Two applications of Min-Max-Jump distance

Anonymous Author(s)
Affiliation
Address
email

Abstract

We explore two applications of Min-Max-Jump distance (MMJ distance): MMJ-based K-means and MMJ-based internal clustering evaluation index. K-means and its variants are possibly the most popular clustering approach. A key drawback of K-means is that it cannot deal with data sets that are not the union of well-separated, spherical clusters. MMJ-based K-means proposed in this paper overcomes this demerit of K-means, so that it can handle irregularly shaped clusters. Evaluation (or "validation") of clustering results is fundamental to clustering and thus to machine learning. Popular internal clustering evaluation indices like Silhouette coefficient, Davies–Bouldin index, and Calinski-Harabasz index performs poorly in evaluating irregularly shaped clusters. MMJ-based internal clustering evaluation index uses MMJ distance and Semantic Center of Mass (SCOM) to revise the indices, so that it can evaluate irregularly shaped data. An experiment shows introducing MMJ distance to internal clustering evaluation index, can systematically improve the performance. We also devise two algorithms for calculating MMJ distance.

1 Introduction

Distance is a numerical measurement of how far apart objects or points are. It is usually formalized in mathematics using the notion of a metric space. A metric space is a set together with a notion of distance between its elements, usually called points. The distance is measured by a function called a metric or distance function. Metric spaces are the most general setting for studying many of the concepts of mathematical analysis and geometry.

In this paper, we introduce two algorithms for calculating Min-Max-Jump distance (MMJ distance) and explore two applications of it. Including MMJ-based K-means (MMJ-K-means) and MMJ-based internal clustering evaluation index.

MMJ-K-means improves K-means, so that it can handle irregularly shaped clusters. We claim MMJ-CH is the SOTA (state-of-the-art) internal clustering evaluation index, which achieves an accuracy of $90/145$. MMJ-CH is one of the MMJ-based internal clustering evaluation indices.

2 RELATED WORK

2.1 Different distance metrics

Many distance measures have been proposed in literature, such as Euclidean distance or cosine similarity. These distance measures often be found in algorithms like k-NN, UMAP, HDBSCAN, etc. The most common metric is Euclidean distance. Cosine similarity is often used as a way to counteract Euclidean distance's problem in high dimensionality. The cosine similarity is the cosine of the angle between two vectors.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Hamming distance is the number of values that are different between two vectors. It is typically used to compare two binary strings of equal length (1).

Manhattan distance is a geometry whose usual distance function or metric of Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates (2).

Chebyshev distance is defined as the greatest of difference between two vectors along any coordinate dimension (3).

Minkowski distance or Minkowski metric is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance (4).

Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets (5).

Haversine distance is the distance between two points on a sphere given their longitudes and latitudes. It is similar to Euclidean distance in that it calculates the shortest path between two points. The main difference is that there is no straight line, since the assumption is that the two points are on a sphere (6).

2.2 K-means

K-means (7) and its variants (8; 9; 10) are possibly the most well-liked clustering approach. K-means divides the data into K groups, where K is a hyper-parameter to be optimized. It aims to reduce the within-cluster dissimilarity. While popular, K-means and its variants perform poorly for data sets that are not the union of well-separated, spherical clusters. MMJ-based K-means (MMJ-K-means) proposed in this paper overcomes this demerit of K-means, so that it can handle irregularly shaped clusters.

2.3 Internal clustering evaluation index

Evaluation (or "validation") of clustering results is as difficult as the clustering itself (11). Popular approaches involve "internal" evaluation and "external" evaluation. In internal evaluation, a clustering result is evaluated based on the data that was clustered itself. Popular internal evaluation indices are Davies-Bouldin index (12), Silhouette coefficient (13), Dunn index (14), and Calinski-Harabasz index (15) etc. In external evaluation, the clustering result is compared to an existing "ground truth" classification, such as the Rand index (16). However, knowledge of the ground truth classes is almost never available in practice.

In Section 5.2, an experiment shows introducing Min-Max-Jump (MMJ) distance to internal clustering evaluation index, can systematically improve the performance.

