Clustering ideas

TODO[a,*]

[a]Technical University of Cluj-Napoca, Str. Memorandumului, Nr. 28, Cluj-Napoca, 400114, Romania

Abstract

TODO

Keywords:

1. Main Result

This section formalizes the evaluation framework and sets up the geometric argument behind why Euclidean-distance-based CVIs fail and how Arboris Distance resolves this

1.1. Evaluation Pipeline for Geometry-Aware Clustering

We formalize a minimal but complete framework for evaluating clustering algorithms that (i) compares predicted clusterings against ground truth, (ii) makes explicit the geometric assumptions induced by the data representation, and (iii) enables the analysis of out-of-sample behavior, smoothness, and sensitivity. This framework serves as the basis for our main theoretical result on the limitations of distance-based cluster validity indices.

Step 1: Ground-truth dataset.. Let

$$X = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^d, \qquad y_{\text{true}} \in \{1, \dots, K\}^N$$

denote a finite dataset together with its ground-truth labeling. The true clusters are given by

$$C_k^{\text{true}} = \{i \mid y_{\text{true}}(i) = k\}, \qquad k = 1, \dots, K.$$

Step 2: Clustering on the finite sample.. A clustering algorithm, possibly parameterized by θ, is modeled as a mapping

$$\mathcal{A}_\theta : X \longmapsto y \in \{1, \dots, \hat{K}\}^N,$$

which assigns a label to each element of the observed dataset. This induces predicted clusters

$$C_m = \{i \mid y(i) = m\}, \qquad m = 1, \dots, \hat{K}.$$

Importantly, $\mathcal{A}_\theta$ acts only on the finite set X and does not, in general, define labels outside the observed samples.

Step 3: Geometry induced by the data.. To evaluate clustering quality, we associate the dataset X with a geometry. This geometry is encoded either by

- a distance function $d_X : X \times X \to \mathbb{R}_{\geq 0}$, or equivalently
- a set of nonnegative pairwise weights $w_{ij}(X)$ satisfying $\sum_{i<j} w_{ij} = 1$, where w_{ij} reflects the notion of proximity or similarity between samples x_i and x_j.

Throughout this work, the choice of geometry is understood to be fully captured by the weights $w_{ij}(X)$. This choice implicitly defines an embedding in which clustering quality is assessed.

⋆This work was been financially supported by project DECIDE, no. 57/14.11.2022, contract number 760069, funded under the PNRR I8 scheme by the Romanian Ministry of Research, Innovation, and Digitisation.

*Corresponding author.

Email address: Mircea.Susca@aut.utcluj.ro (TODO)

Step 4: In-sample clustering evaluation.. Given the ground-truth labeling y_{true} and a predicted labeling y, we evaluate clustering quality using a geometry-aware cluster validity index

$$f(X, y_{\text{true}}, y) \in [0, 1],$$

satisfying the normalization properties

$$f(X, y_{\text{true}}, y_{\text{true}}) = 1, \qquad \mathbb{E}_{y_{\text{rand}}}[f(X, y_{\text{true}}, y_{\text{rand}})] = 0,$$

where the expectation is taken with respect to a specified random labeling model. Such indices compare the agreement between y and y_{true} while weighting pairwise contributions according to the geometry induced by X.

Step 5: Out-of-sample assignment.. Since most clustering algorithms do not naturally assign labels to unseen data, we explicitly introduce an out-of-sample extension rule

$$\phi : (X, y, x_{\text{new}}) \longmapsto \hat{y}_{\text{new}}, \qquad x_{\text{new}} \in \mathbb{R}^d,$$

which assigns a cluster label to a new point. Typical choices include nearest-neighbor, centroid-based, or graph-based rules. This separation between $\mathcal{A}_\theta$ and ϕ is essential for analyzing out-of-sample behavior and geometric consistency.

Step 6: Local correctness and smoothness (optional).. If a continuous or probabilistic ground-truth description is available, we define a truth membership field $\pi(x) \in \Delta^{K-1}$ and a predicted membership $\rho(x) \in \Delta^{\hat{K}-1}$. A local correctness score at x_{new} may then be defined as

$$f_{\text{local}}(x_{\text{new}}) = 1 - \frac{D(\pi(x_{\text{new}}), \rho(x_{\text{new}}))}{D(\pi(x_{\text{new}}), u)},$$

where $D(\cdot, \cdot)$ is a bounded divergence (e.g., Jensen–Shannon) and u denotes the uniform distribution. Studying the variation of $f_{\text{local}}(x)$ as x varies enables the analysis of smoothness, sensitivity near decision boundaries, and robustness with respect to the chosen geometry.

Taken together, this pipeline makes explicit the role of geometry in clustering evaluation and provides the formal setting in which we analyze the failure modes of distance-based cluster validity indices.

