CVIK: A MATLAB-based Cluster Validity Index Toolbox for Automatic Data Clustering Applications

SUPPLEMENTARY MATERIAL

Adán José-García*,[1] and Wilfrido Gómez-Flores[2]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[2]Cinvestav-IPN, Unidad Tamaulipas, 87130, Cd. Victoria, Tamaulipas, México

1 CLUSTER VALIDITY INDICES

Description of several cluster validity indices (CVIs) to address the automatic data clustering problem.

1.1 *Definitions*

Some basic definitions are given before presenting the CVIs:

- A *pattern* (or *object*) $\mathbf{x}$ is a single data item represented by a vector of measurements $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$, where $x_i \in \mathbb{R}$ is a *feature* (or *attribute*) and D denotes the *dimensionality*.
- A *dataset* is denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^D$, where N is the total number of patterns in the dataset.
- A *clustering*, denoted by $\mathbf{C} = \{\mathbf{c}_k \mid k = 1, \ldots, K\}$, refers to a set of mutually disjoint clusters that partition $\mathbf{X}$ into K groups.
- The *number of objects* in a cluster $\mathbf{c}_k$ is denoted by $n_k = |\mathbf{c}_k|$.
- The *centroid of a cluster* (or *prototype*) is expressed as $\bar{\mathbf{c}}_k = {}^1/_{n_k} \sum_{\mathbf{x}_i \in \mathbf{c}_k} \mathbf{x}_i$, whereas the *centroid of a dataset* $\mathbf{X}$ is denoted by $\bar{\mathbf{X}} = {}^1/_N \sum_{\mathbf{x}_i \in \mathbf{X}} \mathbf{x}_i$.
- A *distance measure* is a metric (or quasi-metric) used to quantify the proximity between two patterns. Probably the most common distance metric is the Euclidean distance, expressed as $d_e(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$.
- The *point symmetry distance* [2] between the object $\mathbf{x}_i$ and the cluster $\mathbf{c}_k$ is defined by

$$d_{ps}(\mathbf{x}_i, \mathbf{c}_k) = \frac{1}{k_n} \sum \min_{\mathbf{x}_j \in \mathbf{c}_k} (k_n) \left\{ d_e \left(2\bar{\mathbf{c}}_k - \mathbf{x}_i, \mathbf{x}_j \right) \right\}. \tag{1}$$

  The point $2\bar{\mathbf{c}}_k - \mathbf{x}_i$ is called the symmetrical (reflected) point of $\mathbf{x}_i$ with respect to the centroid of $\mathbf{c}_k$ and $k_n = 2$ are its unique nearest neighbors. The function $\sum \min$ can be seen as an adaptation of the min function, where $\sum \min(n)$ computes the sum of the n lowest arguments.

*Corresponding author: adan.josegarcia@univ-lille.fr

Table 1: List of the CVIs included in the CVIK toolbox.