2.4 Path-based distances

Euclidean distances are frequently used in machine learning and clustering methods to compare points. However, the distance is data-independent, and not tailored to the geometry of the data. Many metrics that are data-dependent have been devised, such as diffusion distances (17) and path-based distances (18; 19). MMJ distance is a path-based distance.

3 Definition of Min-Max-Jump

Definition 1. *Min-Max-Jump distance (MMJ distance)*

Ω is a set of points (at least one). For any pair of points $p, q \in \Omega$, the distance between p and q is defined by a distance function d(p,q) (such as Euclidean distance). $i, j \in \Omega$, $\Psi_{(i,j,n,\Omega)}$ is a path from point i to point j, which has length of n points (see Table 1). $\Theta_{(i,j,\Omega)}$ is the set of all paths from point i to point j. Therefore, $\Psi_{(i,j,n,\Omega)} \in \Theta_{(i,j,\Omega)}$. $max\_jump(\ \Psi_{(i,j,n,\Omega)}\ )$ is the maximum jump in path $\Psi_{(i,j,n,\Omega)}$.

The Min-Max-Jump distance between a pair of points i, j, which belong to Ω, is defined as:

Table 1: Table of notations

Ω	A set of N points, with each point indexed from 1 to N;
$\Omega_{[1,n]}$	The first n points of Ω, indexed from 1 to n;
Ω_{n+1}	The $(n+1)$th point of Ω;
C_i	A cluster of points that is a subset of Ω;
ξ_i	One-SCOM of C_i;
Ω + p	Set Ω plus one new point p. Since $p \notin \Omega$, if Ω has N points, this new set now has $N+1$ points;
$\Psi_{(i,j,n,\Omega)}$	$\Psi_{(i,j,n,\Omega)}$ is a sequence from point i to point j, which has length of n points. All the points in the sequence must belong to set Ω. That is to say, it is a path starts from i, and ends with j. For convenience, the path is not allowed to have loops, unless the start and the end is the same point;
$d(i,j)$	$d(i,j)$ is a distance metric between pair of points i and j, such as Euclidean distance;
$max\_jump(\ \Psi_{(i,j,n,\Omega)}\ )$	$max\_jump(\ \Psi_{(i,j,n,\Omega)}\ )$ is the maximum jump in path $\Psi_{(i,j,n,\Omega)}$. A jump is the distance from two consecutive points p and q in the path;
$\Theta_{(i,j,\Omega)}$	$\Theta_{(i,j,\Omega)}$ is the set of all paths from point i to point j. A path in $\Theta_{(i,j,\Omega)}$ can have arbitrary number of points (at least two). All the points in a path must belong to set Ω;
$MMJ(i,j \mid \Omega)$	$MMJ(i,j \mid \Omega)$ is the MMJ distance between point i and j, where Ω is the $\boldsymbol{Context}$ of the MMJ distance;
$\mathbb{M}_{k,\Omega_{[1,k]}}$	$\mathbb{M}_{k,\Omega_{[1,k]}}$ is the pairwise MMJ distance matrix of $\Omega_{[1,k]}$, which has shape $k \times k$. The MMJ distances are under the $\boldsymbol{Context}$ of $\Omega_{[1,k]}$;
$\mathbb{M}_{\Omega}$	The pairwise MMJ distance matrix of Ω, $\mathbb{M}_{\Omega} = \mathbb{M}_{N,\Omega_{[1,N]}}$;

$$\Pi = \{max\_jump(\epsilon) \mid \epsilon \in \Theta_{(i,j,\Omega)}\} \tag{1}$$

$$MMJ(i,j \mid \Omega) = min(\Pi) \tag{2}$$

Where ϵ is a path from point i to point j, $max\_jump(\epsilon)$ is the maximum jump in path ϵ. Π is the set of all maximum jumps. $min(\Pi)$ is the minimum of Set Π.

Set Ω is called the $\boldsymbol{Context}$ of the Min-Max-Jump distance. It is easy to check $MMJ(i,i \mid \Omega) = 0$.

In summary, Min-Max-Jump distance is the minimum of maximum jumps of all path between a pair of points, under the $\boldsymbol{Context}$ of a set of points.

Similar distances have actually been studied in many places in the literature, including the maximum capacity path problem, the widest path problem, the bottleneck edge query problem, the minimax path problem, the bottleneck shortest path problem, and the longest-leg path distance (LLPD) (20; 21; 22; 23).