Proposition 1 (Geometry-induced failure of distance-based CVIs). *Let $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$ be a dataset with ground-truth labels $y_{\text{true}} \in \{1, \dots, K\}^N$. Consider a cluster validity index of the form*

$$f(X, y_{\text{true}}, y) = \frac{A_w(X, y_{\text{true}}, y) - \mu_w}{1 - \mu_w}, \tag{1}$$

where

$$A_w(X, y_{\text{true}}, y) = \sum_{i<j} w_{ij}(X)\, \mathbf{1}\{S(i, j) = S_{\text{true}}(i, j)\}, \tag{2}$$

with pairwise relations $S(i, j) = \mathbf{1}\{y(i) = y(j)\}$, $S_{\text{true}}(i, j) = \mathbf{1}\{y_{\text{true}}(i) = y_{\text{true}}(j)\}$, and nonnegative weights $w_{ij}(X)$ satisfying $\sum_{i<j} w_{ij} = 1$. Assume that the weights are monotonically decreasing functions of the Euclidean distance, i.e.,

$$w_{ij}(X) = \psi(\|x_i - x_j\|), \quad \psi'(\cdot) \le 0.$$

Then there exist datasets X and ground-truth labelings y_{true} (for example, non-globular cluster structures such as concentric manifolds) for which f assigns a higher score to an incorrect labeling $y \neq y_{\text{true}}$ than to the true labeling y_{true}.

Proof 1 (Proof sketch). *The CVI in* (1) *evaluates agreement between the true and predicted pairwise relations, weighted by proximity in the embedding induced by the chosen distance. Because w_{ij} is a decreasing function of Euclidean distance, pairs of points that are close in Euclidean space contribute more heavily to A_w.*

For datasets whose ground-truth clusters are globular and well separated, Euclidean proximity correlates strongly with true co-membership: most pairs with small $\|x_i - x_j\|$ satisfy $S_{\text{true}}(i, j) = 1$. In this case, $A_w(X, y_{\text{true}}, y_{\text{true}})$ is maximal, and the CVI correctly favors the true labeling.

However, consider datasets in which the ground-truth clusters lie on non-globular manifolds (e.g., concentric rings). In such cases, there exist many pairs (i, j) with small Euclidean distance but different ground-truth labels, and many pairs with large Euclidean distance but identical ground-truth labels. Consequently, the ground-truth labeling y_{true} exhibits high pairwise disagreement on heavily weighted (small-distance) pairs, leading to a reduced value of $A_w(X, y_{\text{true}}, y_{\text{true}})$.

At the same time, there exist alternative labelings $y \neq y_{\text{true}}$ that better align with Euclidean proximity (e.g., partitions producing compact, Euclidean-separated clusters), yielding higher weighted agreement A_w despite being semantically incorrect. Since the normalization in (1) *preserves ordering, $f(X, y_{\text{true}}, y)$ can exceed $f(X, y_{\text{true}}, y_{\text{true}})$. This establishes the claim.*

Intuitively: Any CVI whose weights are monotone in Euclidean distance inherits Euclidean geometry.

Corollary 1 (Restoration via connectivity-aware distance). *If the Euclidean distance in Proposition 1 is replaced by a connectivity-aware distance (e.g., the Arboris Distance), for which within-manifold distances are small and between-manifold distances are separated by large bottlenecks, then the ground-truth labeling becomes smooth with respect to the induced weights w_{ij}, and the mis-ranking phenomenon described above does not occur.*

Remark 1 (Classical CVIs as instances of Eq. (1)**).** *Many widely used internal cluster validity indices can be interpreted as special cases of the general form* (1)*, differing only in the choice of pairwise weighting $w_{ij}(X)$ and in how agreement between pairs is aggregated.*

- ***Dunn index.*** *The Dunn index compares the minimum inter-cluster distance with the maximum intra-cluster diameter. This corresponds to an extremal weighting in which only the smallest inter-cluster distance and the largest intra-cluster distance are retained, implicitly assigning all weight to those pairs. Since both quantities are defined using Euclidean distances, Dunn inherits the Euclidean geometry encoded by d_X.*
- ***Silhouette coefficient.*** *The Silhouette coefficient evaluates, for each point, the difference between its average distance to points in the same cluster and to points in the nearest neighboring cluster. This can be interpreted as a soft pairwise comparison in which $w_{ij}(X)$ decreases monotonically with $\|x_i - x_j\|$, and agreement is measured via relative proximity. Consequently, Silhouette also inherits Euclidean geometry.*
- ***Other internal indices.*** *Indices such as Davies–Bouldin and Calinski–Harabasz similarly rely on averages or extrema of Euclidean distances and therefore correspond to particular choices of distance-based weights within the same general framework.*

Thus, although these indices differ in aggregation strategy, they all implicitly encode the assumption that Euclidean proximity is aligned with cluster membership. Proposition 1 therefore applies uniformly to this class of indices.

