isolation Similarity over four density-based clustering algorithms, i.e., DBSCAN, DP, ReScale and DScale.

In addition, we show for the first time that it is possible to uplift the clustering performance of the classic DBSCAN, through the use of nearest-neighbour-induced Isolation Similarity, to surpass that of DP—the state-of-the-art density-based clustering algorithm.
Acknowledgements

This material is based upon work supported by eSolutions of Monash University (Xiaoyu Qin); and partially supported by the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AOARD) under award number: FA2386-17-1-4034 (Kai Ming Ting).

References


Chang, H., and Yeung, D.-Y. 2008. Robust path-based spec-