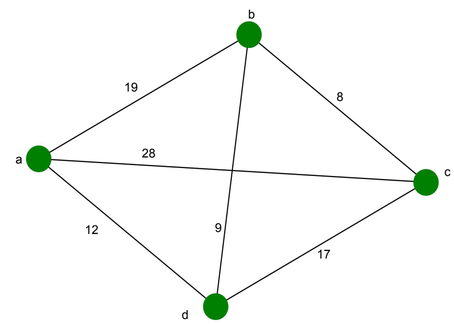


Figure 1: An example

There is a minor difference between Min-Max-Jump distance and other similar distances: Min-Max-Jump distance stresses the *context* of the distance. The *context* is like the condition in conditional probability. The difference becomes non-trivial when we need to calculate the pairwise MMJ distance matrix of a set S, under the context of its superset X, such as in Section 6.3 of (24). A set Ω is a superset of another set B if all elements of the set B are elements of the set Ω.

3.1 An example

Suppose Set Ω is composed of the four points in Figure 1. There are five (non-looped) paths from point a to point c in Figure 1:

1. $a \to c$, the maximum jump is 28;
2. $a \to b \to c$, the maximum jump is 19;
3. $a \to d \to c$, the maximum jump is 17;
4. $a \to b \to d \to c$, the maximum jump is 19;
5. $a \to d \to b \to c$, the maximum jump is 12.

According to Definition 1, $MMJ(a, c \mid \Omega) = 12$.

To understand Min-Max-Jump distance, imagine someone is traveling by jumping in Ω. Suppose $MMJ(i, j \mid \Omega) = \delta$. If the person wants to reach j from i, she must have the ability of jumping at least δ. Otherwise, j is unreachable from i for her. Whether the distance to a point is "far" or "near" is measured by how far (or how high) it requires a person to jump. If the requirement is large, then the point is "far", otherwise, it is "near."

3.2 Properties of MMJ distance

Theorem 1. *Suppose* $i, j, p, q \in \Omega$ *,*

$$MMJ(i, j \mid \Omega) = \delta \tag{3}$$

$$d(i, p) < \delta \tag{4}$$

$$d(j, q) < \delta \tag{5}$$

then,

$$MMJ(p, q \mid \Omega) = \delta \tag{6}$$

where d(x,y) is a distance function (Table 1).

Proof. $MMJ(i, j \mid \Omega) = \delta$ is equivalent to $\exists P \in \Theta_{(i,j,\Omega)}$, such that $M(P) = \delta$, and $\forall T \in \Theta_{(i,j,\Omega)}$, $M(T) \geq \delta$, where $\Theta_{(i,j,\Omega)}$ is the set of all paths from point i to point j under context Ω. $M(P)$ is the maximum jump in path P. We can assume $MMJ(p, q \mid \Omega) > \delta$ and $MMJ(p, q \mid \Omega) < \delta$, then we will arrive to a contradiction in both cases. □

Theorem 2. *Suppose* $r \in \{1, 2, \ldots, n\}$,

$$f(t) = max(d(\Omega_{n+1}, \Omega_t),\ MMJ(\Omega_t, \Omega_r \mid \Omega_{[1,n]})) \tag{7}$$

$$\mathbb{X} = \{f(t) \mid t \in \{1, 2, \ldots, n\}\} \tag{8}$$

then,

$$MMJ(\Omega_{n+1}, \Omega_r \mid \Omega_{[1,n+1]}) = min(\mathbb{X}) \tag{9}$$

For the meaning of $\Omega_t, \Omega_r, \Omega_{[1,n]}, and\ \Omega_{[1,n+1]}$, see Table 1.

Proof. There are n possibilities of the MMJ path from Ω_{n+1} to Ω_r, under the context of $\Omega_{[1,n+1]}$, set $\mathbb{X}$ enumerate them all. Each element of $\mathbb{X}$ is the maximum jump of each possibility. Therefore, according to the definition of MMJ distance, $MMJ(\Omega_{n+1}, \Omega_r \mid \Omega_{[1,n+1]}) = min(\mathbb{X})$. □

Corollary 1. *Suppose* $r \in \{1, 2, \ldots, N\}, p \notin \Omega$,

$$f(t) = max(d(p, \Omega_t),\ MMJ(\Omega_t, \Omega_r \mid \Omega)) \tag{10}$$

$$\mathbb{X} = \{f(t) \mid t \in \{1, 2, \ldots, N\}\} \tag{11}$$

then,

$$MMJ(p, \Omega_r \mid \Omega + p) = min(\mathbb{X}) \tag{12}$$

For the meaning of $\Omega + p$, see Table 1.

