IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 11, 2019, accepted March 19, 2019, date of publication March 22, 2019, date of current version April 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2906949

An Internal Validity Index Based on Density-Involved Distance

LIANYU HU[1] AND CAIMING ZHONG[2]
[1]Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China
[2]College of Science and Technology, Ningbo University, Ningbo 315211, China

Corresponding author: Caiming Zhong (zhongcaiming@nbu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61573235 and Grant 61501270, in part by the Project of Key Disciplinary and New Major and Key Major in Ningbo City, and in part by the K. C. Wong Magna Fund in Ningbo University.

ABSTRACT It is crucial to evaluate the quality of clustering results in cluster analysis. Although many cluster validity indices (CVIs) have been proposed in the literature, they have some limitations when dealing with non-spherical datasets. One reason is that the measure of cluster separation does not consider the impact of outliers and neighborhood clusters. In this paper, a new robust distance measure, one into which density is incorporated, is designed to solve the problem, and an internal validity index based on this separation measure is then proposed. This index can cope with both the spherical and non-spherical structure of clusters. The experimental results indicate that the proposed index outperforms some classical CVIs.

INDEX TERMS Crisp clustering, cluster validity index, arbitrary-shaped clusters.

I. INTRODUCTION

Clustering is one of the most critical problems in machine learning, in which the task is to divide a dataset into structural groups without prior information [1]. Clustering has been widely studied in many research areas, such as image segmentation [2], data mining [3], and bioinformatics [4]. One of the standard components in clustering algorithms is the similarity measure, which is usually estimated by the distance or the density information of objects [5]. On this basis, clustering algorithms can be roughly classified into distance-based, density-based and hybrid clustering. For example, K-means, average-linkage, AP [6], and Ncut [7] use distance-based similarity; DBSCAN [8], ReCon-DBSCAN [9], and OPTICS [10] employ density-based similarity; while DPC [11], SNN-DPC [12] and robust path-based spectral clustering [13] use a mixed similarity. Different clustering algorithms are proposed to solve different types of applications, but there does not exist a unified clustering algorithm can cope with all applications [14]–[18]. As a result, it is necessary to find an effective way to evaluate the goodness of clustering before using a particular algorithm for further use [19], [20].

Clustering validation is a process of estimating how well a partition fits the underlying structure of the dataset. It has been recognized as a vital tool in clustering applications and widely used in clustering ensemble [21]–[23] and multi-objective clustering schemes [24], [25]. Clustering validation can be mainly classified into two types: internal and external. The main difference between internal and external validation is whether some external information, such as class labels, is used in the validation process. Unlike external validation, which is mostly used in supervised learning, internal validation is usually employed in unsupervised learning to evaluate the goodness of a clustering without using any external information. In practice, external information is often not available in clustering processes. Therefore, internal validations are usually the only option for clustering evaluation [26].

The conventional approach for evaluating the internal validation is to use validity indices [27]. Many cluster validity indices (CVIs) have been developed [28], including classic indices such as the Calinski-Harabasz index (CH) [29], Davies-Bouldin index (DB) [30], S_Dbw [31], $\mathcal{I}$ index [32], Dunn index [33], CDbw [34], and Silhouette index [35]; density-based indices [36], [37]; and newly developed indices, such as WB index [38], CVNN [26], RTI [39], CSP [40], and the Local Cores-based Cluster Validity index (LCCV) [41].

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

TABLE 1. CVI representatives.

CVI type	Notions	Selected definition	CVIs
Single center	μ, μ_i	$\mu = mean(X), \mu_i = mean(x \in C_i)$	CH, DB, WB, S_Dbw, $\mathcal{I}$
Multiple centers	μ_i^k	$\mu_i^k = mean(x \in C_i^k)$	RTI
Single non-center	δ_i	$(\delta_i, \delta_j) = \arg\min_{x \in C_i, y \in C_j} d(x, y)$	Dunn
Multiple non-centers	δ_i^k, x	δ_i^k are (i) well-scattered objects in C_i [34] (ii) located at the edge of C_i [26]	CDbw, CVNN, Silhouette

$mean(x)$ is the mean of x.

μ, μ_i, μ_j and μ_i^k are the single center of X, C_i, C_j and multiple centers of C_i, respectively.

δ_i, δ_j and δ_i^k are the single non-central object of C_i, C_j and multiple non-central objects of C_i, respectively.

However, most of them are effective only when applied to a dataset with a simple cluster structure, such as spherical clusters or well-separated clusters, and become degraded when the dataset has a complicated structure, such as arbitrary-shaped clusters or clusters with outliers. For example, the CH index is estimated based on the distances from the objects in a cluster to its centroid and distances from the cluster centroids to the global centroid and may fail when a cluster centroid is outside of the cluster. The Dunn index is based on the nearest neighbor distance and maximum cluster diameter, which is only suitable for clusters well-separated with respect to distance. Furthermore, these indices perform well only for datasets that are composed of spherical shapes or have structures without outliers.

We propose a robust and relatively universal Cluster Validity index based on Density-involved Distance (CVDD), and then use it to choose the best partition and determine the optimal number of clusters.

The remainder of the paper is organized as follows: Section 2 discusses related work. Density-involved distance is proposed in Section 3. The internal cluster validation index based on density-involved distance is presented in Section 4. Experimental results on synthetic and real datasets are presented in Sections 5. Finally, some concluding remarks are given in the last section.

II. RELATED WORK

In this section, some basic concepts of internal validity indices are introduced, and then some representative components of the indices are extracted. With these components, the limitations of the indices are described.

A. COMPONENTS OF INTERNAL VALIDITY INDICES

Since clustering tries to make pairwise objects within a cluster similar and those across clusters dissimilar, internal validity indices are usually designed to measure the intra-cluster compactness and inter-cluster separation simultaneously. Intra-cluster compactness depicts how closely the objects in a cluster are related, and it can be quantified as the overall deviation [24]. Inter-cluster separation, on the other hand, depicts the separation degree of two clusters and can be measured by distances of objects between the two clusters. Normally, a cluster with high compactness and separation is of high quality. The most common measures based on these two aspects for spherical clusters are summarized in [38].

However, some studies pay more attention to the design of inter-cluster separation [42], [43], and some even directly claim that inter-cluster separation is more important than intra-cluster compactness [44].

In general, existing validity indices use some representatives to evaluate the separation [26], where the representatives could be medoids or means. Suppose π is a clustering of dataset $X = \{x_1, x_2, \cdots, x_N\}$, where $\pi = \{C_1, C_2, ..., C_K\}$; $d(x_i, x_j)$ is the Euclidean distance between object x_i and x_j, $|C_i|$ is the number of objects in C_i, N is the number of the objects in X and K is the number of the clusters of π. The separation between C_i and C_j is defined as the distance between representatives, namely $sep(C_i, C_j) = d(a, b)$ where a, b are the representative of C_i and C_j respectively. As shown in Table 1, the validity indices can be categorized into four groups according to the following four kinds of representatives:

1) Single center
2) Multiple centers
3) Single non-center
4) Multiple non-centers

However, the above representatives, which are located in or far from the cluster center, have some drawbacks, and the resulting limitations of CVIs based on these representatives will be discussed in the next subsection.

B. LIMITATIONS OF THE EXISTING REPRESENTATIVES

A single center representative is not enough to represent an entire cluster that has non-spherical structure. To address this problem, RTI divides a cluster into several subclusters, and uses multiple centers to describe the geometrical structure of the cluster [39].