Year	CVI name	Index ID	X	D	Range	Ref.
1973	Dunn index	$\text{Dunn}_{\uparrow}$		✓	$[0,+\infty]$	[8]
1974	Calinski-Harabasz index	$\text{CH}_{\uparrow}$	✓		$[0,+\infty]$	[4]
1976	C index	$\text{CI}_{\downarrow}$		✓	$[0,+\infty]$	[12]
1979	Davies-Bouldin index	$\text{DB}_{\downarrow}$	✓		$[0,+\infty]$	[7]
1987	Silhouette index	$\text{Sil}_{\downarrow}$		✓	$[-1,+1]$	[17]
1991	Xie-Beni index	$\text{XB}_{\downarrow}$	✓		$[0,+\infty]$	[20]
1998	Generalized Dunn-31 index	$\text{gD31}_{\uparrow}$		✓	$[0,+\infty]$	[3]
1998	Generalized Dunn-41 index	$\text{gD41}_{\uparrow}$	✓	✓	$[0,+\infty]$	[3]
1998	Generalized Dunn-51 index	$\text{gD51}_{\uparrow}$	✓	✓	$[0,+\infty]$	[3]
1998	Generalized Dunn-33 index	$\text{gD33}_{\uparrow}$	✓	✓	$[0,+\infty]$	[3]
1998	Generalized Dunn-43 index	$\text{gD43}_{\uparrow}$	✓		$[0,+\infty]$	[3]
1998	Generalized Dunn-53 index	$\text{gD53}_{\uparrow}$	✓		$[0,+\infty]$	[3]
2001	SDbw index	$\text{SDbw}_{\downarrow}$	✓		$[0,+\infty]$	[10]
2002	PBM index	$\text{PBM}_{\uparrow}$	✓		$[0,+\infty]$	[15]
2004	CS index	$\text{CS}_{\downarrow}$	✓	✓	$[0,+\infty]$	[6]
2005	Enhanced Davies-Bouldin index	$\text{DB2}_{\downarrow}$	✓		$[0,+\infty]$	[13]
2007	Score Function index	$\text{SF}_{\uparrow}$	✓		$[0,+\infty]$	[19]
2008	Symmetry index	$\text{Sym}_{\uparrow}$	✓		$[0,+\infty]$	[2]
2009	Davies-Bouldin based on Symmetry index	$\text{SDB}_{\downarrow}$	✓		$[0,+\infty]$	[18]
2009	Dunn based on Symmetry index	$\text{SDI}_{\uparrow}$	✓	✓	$[0,+\infty]$	[18]
2010	COP index	$\text{COP}_{\downarrow}$	✓	✓	$[0,+\infty]$	[9]
2011	SV index	$\text{SV}_{\uparrow}$	✓		$[0,+\infty]$	[22]
2013	Index based on nearest neighbors	$\text{CVNN}_{\downarrow}$		✓	$[0,+2]$	[21]
2014	WB index	$\text{WB}_{\downarrow}$	✓		$[0,+\infty]$	[23]
2014	Density-based index	$\text{DBCV}_{\uparrow}$	✓		$[-1,+1]$	[16]
2019	Index based on density-involved distance	$\text{CVDD}_{\uparrow}$		✓	$[0,+\infty]$	[11]
2019	Index based on local cores	$\text{LCCV}_{\uparrow}$	✓	✓	$[-1,+1]$	[5]
2020	Index based on shapes, sizes densities, and separation distances	$\text{SSDD}_{\downarrow}$	✓	✓	$[0,+1]$	[14]

1.2 *Cluster validity indices*

A CVI defines a relation between intracluster cohesion (within-group scatter) and intercluster separation (between-group scatter) to estimate the quality of a clustering solution [13]. The 28 CVIs included in this toolbox are described in the next. An acronym is defined to identify each CVI, followed by an up arrow (↑) or a down arrow (↓) to indicate whether the index is to be maximized or minimized. A general description is provided for every CVI in terms of criteria of cluster cohesion and separation.

- The Dunn index ($\mathbf{DI}_{\uparrow}$) [8]. This is a ratio-type index in which the cohesion is quantified by the maximum cluster diameter and the separation by the nearest neighbor distance. It is defined by

$$\mathrm{DI}(\mathbf{C}) = \frac{\min_{\mathbf{c}_k \in \mathbf{C}} \left\{ \min_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \{\delta(\mathbf{c}_k, \mathbf{c}_r)\} \right\}}{\max_{\mathbf{c}_k \in \mathbf{C}} \{\Delta(\mathbf{c}_k)\}}, \tag{2}$$

  where

$$\begin{aligned} \delta(\mathbf{c}_k, \mathbf{c}_r) &= \min_{\mathbf{x}_i \in \mathbf{c}_k, \mathbf{x}_j \in \mathbf{c}_r} \{d_e(\mathbf{x}_i, \mathbf{x}_j)\}, \\ \Delta(\mathbf{c}_k) &= \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{c}_k} \{d_e(\mathbf{x}_i, \mathbf{x}_j)\}. \end{aligned}$$

- The Calinski–Harabasz index ($\mathbf{CH}_{\uparrow}$) [4]. This is a ratio-type index in which the cohesion is quantified by the sum of the distances of the patterns to their respective centroids, and the separation is measured by the sum of the distances from the centroids to the global prototype. It is defined by

$$\mathrm{CH}(\mathbf{C}) = \frac{N - K}{K - 1} \times \frac{\sum_{\mathbf{c}_k \in \mathbf{C}} n_k \, d_e^2\left(\bar{\mathbf{c}}_k, \bar{\mathbf{X}}\right)}{\sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e^2(\mathbf{x}_i, \bar{\mathbf{c}}_k)}. \tag{3}$$