Proof. The proof follows the conclusion of Theorem 2. □

Theorem 3. *Suppose* $i, j \in \{1, 2, \ldots, n\}$,

$$x_1 = MMJ(\Omega_i, \Omega_j \mid \Omega_{[1,n]}) \tag{13}$$

$$t_1 = MMJ(\Omega_{n+1}, \Omega_i \mid \Omega_{[1,n+1]}) \tag{14}$$

$$t_2 = MMJ(\Omega_{n+1}, \Omega_j \mid \Omega_{[1,n+1]}) \tag{15}$$

$$x_2 = max(t_1,\ t_2) \tag{16}$$

then,

$$MMJ(\Omega_i, \Omega_j \mid \Omega_{[1,n+1]}) = min(x_1,\ x_2) \tag{17}$$

Proof. There are two possibilities of the MMJ path from Ω_i to Ω_j, under the context of $\Omega_{[1,n+1]}$: Ω_{n+1} is in the path or it is not in the path. x_2 is the min-max jump of the first possibility; x_1 is the min-max jump of the second possibility. Therefore, according to the definition of MMJ distance, $MMJ(\Omega_i, \Omega_j \mid \Omega_{[1,n+1]}) = min(x_1,\ x_2)$. □

4 Calculation of Min-Max-Jump distance

We propose two methods to calculate the pairwise Min-Max-Jump distance matrix of a dataset. There are other methods for calculating or estimating it, such as a modified SLINK algorithm (25), or with Cartesian trees (26; 27), or from a sequence of nearest neighbor graphs (23), or a modified version of the Floyd–Warshall algorithm.

4.1 MMJ distance by recursion

The first method calculates $\mathbb{M}_\Omega$ by recursion. $\mathbb{M}_\Omega$ is the pairwise MMJ distance matrix of Ω (Table 1). $\mathbb{M}_{k,\Omega_{[1,k]}}$ is the MMJ distance matrix of the first k points of Ω (Table 1). Note $\mathbb{M}_{2,\Omega_{[1,2]}}$ is simple to calculate. $\mathbb{M}_\Omega = \mathbb{M}_{N,\Omega_{[1,N]}}$. $\mathbb{M}_\Omega$ is a $N \times N$ symmetric matrix. Rows and columns of $\mathbb{M}_\Omega$ are indexed from 1 to N.

Step 7 of Algorithm 1 can be calculated with the conclusion of Theorem 2; Step 12 of Algorithm 1 can be calculated with the conclusion of Theorem 3.

Algorithm 1 has complexity of $\mathcal{O}(n^3)$, where n is the cardinality of Set Ω.

Algorithm 1 MMJ distance by recursion

Input: Ω
Output: $\mathbb{M}_{\Omega}$

function MMJ_BY_RECURSION(Ω)
    $N \leftarrow length(\Omega)$
    Initialize $\mathbb{M}_{\Omega}$ with zeros
    Calculate $\mathbb{M}_{2,\Omega_{[1,2]}}$, fill in $\mathbb{M}_{\Omega}[1,2]$ and $\mathbb{M}_{\Omega}[2,1]$
    for $n \leftarrow 3$ to N **do**
        for $r \leftarrow 1$ to $n-1$ **do**
            Calculate $MMJ(\Omega_n, \Omega_r \mid \Omega_{[1,n]})$, fill in $\mathbb{M}_{\Omega}[n,r]$ and $\mathbb{M}_{\Omega}[r,n]$
        end for
        for $i \leftarrow 1$ to $n-1$ **do**
            for $j \leftarrow 1$ to $n-1$ **do**
                if $i < j$ **then**
                    Calculate $MMJ(\Omega_i, \Omega_j \mid \Omega_{[1,n]})$, update $\mathbb{M}_{\Omega}[i,j]$ and $\mathbb{M}_{\Omega}[j,i]$
                end if
            end for
        end for
    end for
    return $\mathbb{M}_{\Omega}$
end function

4.2 MMJ distance by calculation and copy

According to the conclusion of Theorem 1, there are many duplicated values in $\mathbb{M}_{\Omega}$. So in the second method we can calculate the MMJ distance value in one position and copy it to other positions in $\mathbb{M}_{\Omega}$.