However, sometimes the distance between center-based representatives of two clusters cannot measure the separation well. In Fig. 1, two groups of pairwise clusters have the same distance between center representatives but different separations, namely, $d(\mu_1, \mu_2) = d(\mu_3, \mu_4)$, $sep(C_1, C_2) > sep(C_3, C_4)$, where $sep(C_i, C_j)$ denotes the separation of C_i and C_j. Thus, center representative based CVIs, such as CH, DB, WB, S_Dbw, $\mathcal{I}$ and RTI, may fail to reflect the separation. But single non-center representatives can depict the separations of the case in Fig. 1, that is, $d(\delta_1, \delta_2) > d(\delta_3, \delta_4)$. Therefore, to measure the separations, a non-center representative is better than a center representative when two

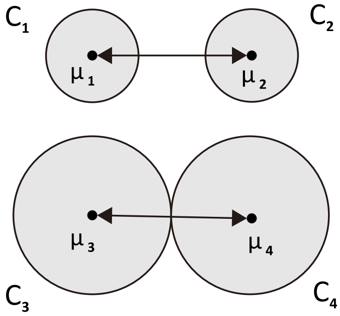


FIGURE 1. Intercluster separation between clusters.

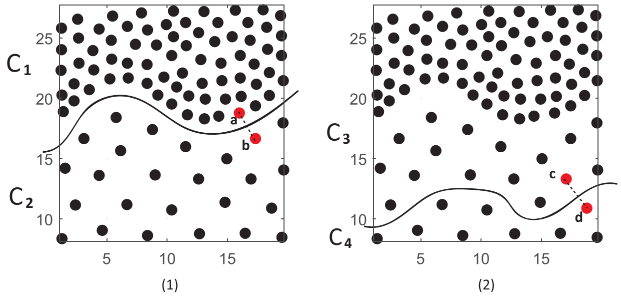


FIGURE 3. Impact of density-separated clusters.

clusters are close but the two corresponding centers are far away.

A single non-center representative does not aim to represent the entire geometrical structure of a cluster, but the geometrical information of local portions especially those that are adjacent to the other clusters.

However, a single non-center representative might be sensitive to outliers. In Fig. 2(1), it is obvious that there exist a gap between clusters, and a and b are the non-center representatives of C_1 and C_2 respectively and $sep(C_1, C_2) = d(a, b)$. In Fig. 2(2), two outliers, p and q, are between C_1 and C_2. The two representatives are selected by Dunn's index as p and q, and the separation $d(p, q)$ is quite different from that in Fig. 2(1). Therefore, the cluster separation estimated by the single non-center may be changed when outliers or noise is introduced.

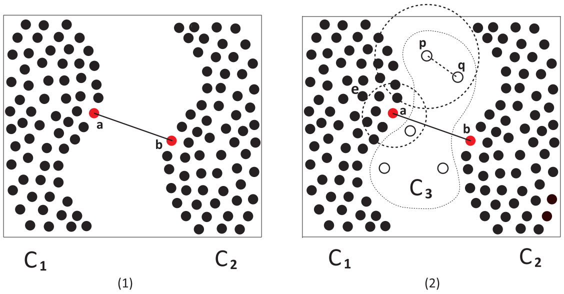


FIGURE 2. Impact of outliers.

To alleviate the above problem, CDbw [34], CVNN [26], and Silhouette scheme [35], VCN [45], and LCCV [41] use multiple non-center representatives to measure the separation more accurately. Unfortunately, this method is only workable when there exist a small number of outliers between the two clusters. If more outliers are selected as representatives, for example in Fig. 2(2), when p, q and the other three outliers are selected, it does not depict the separation well.

Moreover, the above non-center representatives cannot recognize density-separated clusters (also known as density gradient problem [46]). In Fig. 3, there are two clusterings, $\pi_1 = \{C_1, C_2\}$ and $\pi_2 = \{C_3, C_4\}$, where C_1, C_2 are density-wise well-separated and C_3, C_4 are not. Dunn selects a, b and c, d as the corresponding representatives. The separations of C_1, C_2 and C_3, C_4 are represented by $d(a, b)$ and $d(c, d)$, respectively. But $d(a, b) = 2.47$ and $d(c, d) = 2.94$, which means the separation of C_3, C_4 is greater than that of C_1, C_2, but this does not make intuitive sense, since C_1, C_2 constitute a good clustering and C_3, C_4 do not.

To eliminate the impact of outliers and the problems with density-separated clusters in the above separation situations (clusters with lesser degree of *overlap* [47]), we design a new density-involved distance measure in the next section.

III. A NEW DENSITY-INVOLVED DISTANCE

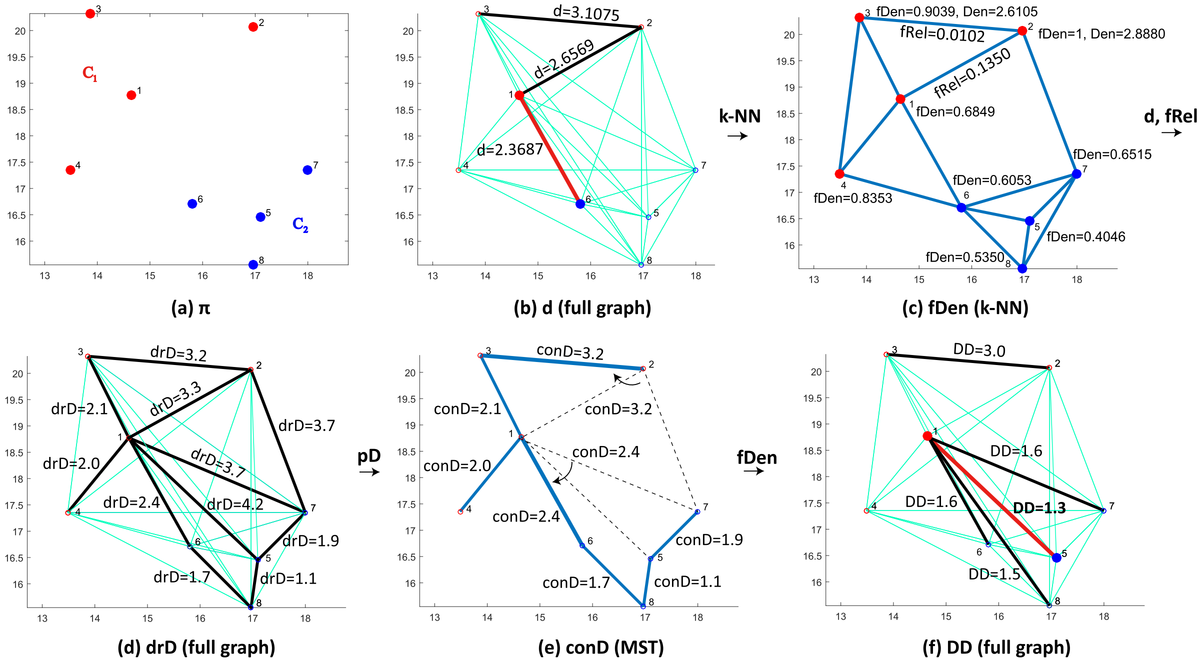
In this section, density information is employed to improve the performance of the distance measure in non-centers representatives. The main idea is to use two concepts from DBSCAN: core objects and density connectivity, in which the first concept is useful to recognize outliers (discussed in subsection A) and the second one is helpful to differentiate the density-separated clusters (discussed in subsection B). The newly defined outlier factor *fDen*, and mutual density factor *fRel* are combined with a graphical-based approach into a novel density-involved distance in subsection C.

An overview of the density-involved distance is illustrated in Fig. 4.

A. USING DENSITY ESTIMATION TO COPE WITH OUTLIERS

As discussed above, one limitation of the existing CVIs is that the definition of the separation may be sensitive to outliers between the two clusters. Normally, any outlier should not act as a representative for the evaluation of the separation.

Definition 1: C is said to be a **region**, if for every $x \in C$, the density of x is approximately equal to a certain value. For example, in Fig. 2(2), every object in C_1 has a similar density to that of a, while objects in C_3 have a similar density to that of p. C_1 and C_3 are different regions with respect to density. The main concept of "region" is based on the density attractor notion [48], which implies dense regions recognized as clusters are surrounded by regions with low density. Since outliers usually come from low-density regions [8], which deviate significantly from higher-density regions [49], [50], the representatives should be selected from regions of high density. Therefore, density should be taken into account as a key factor in the design of the separation.