- The C-Index ($\mathbf{CI}_{\downarrow}$) [12]. This is a normalized index in which the cohesion is quantified by the sum of the distances between all the patterns in the same cluster, and the separation is based on the sum of the smallest and largest distances between all the patterns in the dataset. It is computed by

$$\mathrm{CI}(\mathbf{C}) = \frac{S(\mathbf{C}) - S_{\min}(\mathbf{C})}{S_{\max}(\mathbf{C}) - S_{\min}(\mathbf{C})}, \tag{4}$$

  where

$$\begin{aligned} S(\mathbf{C}) &= \sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{c}_k} d_e(\mathbf{x}_i, \mathbf{x}_j), \\ S_{\min}(\mathbf{C}) &= \sum \min(n_w)_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}} \{d_e(\mathbf{x}_i, \mathbf{x}_j)\}, \\ S_{\max}(\mathbf{C}) &= \sum \max(n_w)_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}} \{d_e(\mathbf{x}_i, \mathbf{x}_j)\}. \end{aligned}$$

  and where n_w is the number of pairs of objects in a partition that are in the same cluster: $n_w = \sum_{\mathbf{c}_k \in \mathbf{C}} \binom{n_k}{2}$.

- The Davies–Bouldin index ($\mathbf{DB}_{\downarrow}$) [7]. In this index, the cohesion is quantified by the mean distance of the objects to their respective centroids, and the separation quantifies the distance between centroids. It is expressed by

$$\mathrm{DB}(\mathbf{C}) = \frac{1}{K} \sum_{\mathbf{c}_k \in \mathbf{C}} \max_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \left\{ \frac{S(\mathbf{c}_k) + S(\mathbf{c}_r)}{d_e(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r)} \right\}, \tag{5}$$

  where

$$S(\mathbf{c}_k) = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e(\mathbf{x}_i, \bar{\mathbf{c}}_k).$$

- The Silhouette index ($\mathbf{Sil}_{\uparrow}$) [17]. This index is a normalized summation-type index in which the cohesion is measured by the sum of the distances between all the points in the same cluster, and the separation is based on the nearest neighbor distance between points in different groups. It is defined by

$$\mathrm{SI}(\mathbf{C}) = \frac{1}{N} \sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{x}_i \in \mathbf{c}_k} \frac{b\left(\mathbf{x}_i, \mathbf{c}_k\right) - a\left(\mathbf{x}_i, \mathbf{c}_k\right)}{\max\left\{b\left(\mathbf{x}_i, \mathbf{c}_k\right), a\left(\mathbf{x}_i, \mathbf{c}_k\right)\right\}}, \tag{6}$$

where

$$a\left(\mathbf{x}_i, \mathbf{c}_k\right) = \frac{1}{n_k - 1} \sum_{\mathbf{x}_j \in \mathbf{c}_k} d_e\left(\mathbf{x}_i, \mathbf{x}_j\right),$$
$$b\left(\mathbf{x}_i, \mathbf{c}_k\right) = \min_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \left\{ \frac{1}{n_r} \sum_{\mathbf{x}_j \in \mathbf{c}_r} d_e\left(\mathbf{x}_i, \mathbf{x}_j\right) \right\}.$$

- The Xie–Beni index ($\mathbf{XB}_{\downarrow}$) [20]. This is the ratio of the total variation to the minimum separation of the clusters and is defined by

$$\mathrm{XB}(\mathbf{C}) = \frac{\sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e^2\left(\mathbf{x}_i, \bar{\mathbf{c}}_k\right)}{N \min_{\mathbf{c}_k \in \mathbf{C}} \min_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \left\{ d_e^2\left(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r\right) \right\}}. \tag{7}$$

- The generalized Dunn indices[1] ($\mathbf{gD31}_{\uparrow}$, $\mathbf{gD41}_{\uparrow}$, $\mathbf{gD51}_{\uparrow}$, $\mathbf{gD33}_{\uparrow}$, $\mathbf{gD43}_{\uparrow}$, and $\mathbf{gD53}_{\uparrow}$) [3]. These indices are modifications of the original Dunn index, and include a combination of three variations of δ (a separation criterion)