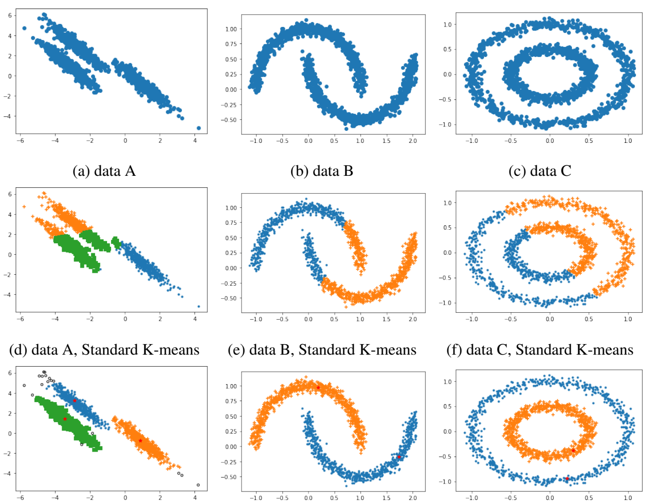
A well-known fact about MMJ distance is: "the path between any two nodes in a minimum spanning tree (MST) is a minimax path." A minimax path in an undirected graph is a path between two vertices v, w that minimizes the maximum weight of the edges on the path. That is to say, it is a MMJ path. By utilizing this fact, we propose Algorithm 2.

Algorithm 2 MMJ distance by Calculation and Copy

Input: Ω
Output: $\mathbb{M}_{\Omega}$

1: **function** MMJ_CALCULATION_AND_COPY(Ω)
2:     Initialize $\mathbb{M}_{\Omega}$ with zeros
3:     Construct a MST of Ω, noted T
4:     Sort edges of T from large to small, generate a list, noted L
5:     **for** e in L **do**
6:         Remove e from T. It will result in two connected sub-trees, T_1 and T_2;
7:         Traverse T_1 and T_2;
8:         For all pair of nodes (p,q), where $p \in T_1$, $q \in T_2$. Fill in $\mathbb{M}_{\Omega}[p,q]$ and $\mathbb{M}_{\Omega}[q,p]$ with the weight of e.
9:     **end for**
10:     **return** $\mathbb{M}_{\Omega}$
11: **end function**

The complexity of Algorithm 2 is $\mathcal{O}(n^2)$. Because the construction of a MST of a complete graph is $\mathcal{O}(n^2)$. During the "for" part (Step 5 to 9) of the algorithm, it accesses each cell of $\mathbb{M}_{\Omega}$ only once. Unlike Algorithm 1, which accesses each cell of $\mathbb{M}_{\Omega}$ for $\mathcal{O}(n)$ times. The merit of the "Calculation and Copy" method is that it is easier to understand than using the Cartesian trees (26; 27).



(a) data A (b) data B (c) data C

(d) data A, Standard K-means (e) data B, Standard K-means (f) data C, Standard K-means

(g) data A, MMJ-K-means (h) data B, MMJ-K-means (i) data C, MMJ-K-means

Figure 2: Standard K-means vs. MMJ-K-means

5 Applications of Min-Max-Jump distance

We explore two applications of MMJ distance, and test the applications with experiments. All the MMJ distances in the experiments are calculated with Algorithm 1.

5.1 MMJ-based K-means

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster center or centroid), serving as a prototype of the cluster (28). Standard K-means uses Euclidean distance. We can revise K-means to use Min-Max-Jump distance, with the cluster centroid replaced by the Semantic Center of Mass (SCOM) (particularly, One-SCOM) of each cluster. For the definition of SCOM, see a previous paper (29). One-SCOM is like medoid, but has some difference from medoid. Section 6.3 of (29) compares One-SCOM and medoid. In simple terms, the One-SCOM of a set of points, is the point which has the smallest sum of squared distances to all points in the set.