(a) π

(b) d (full graph)

(c) fDen (k-NN)

(d) drD (full graph)

(e) conD (MST)

(f) DD (full graph)

FIGURE 4. **The scheme of Density-involved distance. (a) shows a partition of $\pi = \{C_1, C_2\}$ where $C_1 = \{x_1, x_2, x_3, x_4\}$, $C_2 = \{x_5, x_6, x_7, x_8\}$. (b) the Euclidean distances where $\min d(C_1, C_2) = d(x_1, x_6)$. (c) includes density information such as outlier factor *fDen* (normalized *Den*) estimated by 3-NN graph and mutual density factor *fRel*. (d) is the directly density-reachable distance *drD* estimated from *Den*, *d* and *fRel*. (e) is the connectivity distance *conD* estimated from the *drD*'s path-based *pD* transformation based on a minimum spanning tree. (f) is the final density-involved distance *DD* with the combination of *conD*, *fDen*. where $\min DD(C_1, C_2) = DD(x_1, x_5)$.**

Definition 2: Let $\mathcal{N}_i$ be the k nearest neighbors (k-NN) of object x_i, $d(x_i, x_j)$ be the Euclidean distance between object x_i and x_j, the **density estimation** of object x_i is defined as:

$$Den(x_i) = \frac{1}{k} \sum_{x_j \in \mathcal{N}_i} d(x_i, x_j). \tag{1}$$

This density estimation is a neighbor-based approach [51] used in [52]. The smaller the $Den(x_i)$, the denser the x_i; e.g., in Fig. 2(2), the size of a's neighborhoods are smaller than those of p's while a is denser than p. Accordingly, the density of a region C can be evaluated as $Den(C) = \frac{1}{|C|} \Sigma_{x_i \in C} Den(x_i)$.

Assumption 1: Let $dis(x_i, x_j)$ be the dissimilarity of objects x_i and x_j. Suppose $d(x_i, x_j) = d(x_p, x_q)$, $x_i, x_j \in C_1$, $x_p, x_q \in C_2$. If $Den(C_1) < Den(C_2)$ then $dis(x_i, x_j) < dis(x_p, x_q)$.

The above assumption is based on an intuition: Dense regions are more likely to be recognized as clusters. As clusters are surrounded by sparse regions, objects in a dense region are more likely to be compact or similar while objects in a sparse region are more separated. When $Den(C_1) < Den(C_2)$, even if $d(x_i, x_j) = d(x_p, x_q)$, x_i and x_j are more likely density-connected than x_p and x_q. In Fig. 2(2), for example, two clusters C_1 and C_2 are separated by a sparse region C_3, objects a and e are in C_1, while objects p and q are in C_3. When the number of nearest neighbors k is set to 8, we have $Den(C_1) < Den(C_3)$. Since $d(a, e) = d(p, q)$, the dissimilarity $dis(p, q)$ is larger than $dis(a, e)$. Accordingly, this assumption is reasonable.

Definition 3: The normalized inverse of $Den(x_i)$ is called its **outlier factor**:

$$fDen(x_i) = \frac{Den(x_i)}{\max_{x_j \in X} Den(x_j)}. \tag{2}$$

The outlier factor, $fDen(x_i) \in (0, 1]$, reflects the score of how likely x_i can be viewed as an outlier.

Definition 4: The outlier-penalized distance, abbreviated as ***out-distance***, of object x_i and x_j is defined as in:

$$outD(x_i, x_j) = \sqrt{fDen(x_i) \cdot fDen(x_j)} \cdot d(x_i, x_j). \tag{3}$$

The above defined out-distance can be used as a dissimilarity measure or distance measure.

Suppose x_i and x_j are in a dense region, x_p and x_q are in a sparse region, and $d(x_i, x_j) = d(x_p, x_q)$. We have $\sqrt{fDen(x_p) \cdot fDen(x_q)} > \sqrt{fDen(x_i) \cdot fDen(x_j)}$, and $outD(x_p, x_q) > outD(x_i, x_j)$. At the same time, according to **Assumption 1**, $dis(x_p, x_q) > dis(x_i, x_j)$.

Moreover, suppose $x_i, x_j \in C_1$, $x_p \notin C_1$, $Den(x_p) > Den(C_1)$, and $d(x_i, x_j) = d(x_p, x_i)$. As $\sqrt{fDen(x_p)} > \sqrt{fDen(x_j)}$, we have $outD(x_p, x_i) > outD(x_i, x_j)$. Similarly, if $d(x_i, x_j) = d(x_p, x_j)$, then $outD(x_p, x_j) > outD(x_i, x_j)$. Therefore, under **Assumption 1**, $outD(x_i, x_j)$ can be viewed as $dis(x_i, x_j)$.

Obviously, out-distance $outD(x_i, x_j)$ satisfies:

1) $outD(x_i, x_j) \geq 0$
2) $outD(x_i, x_j) = 0 \leftrightarrow x_i = x_j$
3) $outD(x_i, x_j) = outD(x_j, x_i)$

but does not satisfy (iv) $outD(x_i, x_z) \leq outD(x_i, x_j) + outD(x_j, x_z)$, since $outD(x_i, x_j)$ is not linearly defined. Therefore, $outD(x_i, x_j)$ is a distance measure but not a metric.

The main advantage of out-distance is that this distance measure is robust to outliers, and this performance is illustrated in Fig. 5. We use two-dimension multidimensional scaling (MDS)[1] to generate the transformed space by the distance measure and verify the performance. The original dataset contains six outliers and is shown in Fig. 5(a). The transformed dataset by MDS with out-distance is shown in Fig. 5(b). The outliers between two clusters in original space are removed, and the separation in the transformed space is more prominent compared with the separation in the original space.

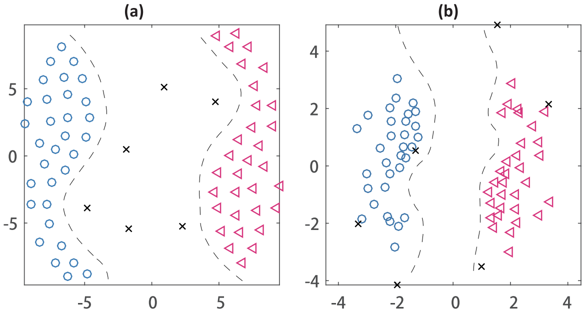


FIGURE 5. MDS transformation of distance with the impact of outliers.

B. USING MUTUAL DENSITY TO COPE WITH DENSITY-SEPARATED CLUSTERS

In general, two density-separated clusters should not be merged into one cluster, as they are not directly density-reachable [8]. However, separation measures in the existing CVIs usually favor distance-separated clusters but not density-separated clusters, since factors of density separation are not considered well. In this subsection, a density separation measure is carefully defined to recognize the density-separated clusters.

Definition 5: The relative density of x_i with respect to x_j is defined as:

$$Rel(x_i, x_j) = \frac{Den(x_i)}{Den(x_j)}. \tag{4}$$

If $Den(x_i) > Den(x_j)$, the relative density $Rel(x_i, x_j)$ is greater than 1, which means x_i is of a relatively high density.

Definition 6: The **mutual density factor** between x_i and x_j is defined as:

$$fRel(x_i, x_j) = 1 - e^{-[Rel(x_i,x_j)+Rel(x_j,x_i)-2]}. \tag{5}$$

The mutual density factor, $fRel(x_i, x_j) \in [0, 1)$, reflects the score of how likely the x_i and x_j are density-separated. If and only if $Den(x_i) = Den(x_j)$, then $fRel(x_i, x_j) = 0$. The power $-[Rel(x_i, x_j) + Rel(x_j, x_i) - 2]$ is always less than or equal to 0, which makes the factor change with the relative density but not with respect to the individual density of x_i or x_j.

As $fRel(x_i, x_j)$ measures the difference between x_i and x_j respect to relative density, one may use it to define a new density-reachable distance.