$$\begin{aligned} \delta_3\left(\mathbf{c}_k, \mathbf{c}_r\right) &= \frac{1}{n_k \times n_r} \sum_{\mathbf{x}_i \in \mathbf{c}_k, \mathbf{x}_j \in \mathbf{c}_r} d_e\left(\mathbf{x}_i, \mathbf{x}_j\right), \\ \delta_4\left(\mathbf{c}_k, \mathbf{c}_r\right) &= d_e\left(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r\right), \\ \delta_5\left(\mathbf{c}_k, \mathbf{c}_r\right) &= \frac{1}{n_k + n_r} \left[\sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e\left(\mathbf{x}_i, \bar{\mathbf{c}}_k\right) + \sum_{\mathbf{x}_j \in \mathbf{c}_r} d_e\left(\mathbf{x}_j, \bar{\mathbf{c}}_r\right) \right] \end{aligned} \tag{8}$$

and two variations of Δ (a cohesion criterion)

$$\Delta_1\left(\mathbf{c}_k\right) = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{c}_k} \left\{ d_e\left(\mathbf{x}_i, \mathbf{x}_j\right) \right\},$$
$$\Delta_3\left(\mathbf{c}_k\right) = \frac{2}{n_k} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e\left(\mathbf{x}_i, \bar{\mathbf{c}}_k\right).$$

- The S_Dbw index ($\mathbf{SDbw}_{\downarrow}$) [10]. This is a ratio-type index that is based on the Euclidean norm $\|\mathbf{x}\| = \left(\mathbf{x}^{\mathsf{T}}\mathbf{x}\right)^{1/2}$, the standard deviation of the dataset $\sigma(\mathbf{X}) = 1/|\mathbf{X}| \sum_{\mathbf{x}_i \in \mathbf{X}} \left(\mathbf{x}_i - \bar{\mathbf{X}}\right)^2$, and the standard deviation of the partition $\mathrm{stdev}(\mathbf{C}) = 1/K \sqrt{\sum_{\mathbf{c}_k \in \mathbf{C}} \|\sigma\left(\mathbf{c}_k\right)\|}$. It is defined by

$$\begin{aligned} \mathrm{Scat}(\mathbf{C}) &= \frac{1}{K} \sum_{\mathbf{c}_k \in \mathbf{C}} \frac{\|\sigma(\mathbf{c}_k)\|}{\|\sigma(\mathbf{X})\|}, \\ \mathrm{Dbw}(\mathbf{C}) &= \frac{1}{K(K-1)} \sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \frac{\mathrm{den}(\mathbf{c}_k, \mathbf{c}_r)}{\max\{\mathrm{den}(\mathbf{c}_k), \mathrm{den}(\mathbf{c}_r)\}}, \\ \mathrm{SDbw}(\mathbf{C}) &= \mathrm{Scat}(\mathbf{C}) + \mathrm{Dbw}(\mathbf{C}), \end{aligned} \tag{9}$$

where

$$\begin{aligned} \mathrm{den}\left(\mathbf{c}_k\right) &= \sum_{\mathbf{x}_i \in \mathbf{c}_k} f\left(\mathbf{x}_i, \bar{\mathbf{c}}_k\right), \\ \mathrm{den}\left(\mathbf{c}_k, \mathbf{c}_r\right) &= \sum_{\mathbf{x}_i \in \mathbf{c}_k \cup \mathbf{c}_r} f\left(\mathbf{x}_i, \frac{\bar{\mathbf{c}}_k + \bar{\mathbf{c}}_r}{2}\right), \end{aligned}$$

[1] Bedeck and Pal proposed 18 variants, but we have selected those that were analyzed by Arbelaitz *et al.* [1], since these showed the best results.

and

$$f(\mathbf{x}_i, \bar{\mathbf{c}}_k) = \begin{cases} 0 & \text{if } d_e(\mathbf{x}_i, \bar{\mathbf{c}}_k) > \text{stdev}(\mathbf{C}) \\ 1 & \text{otherwise} \end{cases}.$$

- The CS index (**CS**$_\downarrow$) [6]. This is a ratio-type index that quantifies the cohesion by using the cluster diameters. The measure of separation is based on the nearest neighbor distance between prototypes. It is computed by

$$\text{CS}(\mathbf{C}) = \frac{\sum_{\mathbf{c}_k \in \mathbf{C}} \Delta(\mathbf{c}_k)}{\sum_{\mathbf{c}_k \in \mathbf{C}} \min_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \{d_e(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r)\}}, \tag{10}$$

where

$$\Delta(\mathbf{c}_k) = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathbf{c}_k} \max_{\mathbf{x}_j \in \mathbf{c}_k} \left\{ d_e(\mathbf{x}_i, \mathbf{x}_j) \right\}.$$