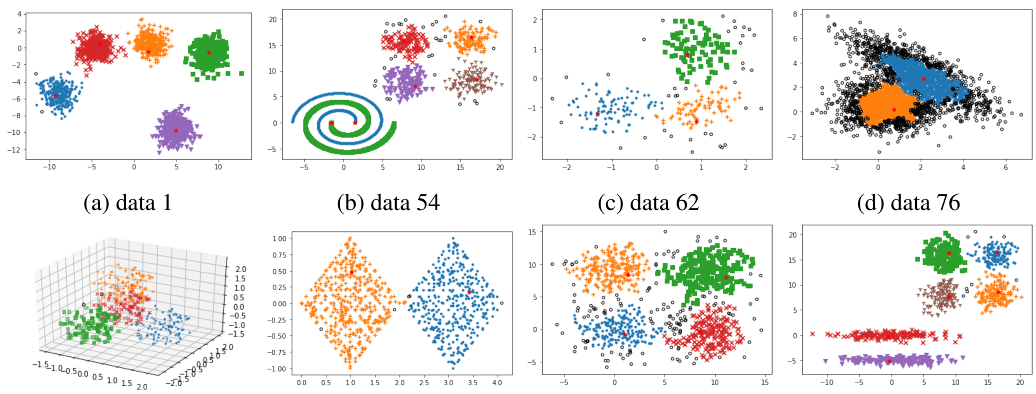
Standard K-means usually cannot deal with non-spherical shaped data, such as the ones in Figure 2. MMJ-based K-means (MMJ-K-means) can cluster such irregularly shaped data. Figure 2 compares Standard K-means and MMJ-K-means, on clustering three data which come from the scikit-learn project (30). Figure 3 are eight more samples of MMJ-K-means. The data sources corresponding to the data IDs can be found at this URL (temporarily hidden for double blind review).

It can be seen MMJ-K-means can (almost) work properly for clustering the 11 data, which have different kinds of shapes. The black circles are Border points (Definition 2), the red stars are the center (One-SCOM) of each cluster. During training of MMJ-K-means, the Border points are randomly allocated to one of its nearest centers.

Definition 2. *Border point*

A point is defined to be a Border point if its nearest mean (center, centroid, or One-SCOM) is not unique.

Compared with other clustering models that can handle irregularly shaped data, such as Spectral clustering or the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), the merit of MMJ-K-means is its simplicity; the logic of MMJ-K-means is as simple as K-means. We just replace the Euclidean distance with MMJ distance, and the centroid with the Semantic Center of Mass (SCOM).



(a) data 1 (b) data 54 (c) data 62 (d) data 76

(e) data 78 (f) data 83 (g) data 89 (h) data 102

Figure 3: Eight more samples of MMJ-K-means

	CH	SC	DB	CDbw	DBCV	VIASCKDE	New	MMJ-SC	MMJ-CH	MMJ-DB
Accuracy	27/145	38/145	42/145	8/145	56/145	11/145	74/145	83/145	90/145	69/145

Table 2: Accuracy of the ten indices

5.2 MMJ-based internal clustering evaluation index

Calinski-Harabasz index, Silhouette coefficient, and Davies-Bouldin index are three of the most popular techniques for internal clustering evaluation. They are used to calculate the goodness of a clustering technique.

The Silhouette coefficient for a single sample is given as:

$$s = \frac{b - a}{max(a, b)}$$

where a is the mean distance between a sample and all other points in the same class. b is the mean distance between a sample and all other points in the next nearest cluster. The Silhouette coefficient for a set of samples is given as the mean of Silhouette coefficient for each sample.

We can also revise Silhouette coefficient to use Min-Max-Jump distance, forming a new internal clustering evaluation index called MMJ-based Silhouette coefficient (MMJ-SC). We tested the performance of MMJ-SC with the 145 datasets mentioned in another paper(31). MMJ-SC obtained a good performance score compared with the other seven internal clustering evaluation indices mentioned in the paper(31). Readers can check Table 2 and compare with Table 5 of Liu's paper(31).