Definition 7: The directly density-reachable distance, abbreviated as ***dr-distance***, between x_i and x_j is defined as:

$$drD(x_i, x_j) = d(x_i, x_j) + relD(x_i, x_j) \tag{6}$$

where $relD(x_i, x_j) = fRel(x_i, x_j) \cdot nD(x_i, x_j)$, and $nD(x_i, x_j)$ is a reference. Addend $relD(x_i, x_j)$ is a penalty distance, and the purpose of its introduction is to make the dr-distance $drD(x_i, x_j)$ larger than $d(x_i, x_j)$ when the mutual density factor $fRel(x_i, x_j)$ is large. That is to say, when the density difference between x_i and x_j is large, the distance between them is mandatorily changed into a large one so that the separability is strengthened. If x_i and x_j have the same density, then $fRel(x_i, x_j) = 0$ and $drD(x_i, x_j) = d(x_i, x_j)$.

We set $nD(x_i, x_j)$ to $Den(x_i) + Den(x_j)$, because on the one hand $nD(x_i, x_j)$ is adaptive with respect to the varied individual densities of x_i and x_j, and on the other hand, $Den(x_i) + Den(x_j)$ approaches the sum of the two k-NN radii and is relatively moderate.

If $d(x_i, x_j) \gg relD(x_i, x_j)$, then $drD(x_i, x_j)$ is not sensitive to $fRel(x_i, x_j)$, and the separability of x_i and x_j is determined by the Euclidean distance rather than density. Moreover, if x_i and x_j are near enough, for example, $x_i \in \mathcal{N}_j$ or $x_j \in \mathcal{N}_i$, then $drD(x_i, x_j)$ is sensitive to $fRel(x_i, x_j)$, and the impact of density-separated clusters will play an important role in $drD(x_i, x_j)$.

The main advantage of dr-distance is that this distance measure is sensitive to density-separated scenarios. This performance is illustrated in Fig. 6. The original dataset contains two density-separated clusters shown in Fig. 6(a). The dataset transformed by classic two-dimensional MDS with dr-distance is shown in Fig. 6(b). The separability in the transformed space is greater than that in the original space.

C. COMBINATION OF OUTLIER FACTOR AND MUTUAL DENSITY FACTOR

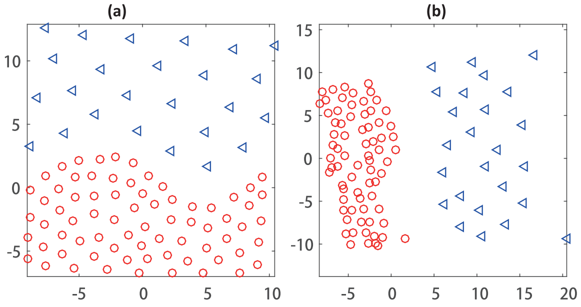


FIGURE 6. MDS transformation of distance with the impact of density-separated clusters.

[1] Multidimensional scaling is a powerful statistical method for visualizing the information contained in a dissimilarity or distance matrix [53]. Generally, the aim of MDS is to place objects in a low-dimensional space while preserving the pairwise dissimilarity between objects as well as possible. Note that the dissimilarities is preserved in the transformed space by MDS while the orientation of mapping is arbitrary.

To cope with the impact of outliers and the density-separated clusters simultaneously, we should consider both the outlier factor and the mutual density factor. We can combine the above **out-distance** and **dr-distance** to form a new density-involved distance. The combination is as follows.

Definition 8: The path-based distance [54] between objects x_i and x_j in C_i is defined as in:

$$pD(x_i, x_j) = \min_{p \in P_{ij}} \left\{ \max_{1 \le h < |p|} d(x_{[h]}, x_{[h+1]}) \right\} \tag{7}$$

where P_{ij} denotes the set of all paths from object x_i to object x_j, $[h]$ denote the $[h]$th object along a path p from x_i to x_j ($P_{ij} = \{p_1, p_2, \ldots, p_t\}$, $p = [x_{[1]}, x_{[2]}, \ldots, x_{[h]}, x_{[h+1]}, \ldots x_{[p]}]$ is a sequence along x_i to x_j, $x_{[1]} = x_i, x_{[p]} = x_j$, and the $[h+1]$th object is the nearest object with $[h]$th object). For each path $p \in P_{ij}$, the maximum weight $d(x_{[h]}, x_{[h+1]})$ along the path p is selected to be the weight of the path p, then the minimum weight from all paths P_{ij} is selected to be path-based distance between x_i and x_j. The path-based distance can be computed from a minimum spanning tree (MST) of d [10], [55].

The main property of the path-based distance is that the similarity or dissimilarity is transitive: "My neighbor's neighbor can also be my neighbor." [56] As the density-reachability between objects is also transitive [10], we use the path-based method to define the following global distance.

Definition 9: The **connectivity distance** between x_i and x_j is defined as:

$$conD(x_i, x_j) = \min_{p \in P_{ij}} \left\{ \max_{1 \le h < |p|} drD(x_{[h]}, x_{[h+1]}) \right\}. \tag{8}$$

The connectivity distance between x_i and x_j is the minimum of the maximum dr-distance weights of all the path P_{ij}. Compared with the definition of path-based distance in **Definition 8**, connectivity distance is obtained by replacing the weight of original Euclidean distance $d(x_{[h]}, x_{[h+1]})$ with the weight of dr-distance $drD(x_{[h]}, x_{[h+1]})$.

This path-based like distance measure satisfies the following assumption of clustering, called cluster assumption: objects are likely in different clusters if there is a path connecting them passing through regions of low density [57], [58]. One can see that neighboring weights or dr-distance weights imply the local cluster structure.

Definition 10: The **density-involved distance** between x_i and x_j is defined as:

$$DD(x_i, x_j) = \sqrt{fDen(x_i) \cdot fDen(x_j)} \cdot conD(x_i, x_j). \tag{9}$$

Compared with the definition of out-distance in **Definition 4**, the density-involved distance is obtained by replacing the original Euclidean distance $d(x_i, x_j)$ with the connectivity distance $conD(x_i, x_j)$.

The performance of density-involved distance is demonstrated in Fig. 7. Fig. 7(a) is the original dataset, which contains two density-separated clusters in the left side and two distance separated clusters containing six outliers in the right side. The dataset is transformed by classic two-dimensional MDS with density-involved distance. The transformed version is shown in Fig. 7(b), from which one can see that the left two clusters are distance well-separated and the right two clusters are also distance well-separated, since the outliers are not in the middle of the two clusters. This performance indicates that the density-involved distance can remove the negative impact of outliers and be sensitive to density-separated clusters simultaneously.

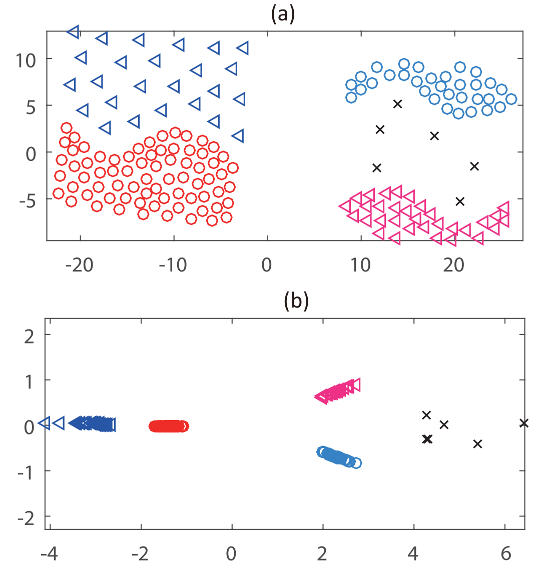
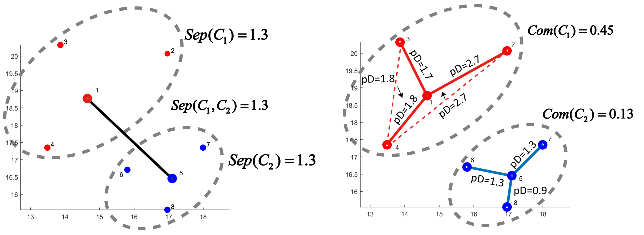


FIGURE 7. MDS transformation of distance with the impact of outliers and density-separated clusters.