- The Enhanced Davies–Bouldin index (**DB2**$_\downarrow$) [13]. This variation of the DB index estimates the separation by the sum of the minimum distances between prototypes. It is defined by

$$\text{DB2}(\mathbf{C}) = \frac{1}{K} \sum_{\mathbf{c}_k \in \mathbf{C}} \frac{\max_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \{S(\mathbf{c}_k) + S(\mathbf{c}_r)\}}{\min_{\mathbf{c}_r \in \mathbf{C} \setminus \mathbf{c}_k} \{d_e(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r)\}}. \tag{11}$$

- The Score Function index (**SF**$_\uparrow$) [19]. This is a summation-type index in which the separation is measured by the sum of the distances from each centroid to the global prototype, and the cohesion is quantified by the mean distance of the objects to their respective centroids. It is computed by

$$\text{SF}(\mathbf{C}) = 1 - \frac{1}{e^{e^{\text{bcd}(\mathbf{C}) - \text{wcd}(\mathbf{C})}}}, \tag{12}$$

where

$$\text{wcd}(\mathbf{C}) = \sum_{\mathbf{c}_k \in \mathbf{C}} \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e(\mathbf{x}_i, \bar{\mathbf{c}}_k),$$

$$\text{bcd}(\mathbf{C}) = \frac{\sum_{\mathbf{c}_k \in \mathbf{C}} n_k \, d_e\left(\bar{\mathbf{c}}_k, \bar{\mathbf{X}}\right)}{N \times K}.$$

- The PBM index (**PBM**$_\uparrow$) [15]. This index is defined as the product of cohesion and separation measures. The former is the sum of all pattern distances in a cluster to their respective centroids, whereas the latter is quantified by the maximum distance between the centroids. The PBM index is given by

$$\text{PBM}(\mathbf{C}) = \left[\frac{\sum_{\mathbf{x}_i \in \mathbf{X}} d_e\left(\mathbf{x}_i, \bar{\mathbf{X}}\right)}{K \sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{x}_i \in \mathbf{c}_k} d_e(\mathbf{x}_i, \bar{\mathbf{c}}_k)} \times \mathcal{D}(\mathbf{C}) \right]^2, \tag{13}$$

where

$$\mathcal{D}(\mathbf{C}) = \max_{\mathbf{c}_k, \mathbf{c}_r \in \mathbf{C}} \{d_e(\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_r)\}.$$

- The Symmetry index ($\mathbf{Sym}_{\uparrow}$) [2]. This ratio-type index is an adaptation of the PBM index [15]: the cohesion is quantified by the sum of the point symmetry distances in the same cluster, and the separation is quantified by the maximum Euclidean distance between the centroids. This index is defined by

$$\mathrm{Sym}(\mathbf{C}) = \frac{\max_{\mathbf{c}_k,\mathbf{c}_r\in\mathbf{C}}\{d_e(\bar{\mathbf{c}}_k,\bar{\mathbf{c}}_r)\}}{K\sum_{\mathbf{c}_k\in\mathbf{C}}\sum_{\mathbf{x}_i\in\mathbf{c}_k} d_{ps}(\mathbf{x}_i,\mathbf{c}_k)}. \tag{14}$$

- Point Symmetry-Distance indices[2] ($\mathbf{SDB}_{\downarrow}$ and $\mathbf{SDunn}_{\uparrow}$) [18]. These are based on the PS distance and modify the cohesion criterion of the Davies–Bouldin and Dunn indices. The SDB index is computed the same way as the DB, but the computation of S is refined as follows:

$$S(\mathbf{c}_k) = \frac{1}{n_k}\sum_{\mathbf{x}_i\in\mathbf{c}_k} d_{ps}(\mathbf{x}_i,\bar{\mathbf{c}}_k). \tag{15}$$

  On the other hand, SDunn is a modification of DI, where the cohesion criterion Δ is defined by

$$\Delta(\mathbf{c}_k) = \max_{\mathbf{x}_i\in\mathbf{c}_k}\left\{d_{ps}(\mathbf{x}_i,\bar{\mathbf{c}}_k)\right\}. \tag{16}$$