MMJ-based Calinski-Harabasz index (MMJ-CH) and MMJ-based Davies-Bouldin index (MMJ-DB) were also tested. In calculation of these two indices, besides using MMJ distance, the center/centroid of a cluster is replaced by the One-SCOM of the cluster again, as in MMJ-K-means. It can be seen that MMJ distance systematically improves the three internal clustering evaluation indices (Table 2). The best performer is MMJ-CH, which achieves an accuracy of $90/145$. The accuracy of an index is computed by evaluating the index's ability of recognizing the best partition of a dataset from hundreds of candidate partitions(31).

5.2.1 Using MMJ-SC in CNNI

The Clustering with Neural Network and Index (CNNI) model uses a Neural Network to cluster data points. Training of the Neural Network mimics supervised learning, with an internal clustering evaluation index acting as the loss function (24). CNNI with standard Silhouette coefficient as the internal clustering evaluation index, cannot deal with non-flat geometry data, such as data B and data C in Figure 2. MMJ-SC gives CNNI model the capability of processing non-flat geometry data. E.g., Figure 4 is the clustering result and decision boundary of data B by CNNI using MMJ-SC. It uses Neural Network C of the CNNI paper (24). CNNI equipped with MMJ-SC, achieves the

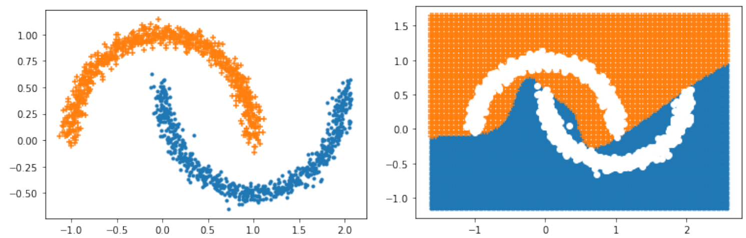


Figure 4: Clustering result and decision boundary of data B by CNNI using MMJ-SC

first inductive clustering model that can deal with non-flat geometry data (24). For the definition of non-flat geometry data, see this[1] Stackexchange question.

6 Discussion

6.1 Using PAM

Since One-SCOM is like medoid, in MMJ-K-means, we can also use the Partitioning Around Medoids (PAM) algorithm or its variants to find the One-SCOMs (32).

6.2 Multiple One-SCOMs in one cluster

There might be multiple One-SCOM points in a cluster, which have the same smallest sum of squared distances to all the points in the cluster. Usually they are not far from each other. We can arbitrarily choose one or keep them all. If we keep them all, then the One-SCOM of a cluster is not a point, but a set of points. If the One-SCOM is a set, when calculating a point's MMJ distance to the One-SCOM of a cluster, we can select the minimum of the point's MMJ distances to all the One-SCOM points.

6.3 Differentiating border points

Border points defined in Definition 2 can further be differentiated as weak and strong border points.

Definition 3. *Weak Border Point (WBP)*

A point is defined to be a WBP if its nearest mean (center or One-SCOM) is not unique but less than K, where K is the number of clusters.

Definition 4. *Strong Border Point (SBP)*

A point is defined to be a SBP if its nearest mean (center or One-SCOM) is not unique and equals K, where K is the number of clusters.

Then we can process different kinds of border points with different strategies. E.g., deeming the Strong Border Points as outliers and removing them.

7 Conclusion and Future Works

We proposed two algorithms for calculating Min-Max-Jump distance (MMJ distance), and tested two applications of it: MMJ-based K-means and MMJ-based internal clustering evaluation index. MMJ-K-means overcomes a big drawback of K-means, improving its ability of clustering, so that it can handle irregularly shaped clusters. We claim MMJ-CH is the SOTA (state-of-the-art) internal clustering evaluation index, which achieves an accuracy of $90/145$. To thoroughly test the internal clustering evaluation indices, we conducted an experiment on a set of 145 datasets. A normal Machine Learning paper usually uses several or dozens of datasets to test their models or algorithms. In summary, MMJ distance has good capability and potentiality in Machine Learning. Further research may test its applications in other models, such as other clustering evaluation indices.