IV. CVDD: A NEW CLUSTERING VALIDITY INDEX

As $DD(x_i, x_j)$ is robust to outliers and sensitive to density-separated clusters, it can be used to measure the separability of two clusters in an internal validity index. An overview of CVDD is seen in Fig. 8.



$$CVDD(\pi) = \frac{Sep(C_1) + Sep(C_2)}{Com(C_1) + Com(C_2)} = 4.48$$

FIGURE 8. The scheme of CVDD between two clusters. The π is from Fig. 4(a). The separation of π is based on minimum *DD* between clusters. The compactness of π is based on the weights of C_1, C_2's *pD*.

Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ be M clusterings on a dataset X, $\pi_i = \{C_{i1}, C_{i2}, \ldots, C_{iK_i}\}$ a clustering from Π, C_{iK_i} a cluster of π_i and K_i the number of the clusters of π_i.

Definition 11: The separation between C_i and C_j is defined as:

$$sep(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} DD(x_i, x_j). \tag{10}$$

Like the Dunn index [33], we use the minimum pairwise distance between clusters to represent the separation between clusters.

Then, the separation between C_i and other clusters is defined as:

$$sep(C_i) = \min_{x_j \in \pi, x_j \notin C_i} sep(C_i, C_j). \tag{11}$$

When evaluating the compactness of a cluster, we do not take the density factors into account, as compactness is usually irrelevant to outliers and density separation. But we use the path-based distance defined in **Definition 8** as a measure to evaluate the compactness; the reason is that the path-based distance is transferable within a cluster. Some statistical properties of path-based distance are used to define the compactness as follows.

Definition 12: The compactness of C_i is defined as:

$$com(C_i) = \frac{1}{|C_i|} \cdot STD(C_i) \cdot Mean(C_i) \tag{12}$$

where $Mean(C_i) = \frac{1}{|C_i|} \cdot \sum_{x_i, x_j \in C_i} pD(x_i, x_j)$, $STD(C_i) = \sqrt{\frac{1}{|C_i|-1} \cdot \sum_{x_i, x_j \in C_i} (pD(x_i, x_j) - Mean(C_i))}$.

In the above definition, three factors are included. $Mean(C_i)$ and $STD(C_i)$ are the mean and the standard deviation respectively of a cluster with respect to the path-based distance, and $\frac{1}{|C_i|}$ acts as a penalty factor.

It is intuitive that a compact cluster should have a relatively small mean and standard deviation for its path-based distances. While $\frac{1}{|C_i|}$ means that when two clusters have the same mean and standard deviation, the cluster with more objects should be relatively more compact.

Definition 13: The CVDD index is defined as in:

$$CVDD(\pi) = \frac{\sum_{i=1}^{K} sep(C_i)}{\sum_{i=1}^{K} com(C_i)}. \tag{13}$$

The above definition is the proposed internal validity index. It is the average performance of π with respect to individual clusters. Another option of designing the final index is to compute the performance of a single cluster first and then average the performances as the whole performance of π. But this option may be sensitive to single clusters with extreme cases such as when the compactness approaches 0.

Obviously, the larger $CVDD(\pi)$, the better quality of the partition π.

To sum up,[2] the process of computing the CVDD of a given partition is shown in Algorithm 1

Algorithm 1 CVDD

```
Input: $\pi = \{C_1, C_2, ..., C_K\}$ of $X$, number of nearest
       neighbor $k$
Output: Validity index $CVDD(\pi)$

$CVDD(\pi) \leftarrow null$
Compute the density estimation $Den$ as in Eq. 1
Compute the outlier factor $fDen$ as in Eq. 2
Compute the mutual density factor $fRel$ as in Eq. 5
Compute the density-involved distance $DD$ as in Eq. 9
for $i \leftarrow 1$ to $K$ do
    Compute the separation $sep[i]$ of $C_i$ as in Eq. 11
    Compute the compactness $com[i]$ of $C_i$ as in Eq. 12
Compute the validity index $CVDD$ of $\pi$ as in Eq. 13.
```

Algorithm for determining the optimal partition. For a set of partitions Π, the optimal clustering π_{OP} can be given as:

$$\pi_{OP} = \arg \max_{1 \leq i \leq M} \{CVDD(\pi_i)\} \tag{14}$$

The optimal partition algorithm is described in Algorithm 2.

Algorithm 2 CVDD-OP

```
Input: A set of partitions $\Pi = \{\pi_1, \pi_2, ..., \pi_M\}$
Output: The optimal partition $\pi_{OP}$

$\pi_{OP} \leftarrow null$
$cvi \leftarrow 0$
for $i \leftarrow 1$ to $M$ do
    if $cvi < CVDD(\pi_i)$ then
        $cvi \leftarrow CVDD(\pi_i)$
        $\pi_{OP} \leftarrow \pi_i$
```

To obtain the CVDD index, two main graph constructions are required: k-NN graph and MST. The time complexity of k-NN and MST are both $O(N^2)$. In computing density-involved distance, k-NN is used for density estimation and MST is used to figure out the path-based distance. Thus, the time complexity of DD is $O(N^2)$. For M partitions, MST is used for computing the compactness of each partition while DD is global for all partitions. Therefore, the overall time complexity of computing CVDD index for M partitions is $O(N^2 + MN^2)$.

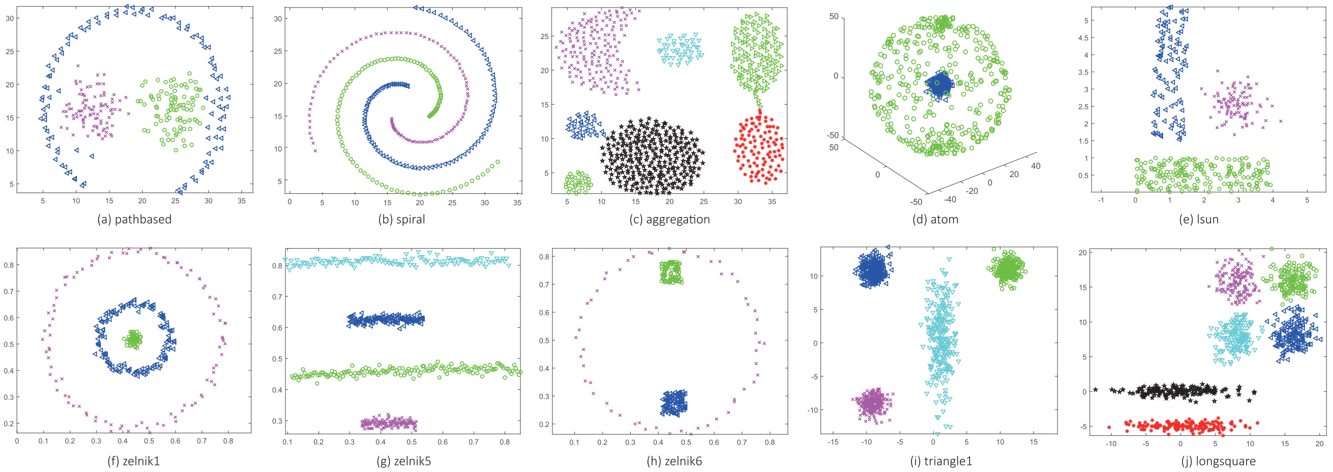
V. EXPERIMENTAL RESULTS

As CVIs can usually be used to select the optimal partition from multiple candidates and determine the optimal number of clusters (K), we evaluate the performance of CVDD from these two aspects. Eight well-known CVIs are selected for comparison and shown in Table 2.