- The COP index ($\mathbf{COP}_{\downarrow}$) [9]. This is a ratio-type index in which the cohesion is quantified by the sum of the distances of the patterns to their respective centroids, and the separation is based on the furthest neighbor distance between patterns. It is defined by

$$\mathrm{COP}(\mathbf{C}) = \frac{1}{N}\sum_{\mathbf{c}_k\in\mathbf{C}} n_k \frac{\frac{1}{n_k}\sum_{\mathbf{x}_i\in\mathbf{c}_k} d_e(\mathbf{x}_i,\bar{\mathbf{c}}_k)}{\min_{\mathbf{x}_i\notin\mathbf{c}_k}\max_{\mathbf{x}_j\in\mathbf{c}_k} d_e(\mathbf{x}_i,\mathbf{x}_j)}. \tag{17}$$

- The SV index ($\mathbf{SV}_{\uparrow}$) [22]. In this ratio-type index, the measure of cohesion is based on the distances from the border patterns to their respective centroids, and the measure of separation is based on the nearest neighbor distance between prototypes. It is computed by

$$\mathrm{SV}(\mathbf{C}) = \frac{\sum_{\mathbf{c}_k\in\mathbf{C}}\min_{\mathbf{c}_r\in\mathbf{C}\setminus\mathbf{c}_k}\{d_e(\bar{\mathbf{c}}_k,\bar{\mathbf{c}}_r)\}}{\sum_{\mathbf{c}_k\in\mathbf{C}}\left(\frac{10}{n_k}\right)\sum\max_{\mathbf{x}_i\in\mathbf{c}_k}\left(\frac{n_k}{10}\right)\{d_e(\mathbf{x}_i,\bar{\mathbf{c}}_k)\}}. \tag{18}$$

REFERENCES

[1] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognition*, 46(1):243–256, 2013.

[2] Sanghamitra Bandyopadhyay and Sriparna Saha. A Point Symmetry-Based Clustering Technique for Automatic Evolution of Clusters. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1441–1457, 2008.

[3] James C. Bezdek and N. R. Pal. Some New Indexes of Cluster Validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):301–15, 1998.

[4] T. Calinski and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974.

[5] Dongdong Cheng, Qingsheng Zhu, Jinlong Huang, Quanwang Wu, and Lijun Yang. A Novel Cluster Validity Index based on Local Cores. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):985–999, 2019.

[2] Saha and Bandyopadhyay presented nine indices but we have selected the two most representative.

[6] Chien Hsing Chou, M. C. Su, and E. Lai. A New Cluster Validity Measure and its Application to Image Compression. *Pattern Analysis and Applications*, 7(2):205–220, 2004.

[7] David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[8] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[9] Ibai Gurrutxaga, Iñaki Albisua, Olatz Arbelaitz, José I. Martín, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. SEP/COP: An Efficient Method to Find the Best Partition in Hierarchical Clustering Based on a New Cluster Validity Index. *Pattern Recognition*, 43(10):3364–3373, 2010.

[10] M. Halkidi and M. Vazirgiannis. Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 187–194, California, USA, 2001. IEEE.

[11] Lianyu Hu and Caiming Zhong. An internal validity index based on density-involved distance. *IEEE Access*, 7:40038–40051, 2019.

[12] L. J. Hubert and J. R. Levin. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. *Psychological Bulleting*, 83:1072–1080, 1976.

[13] Minho Kim and R. S. Ramakrishna. New Indices for Cluster Validity Assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[14] Shaoyi Liang, Deqiang Han, and Yi Yang. Cluster Validity Index for Irregular Clustering Results. *Applied Soft Computing*, 95:106583, 2020.

[15] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[16] Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-Based Clustering Validation. In *SIAM International Conference on Data Mining*, pages 839–847, Philadelphia, PA, 2014. Society for Industrial and Applied Mathematics.

[17] Peter J. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[18] Sriparna Saha and Sanghamitra Bandyopadhyay. Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(4):420–425, 2009.

[19] Sandro Saitta, Benny Raphael, and Ian F. C. Smith. A Bounded Index for Cluster Validity. In *Machine Learning and Data Mining in Pattern Recognition*, pages 174–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[20] Xuanli Lisa Xie and Gerardo Beni. A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.

[21] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and Enhancement of Internal Clustering Validation Measures. *IEEE Transactions on Cybernetics*, 43(3):982–994, 2013.

[22] Krista Rizman Žalik and Borut Žalik. Validity Index for Clusters of Different Sizes and Densities. *Pattern Recognition Letters*, 32(2):221–234, 2011.

[23] Qinpei Zhao and Pasi Fränti. WB-index: A sum-of-squares based index for cluster validity. *Data & Knowledge Engineering*, 92:77–89, 2014.