[1] `https://datascience.stackexchange.com/questions/52260/terminology-flat-geometry-in-the-context-of-clustering`

References

[1] S. Z. Li and A. Jain, Eds., *Hamming Distance*. Boston, MA: Springer US, 2009, pp. 668–668. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_956

[2] D. Sinwar and R. Kaushik, "Study of euclidean and manhattan distance metrics using simple k-means clustering," *Int. J. Res. Appl. Sci. Eng. Technol*, vol. 2, no. 5, pp. 270–274, 2014.

[3] R. Coghetto, "Chebyshev distance," 2016.

[4] P. J. Groenen and K. Jajuga, "Fuzzy clustering with squared minkowski distances," *Fuzzy Sets and Systems*, vol. 120, no. 2, pp. 227–237, 2001.

[5] S. Fletcher, M. Z. Islam *et al.*, "Comparing sets of patterns with the jaccard index," *Australasian Journal of Information Systems*, vol. 22, 2018.

[6] N. R. Chopde and M. Nichat, "Landmark based shortest path detection by using a* and haversine formula," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, no. 2, pp. 298–302, 2013.

[7] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[8] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of lloyd-type methods for the k-means problem," *Journal of the ACM (JACM)*, vol. 59, no. 6, pp. 1–22, 2013.

[9] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[10] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[11] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," *Knowledge and Information Systems*, vol. 19, no. 3, pp. 361–394, 2009.

[12] S. Petrović, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters," 2006.

[13] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *International conference on computational intelligence and multimedia applications (ICCIMA 2007)*, vol. 2. IEEE, 2007, pp. 13–17.

[14] J. C. Bezdek and N. R. Pal, "Cluster validation with generalized dunn's indices," in *Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems*. IEEE Computer Society, 1995, pp. 190–190.

[15] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[16] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[17] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the national academy of sciences*, vol. 102, no. 21, pp. 7426–7431, 2005.

[18] B. Fischer and J. M. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 4, pp. 513–518, 2003.

[19] H. Chang and D.-Y. Yeung, "Robust path-based spectral clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 191–203, 2008.

[20] M. Pollack, "The maximum capacity through a network," *Operations Research*, vol. 8, no. 5, pp. 733–736, 1960.

[21] T. Hu, "The maximum capacity route problem," *Operations Research*, vol. 9, no. 6, pp. 898–900, 1961.

[22] P. M. Camerini, "The min-max spanning tree problem and some extensions," *Information Processing Letters*, vol. 7, no. 1, pp. 10–14, 1978.

[23] A. V. Little, M. Maggioni, and J. M. Murphy, "Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms," *J. Mach. Learn. Res.*, vol. 21, pp. 6:1–6:66, 2020. [Online]. Available: http://jmlr.org/papers/v21/18-085.html

[24] G. Liu, "Clustering with neural network and index," *arXiv preprint arXiv:2212.03853*, 2022.

[25] R. Sibson, "Slink: an optimally efficient algorithm for the single-link cluster method," *The computer journal*, vol. 16, no. 1, pp. 30–34, 1973.

[26] N. Alon and B. Schieber, *Optimal preprocessing for answering on-line product queries*. Citeseer, 1987.

[27] E. D. Demaine, G. M. Landau, and O. Weimann, "On cartesian trees and range minimum queries," *Algorithmica*, vol. 68, pp. 610–625, 2014.

[28] H.-H. Bock, "Clustering methods: a history of k-means algorithms," *Selected contributions in data analysis and classification*, pp. 161–172, 2007.

[29] G. Liu, "Topic model supervised by understanding map," *arXiv preprint arXiv:2110.06043*, 2021.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] G. Liu, "A new index for clustering evaluation based on density estimation," *arXiv preprint arXiv:2207.01294*, 2022.

[32] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O (k) runtime improvement of the pam, clara, and clarans algorithms," *Information Systems*, vol. 101, p. 101804, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: We have used 145 datasets to test the models in the paper. Maybe it is not enough, we need more datasets to test the models.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs of theoretical results have been provided in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have fully disclosed all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an URL to data and code of the paper, to reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Full details are provided with the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not contain statistical experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not discuss about the efficiency of the models, only effectiveness and time complexity of the algorithms. Because efficiency can be affected by a lot of factors, e.g., using C++ to implement is much faster than using python, and some minor optimization of the codes may drastically improve the speed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets used in the paper are properly credited. The license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.