A. DATASETS USED

Thirty datasets are used to test the CVDD algorithm. Ten non-spherical datasets are downloaded from

[2] The source code of this paper is available at https://github.com/hulianyu/CVDD



(a) pathbased (b) spiral (c) aggregation (d) atom (e) lsun

(f) zelnik1 (g) zelnik5 (h) zelnik6 (i) triangle1 (j) longsquare

FIGURE 9. **Ten non-spherical clusters.**

TABLE 2. **CVIs compared. The middle column denotes that an index is optimal when its value is minimum or maximum.**

Name	Optimal value	Reference
CVNN	Min	[26]
WB	Min	[38]
Silhouette	Max	[35]
CH	Max	[29]
DB	Min	[30]
Dunn	Max	[33]
S_Dbw	Min	[31]
$\mathcal{I}$	Max	[32]

TABLE 3. **The description of the data sets.**

Datasets	Classes (K)	Objects (N)	Dimensions	Sources
non-spherical clusters				
pathbased	3	300	2	[13]
spiral	3	312	2	[13]
aggregation	7	788	2	[63]
atom	2	800	3	FCPS [59]
lsun	3	400	2	FCPS [59]
zelnik1	3	299	2	[64]
zelnik5	4	512	2	[64]
zelnik6	3	238	2	[64]
triangle1	4	1000	2	[65]
longsquare	6	900	2	[24]
spherical clusters				
s1	15	5000	2	[66]
s2	15	5000	2	[66]
s3	15	5000	2	[66]
s4	15	5000	2	[66]
dim32	16	1024	32	[67]
dim64	16	1024	64	[67]
dim128	16	1024	128	[67]
a1	20	3000	2	[68]
a2	35	5250	2	[68]
unbalance	8	6500	2	[69]
classification				
Iris	3	150	4	UCI [61]
Ionosphere	2	351	33	UCI [61]
Wine	3	178	13	UCI [61]
Glass	6	214	9	UCI [61]
WDBC	2	569	30	UCI [61]
Movement	15	360	90	UCI [61]
Vertebral	3	310	6	UCI [61]
Yeast	10	1484	8	UCI [61]
Leukemia	3	72	1081	[62]
Seeds	3	210	7	UCI [61]

Shape sets,[3] Fundamental clustering problem suite (FCPS) [59][4] and Tomas Barton's clustering benchmark[5] and illustrated in Fig. 9. The ten spherical datasets are all from the Clustering basic benchmark [47][3], and the ten real datasets or classification datasets are from UCI [60] and [61]. In Fig. 9, we can see that the separation of pathbased, aggregation, zelnik6 and longsquare datasets may be affected from outliers and density-separated clusters. Detailed information on these datasets such as the number of clusters (K), the number of objects (N) and the number of dimensions are depicted in Table 3.

B. PERFORMANCE MEASURES

If the ground truth partition of X is available, it is usually used to evaluate the performance of a partition generated by a clustering algorithm. Suppose $\pi^* = \{C_1^*, C_2^*, \ldots, C_K^*\}$ is the ground truth of X. Taking π^* as a reference, we measure the performance of CVIs by measuring the quality of the best partition (π_{OP}), which is determined by CVIs.

We use cluster-level **Centroid Index (CI)** [69] and point-level **Purity** [70] to measure the quality of π_{OP}, and the similarity between π_{OP} and ground truth. CI provides a clear interpretation about the dissimilarity between π_{OP} and π^* in cluster-level structure (e.g., CI $= 1$ demonstrates one cluster difference in the global allocation of π_{OP}, π^*). Purity is equivalent to Accuracy [71] and widely cited as

[3] http://cs.uef.fi/sipu/datasets/

[4] https://www.uni-marburg.de/fb12/arbeitsgruppen/datenbionik/data

[5] https://github.com/deric/clustering-benchmark

TABLE 4. **The results measured by CI. In the second column, the best partitions are determined by CI from 40,000 partitions generated by Ncut with different parameters. The remaining nine columns are the values of CI of the best partitions π_{OP} selected by corresponding CVIs. The numbers in bold font means the best in all of the CVIs for a dataset.**

Data sets	Best partitions	CVIs								
		CVDD	CVNN	WB	Silhouette	CH	DB	Dunn	S_Dbw	$\mathcal{I}$
non-spherical clusters										
pathbased	0	**0**	1	1	1	1	1	**0**	1	1
spiral	0	**0**	2	1	1	1	1	**0**	1	2
aggregation	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
atom	0	**0**	1	1	1	1	1	**0**	1	1
lsun	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**
zelnik1	0	**0**	1	2	1	2	2	**0**	2	1
zelnik5	0	**0**	2	2	2	2	3	**0**	2	2
zelnik6	0	**0**	1	1	1	1	1	**0**	1	1
triangle1	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
longsquare	0	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**
classification										
Iris	0	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Ionosphere	0	**0**	1	**0**	1	**0**	1	1	1	**0**
Wine	0	**0**	**0**	**0**	1	**0**	**0**	1	1	**0**
Glass	2	3	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**
WDBC	0	**0**	1	**0**	1	**0**	1	1	**0**	1
Movement	2	**5**	6	6	6	6	**5**	6	6	6
Vertebral	1	**1**	2	**1**	**1**	**1**	**1**	**1**	2	**1**
Yeast	3	**3**	4	4	6	4	7	7	4	**3**
Leukemia	0	**0**	1	1	1	1	2	2	1	2
Seeds	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Number of the best partitions		**19**	6	11	8	11	10	14	7	12

classification accuracy [68] in current clustering research, such as [72], [73], [74].

Let c_i and c_i^* denote the clustering label and the ground truth of x_i respectively, then we have

$$Purity(\pi, \pi^*) = \frac{\sum_i \Delta(c_i, map(c_i^*))}{N} \tag{15}$$

where $\Delta(c_i, c_i^*) = 1$ if $c_i = c_i^*$, and $\Delta(c_i, c_i^*) = 0$ otherwise, and $map(c_i^*)$ is the best mapping function that permutes clustering labels to match the ground truth labels [71]. Intuitively, $Purity \in (0, 1]$. Normally, larger $Purity(\pi_{OP}, \pi^*)$ values indicate better quality of the selected partition.

Using the above mapping function, the CI [69] is generalized for arbitrary-shaped data [75], then we have

$$orphan(C_j^*) = \Delta(\pi, map(C_j^*)) \tag{16}$$

where $\Delta(\pi, map(C_j^*)) = 1$ if no cluster in π is matched C_j^*, and $\Delta(\pi, map(C_j^*)) = 0$ otherwise, and

$$CI_1(\pi, \pi^*) = \sum_j orphan(C_j^*) \tag{17}$$

$$CI(\pi, \pi^*) = \max\{CI_1(\pi, \pi^*), CI_1(\pi^*, \pi)\} \tag{18}$$

Intuitively, $CI \in [0, K)$. Normally, smaller $CI(\pi, \pi^*)$ values indicate better quality of the selected partition.

C. COMPARE THE QUALITY OF OPTIMAL PARTITION

Since Ncut [7] works on the similarity matrix W of X and can analyse datasets with complex structures, we use Ncut with varied W to produce different partitions Π. The form of element w in W is defined by exponential function as:

$$w_{ij} = \mathrm{e}^{-|filter_{ij}|} \tag{19}$$

where $filter_{ij} = [d(x_i, x_j)]^{order}/(s \cdot \max_{i,j}\{d(x_i, x_j)\})$, $order$ and s are constant numbers. We set s to $[0.0005 : 1]$ with step 0.0005 and $order$ to $[1 : 20]$. Then we have a set of partitions $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$, where $M = 40,000$.

To determine the optimal partition, we fix the number of clusters at the correct K. The M partitions[6] of *non-spherical clusters* and *classification* datasets are tested and the results are shown in Tables 4 and 5.

In Table 4, from the second column, one can see that 16 ground truth partitions in cluster-level (CI=0) are discovered by Ncut with M groups of parameters, except Glass, Movement, Vertebral and Yeast. The performance of CVDD is shown in the third column, from which one can see that 19 partitions out of 20 are the best compared with the other CVIs, and CVDD ranks the first out of the methods tested. Dunn ranks the second, as it can find 14 best partitions. The third-ranking CVI is $\mathcal{I}$, since using it the 12 best partitions

[6] In the experiments, we discard some extreme partitions from M partitions with Ncut's mode isolation (Breiman's bias [76]), namely, partitions possess clusters that are composed of few objects.