Example 1 (Concentric clusters). *Consider a dataset consisting of two concentric manifolds in $\mathbb{R}^2$, with ground-truth labels corresponding to the inner and outer structures. Points belonging to different ground-truth clusters may be arbitrarily close in Euclidean distance, while points in the same cluster may be far apart.*

In this setting, Euclidean-distance-based CVIs such as Dunn or Silhouette may assign higher scores to geometrically compact but semantically incorrect partitions than to the true clustering. This behavior is observed empirically in Section X and follows directly from Theorem **??**.

1.2. Proposed Solution: Geometry Replacement via Arboris Distance

The analysis in Proposition 1 shows that the failure of many classical cluster validity indices does not stem from their functional definition, but from the geometric assumptions encoded by the distance used to define pairwise relationships. In particular, when the weights $w_{ij}(X)$ are monotone functions of the Euclidean distance, the resulting CVI implicitly evaluates clustering quality in a Euclidean embedding, which is incompatible with non-globular or manifold-based cluster structures.

We therefore propose a minimal and principled modification of the evaluation pipeline: *replace the Euclidean distance by a connectivity-aware distance*, while keeping the CVI definition unchanged.

Graph construction.. Let $G = (V, E)$ be a proximity graph constructed on the dataset $X = \{x_i\}_{i=1}^N$, where each vertex corresponds to a sample and edges are weighted by Euclidean distance,

$$w_{ij}^{\text{Euc}} = \|x_i - x_j\|.$$

Typical choices include the k-nearest neighbor graph or the complete graph restricted to a minimum spanning tree (MST).

Arboris Distance.. Let $T \subseteq G$ denote a spanning tree (or forest) of G. For any two samples $x_i, x_j \in X$, the Arboris Distance is defined as

$$d_{\text{AD}}(x_i, x_j) := \max_{(u,v)\in\mathcal{P}_{ij}} w_{uv}^{\text{Euc}}, \tag{3}$$

where $\mathcal{P}_{ij}$ denotes the unique path between x_i and x_j in T. Thus, $d_{\text{AD}}(x_i, x_j)$ measures the minimum bottleneck required to connect x_i and x_j through the data graph.

This distance captures data connectivity rather than ambient proximity: points lying on the same intrinsic structure are connected by paths with small bottlenecks, whereas points belonging to different structures are separated by large values of d_{AD}.

Geometry-aware weighting.. The Arboris Distance induces a new geometry on X by defining pairwise weights as

$$w_{ij}(X) = \frac{\psi\big(d_{\text{AD}}(x_i, x_j)\big)}{\sum_{p<q} \psi\big(d_{\text{AD}}(x_p, x_q)\big)}, \qquad \psi'(\cdot) \le 0, \tag{4}$$

where ψ is a monotone decreasing function (e.g., Gaussian or inverse polynomial). These weights replace the Euclidean-based weights in Eq. (1).

Effect on the CVI.. Substituting (4) into the definition of the weighted agreement

$$A_w(X, y_{\text{true}}, y) = \sum_{i<j} w_{ij}(X)\, \mathbf{1}\{S(i, j) = S_{\text{true}}(i, j)\},$$

yields a CVI that evaluates clustering quality with respect to the connectivity-induced geometry of the data. For clusterings that are non-smooth under Euclidean distance but consistent with the intrinsic data structure, the ground-truth labeling becomes smooth with respect to d_{AD}, restoring maximal agreement $A_w(X, y_{\text{true}}, y_{\text{true}})$.

Importantly, this modification does not alter the normalization, interpretation, or range of the CVI. It replaces only the geometric component of the evaluation, thereby extending the applicability of existing CVIs beyond Euclidean, globular cluster structures without introducing new hyperparameters or heuristic corrections.

Remark 2 (Euclidean distance as a degenerate case of Arboris Distance). *The Euclidean distance can be viewed as a special, degenerate instance of the Arboris Distance. Indeed, if the proximity graph G is chosen as the complete graph on X and the spanning tree T contains the direct edge (i, j) for every pair of samples, then the unique path $\mathcal{P}_{ij}$ between x_i and x_j consists of a single edge with weight $\|x_i - x_j\|$. In this case, the Arboris Distance defined in* (3) *reduces to*

$$d_{\text{AD}}(x_i, x_j) = \|x_i - x_j\|.$$

More generally, Euclidean distance corresponds to evaluating proximity using only direct edges in the ambient space, thereby ignoring the connectivity structure of the data. The Arboris Distance extends this notion by allowing proximity to be mediated by paths through the data graph, recovering Euclidean distance when such paths are restricted to single edges.