TABLE 5. The results measured by Purity. In the second column, the best partitions are determined by purity from 40,000 partitions generated by Ncut with different parameters. The remaining 9 columns are the values of Purity of the best partitions π_{OP} selected by corresponding CVIs. The numbers in bold font means the best in all of the CVIs for a dataset.

Data sets	Best partitions	CVIs								
		CVDD	CVNN	WB	Silhouette	CH	DB	Dunn	S_Dbw	$\mathcal{I}$
non-spherical clusters										
pathbased	0.980	**0.977**	0.627	0.737	0.733	0.737	0.733	0.760	0.627	0.750
spiral	1.000	**1.000**	0.404	0.343	0.349	0.343	0.353	**1.000**	0.346	0.359
aggregation	0.996	**0.996**	**0.996**	0.895	0.931	0.895	**0.996**	0.898	**0.996**	0.977
atom	1.000	**1.000**	0.601	0.715	0.681	0.715	0.553	**1.000**	0.553	0.586
lsun	1.000	**1.000**	0.940	0.758	0.765	0.758	0.948	**1.000**	0.750	0.923
zelnik1	1.000	**1.000**	0.669	0.482	0.712	0.482	0.656	**1.000**	0.482	0.716
zelnik5	1.000	**1.000**	0.609	0.719	0.707	0.719	0.381	**1.000**	0.576	0.582
zelnik6	0.929	**0.853**	0.765	0.828	0.819	0.828	0.798	0.849	0.798	0.815
triangle1	1.000	**1.000**	**1.000**	0.996	0.996	0.996	0.996	**1.000**	0.996	0.998
longsquare	0.999	**0.999**	0.868	0.868	0.836	0.868	0.868	0.994	0.868	0.871
classification										
Iris	0.973	**0.967**	0.667	0.893	0.893	0.893	0.840	0.907	0.840	0.900
Ionosphere	0.920	**0.843**	0.641	0.715	0.667	0.715	0.641	0.664	0.641	0.715
Wine	0.719	**0.719**	**0.719**	0.708	0.685	0.708	**0.719**	0.691	0.680	0.691
Glass	0.593	0.533	0.547	0.551	0.537	0.551	0.547	0.514	0.547	**0.570**
WDBC	0.912	**0.866**	0.627	0.854	0.805	0.854	0.654	0.661	0.859	0.654
Movement	0.567	**0.514**	0.483	0.492	0.458	0.492	0.481	0.492	0.469	0.464
Vertebral	0.777	**0.748**	0.535	0.726	0.713	0.726	0.706	0.719	0.535	0.706
Yeast	0.565	**0.555**	0.510	0.491	0.379	0.491	0.437	0.436	0.510	0.542
Leukemia	0.819	**0.806**	0.708	0.708	0.708	0.708	0.528	0.528	0.708	0.528
Seeds	0.914	**0.905**	0.890	0.890	0.890	0.890	0.890	0.876	0.890	0.862
Number of the best partitions		**19**	3	0	0	0	2	6	1	2

are found. In addition, the CI of CVDD's π_{OP} on Glass is not much worse than the best $CI = 2$.

In Table 5, from the second column, one can see that six ground truth partitions (the same to the GT's labels) in point-level ($Purity = 1$) are discovered by Ncut with M groups of parameters, and all partitions with high quality are discovered in *non-spherical clusters*. The performance of CVDD is shown in the third column, from which one can see that 19 partitions out of 20 are the best compared with the other CVIs, and the method ranks first among the methods tested. Dunn ranks the second, as it did not find one best partition in *classification* but could find the six best partitions that are well-separated in distance. Therefore, no CVIs except CVDD can cope with outliers and density-separated in complex structures. For example, in pathbased datasts, only CVDD could cluster the three structures shown in Fig. 10. In addition, the Purity of CVDD's π_{OP} on Glass is not much worse than the best $Purity = 0.57$.

To sum up, CVDD has attractive performance when it is used to determine the optimal partition from complex candidates generated with fixed K.

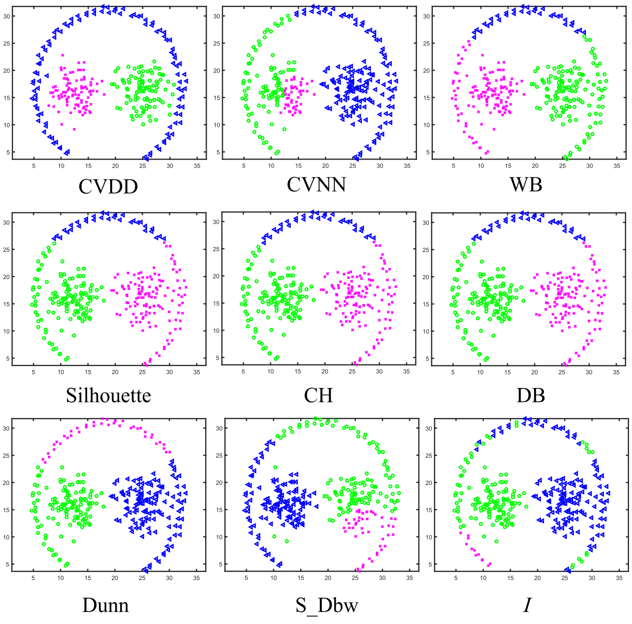


FIGURE 10. CVIs' π_{OP} on pathbased datasets.

D. DETERMINING THE OPTIMAL CLUSTER NUMBER

The other function of an internal validity index is to determine the optimal K. This is a tough task in clustering community, as we usually do not have any a priori knowledge about the dataset and several factors may affect the determination of the number, such as internal validity index, clustering algorithm and parameter estimations. There is no guarantee that the clustering algorithms in optimal K provide

TABLE 6. The optimal K of π_{OP} determined from partitions generated by RS, where partitions with correct K are replaced by the ground truths.

Data sets	Correct K	K detected								
		CVDD	CVNN	WB	Silhouette	CH	DB	Dunn	S_Dbw	$\mathcal{I}$
non-spherical clusters										
pathbased	3	**3**	2	17	2	17	17	16	17	2
spiral	3	**3**	2	17	13	17	13	**3**	17	4
aggregation	7	**7**	2	28	3	27	**7**	26	28	4
atom	2	**2**	**2**	28	12	28	21	**2**	28	3
lsun	3	**3**	2	12	5	6	4	**3**	20	4
zelnik1	3	**3**	2	17	17	17	17	**3**	17	14
zelnik5	4	**4**	2	22	9	22	10	2	22	20
zelnik6	3	**3**	2	6	4	6	2	6	6	2
triangle1	4	**4**	**4**	7	**4**	6	**4**	**4**	19	**4**
longsquare	6	**6**	2	30	2	30	5	**6**	30	7
classification										
Iris	3	**3**	2	6	2	4	2	4	7	**3**
Ionosphere	2	**2**	**2**	3	4	3	7	5	7	3
Wine	3	2	2	5	2	5	2	5	5	5
Glass	6	2	2	3	3	2	3	3	3	3
WDBC	2	**2**	**2**	7	3	7	3	3	7	5
Movement	15	2	2	6	16	2	14	16	16	2
Vertebral	3	2	2	2	2	2	2	2	**3**	2
Yeast	10	2	2	6	5	2	5	6	13	5
Leukemia	3	2	2	2	2	2	2	2	**3**	2
Seeds	3	2	2	9	2	2	2	6	9	**3**
Number of optimal Ks detected		**13**	4	0	1	0	2	6	2	3

meaningful results unless we follow the Algorithm Correctness Assumption [77].[7]

In this section, we evaluate the performance of CVDD on determining the optimal K. Normally, one can generate partitions with different K, and measure the quality of these partitions with an internal validity index. The best partition is then selected, and the K of this partition can be viewed as the optimal K. To generate different partitions only by changing K, we do not employ Ncut. Because the inputs W and K of Ncut, both can change a partition while we focus on the affection of different K on partitions. Meanwhile, it is not applicable to tune the correct K in Ncut if we have no idea of selected W. Furthermore, clustering algorithms optimizing sum-of-squared criterion are usually used to generate natural clusters. Therefore, we employ Random Swap (RS) [78] which is shown to find better clustering results than K-means. The different numbers of clusters are varied from 2 to $\lfloor \sqrt{N} \rfloor$, according to the commonly used rule [79].

The RS[8] consists of K-means while the quality of the result depends on the initialization [47], so we run K-means 20 times in each iteration of RS and set the iteration to 1000 for converge. From the varied partitions with different Ks, one can select π_{OP} and take the K of π_{OP} as optimal K.

For *non-spherical clusters* and *classification* datasets, RS does not perform well in correct K. Thus, to follow the Algorithm Correctness Assumption, we put the ground truths of the twenty datasets into the candidate partitions, and replace those partitions with correct K. The experimental results are shown in Table 6. From the table, one can observe that CVIs find the correct K on 16 datasets except Wine, Glass, Movement and Yeast while CVDD detects the correct K on 13 datasets out of 16 except Vertebral, Leukemia and Seeds. In addition, CVDD finds all correct K in *non-spherical clusters* while the second-best, Dunn, finds six of them and none of the *classification* datasets.

For *spherical clusters* datasets, RS performs very well in correct K [78]. Thus, the selected partitions generalized by RS in correct K makes sense. The experimental results are shown in Table 7. From the table, one can observe that the sum-of-squares based indices [38] CH and WB find all the correct K on 10 datasets. CVDD finds 6 correct K, which is better than the performance of Dunn, CVNN and S_Dbw. It is surprising that CVDD has a good performance on both a1 and a2 datasets [47]. Furthermore, CVDD is degraded on the unbalanced datasets and high overlap datasets (s2, s3, s4 datasets) [47]. In addition, as the overlap of a1, a2 is 20%, the overlap of s1 is 9%, CVDD may be able to cope with less-overlapped datasets but possibly be discard on higher- overlapped datasets (the overlap of s2, s3, s4 is 22%, 41%, 44%.)

From the above two situations, one can see CVDD outperforms the compared CVIs excepting overlap and unbalance towards determination of the number of clusters.

[7] The main idea is that the clustering algorithm works well on a set of Ks, and the partition with correct K is the one that best fits the dataset. In other words, only if the optimal partition is in the candidates, does the evaluation makes sense.

[8] The source code of Random Swap (RS) algorithm is available at http://www.uef.fi/web/machine-learning/software

TABLE 7. The optimal K of π_{OP} determined from partitions generated by RS.

Data sets	Correct K	K detected								
		CVDD	CVNN	WB	Silhouette	CH	DB	Dunn	S_Dbw	$\mathcal{I}$
spherical clusters										
s1	15	**15**	2	**15**	**15**	**15**	**15**	**15**	70	**15**
s2	15	2	2	**15**	**15**	**15**	**15**	7	70	**15**
s3	15	2	2	**15**	**15**	**15**	**15**	70	70	5
s4	15	2	2	**15**	**15**	**15**	**15**	57	68	2
dim32	16	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**
dim64	16	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**
dim128	16	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**
a1	20	**20**	2	**20**	**20**	**20**	**20**	**20**	54	3
a2	35	**35**	2	**35**	**35**	**35**	**35**	51	72	2
unbalance	8	2	4	**8**	2	**8**	4	2	42	**8**
Number of optimal Ks detected		6	3	**10**	9	**10**	9	5	3	6

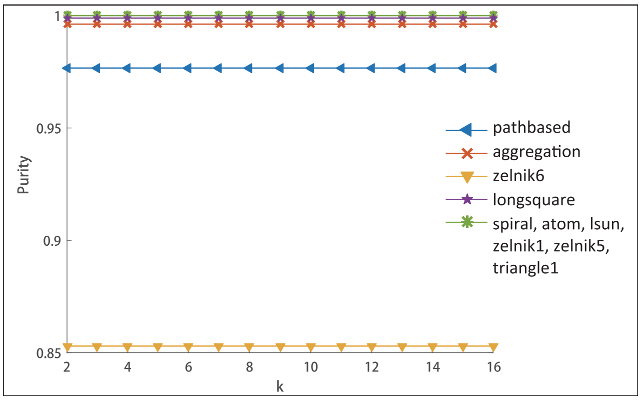


FIGURE 11. CVDD selects optimal partitions with different k on non-spherical datasets. Horizontal axis represents the k value and vertical axis represents the corresponding *Purity* of selected partitions.

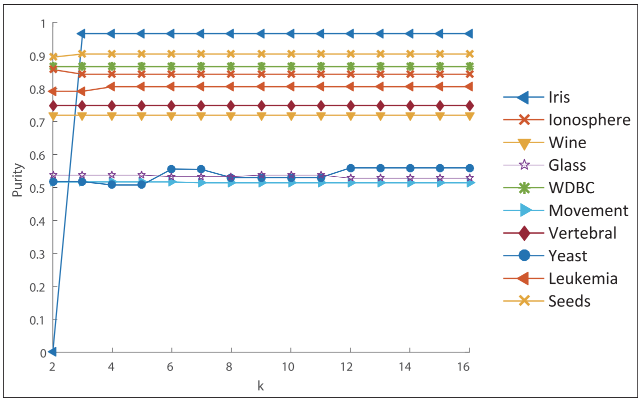


FIGURE 12. CVDD selects optimal partitions with different k on classification datasets. Horizontal axis represents the k value and vertical axis represents the corresponding *Purity* of selected partitions.

E. DISCUSSION OF PARAMETERS IN DENSITY ESTIMATION

In proposed CVDD index, one parameter is used: the number of nearest neighbors k. This is similar to the CVNN index. As this parameter may affect the performance of CVDD, we experimentally discuss it as follows.

We test CVDD with different ks on the twenty *non-spherical clusters* and *classification* datasets. Thus, k is set from 2 to 16, and the performances of selecting the optimal partition on synthetic and real datasets are shown in Figs. 11 and 12, respectively. One can see the performances of CVDD is not sensitive to this parameter except in the datasets Iris and Yeast, where Iris has a large difference between $k = 2$ and the others and Yeast has a small difference between $k = 5$ and $k = 6$ and the others. In all the experiments above, we set k to 7.

VI. CONCLUSION

There does not exist a unified clustering validity index that can cope with varied partitions. The classical and some newly developed ones still cannot cope with complex structures, because they mainly ignore the important information of density connectivity. The newly refined density-involved distance is effectively capture this. The experimental results indicate CVDD has a significance performance on arbitrary clusters. As our index is designed for less-overlapped datasets at the beginning, we will generalize it into higher-overlapped datasets in future work. Furthermore, we will find a better tool to enhance the computational efficiency of density estimation and compactness weights.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[2] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1101–1113, Nov. 1993.

[3] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[4] Z. Yu, H.-S. Wong, and H. Wang, "Graph-based consensus clustering for class discovery from gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2888–2896, 2007.

[5] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[6] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96. 1996, pp. 226–231.

[9] Y. Zhu, K. M. Ting, and M. J. Carman, "Density-ratio based clustering for discovering clusters with varying densities," *Pattern Recognit.*, vol. 60, pp. 983–997, Dec. 2016.

[10] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD. Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.

[11] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[12] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks," *Inf. Sci.*, vol. 450, pp. 200–226, Jun. 2018.

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

[14] U. Von Luxburg, R. C. Williamson, and I. Guyon, "Clustering: Science or art?" in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 65–79.

[15] J. M. Kleinberg, "An impossibility theorem for clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 463–470.

[16] J. Hou and A. Zhang, "Enhanced dominant sets clustering by cluster expansion," *IEEE Access*, vol. 6, pp. 8916–8924, 2018.

[17] J. Wang, C. Zhu, Y. Zhou, X. Zhu, Y. Wang, and W. Zhang, "From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases," *IEEE Access*, vol. 6, pp. 1718–1729, 2018.

[18] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *IEEE Access*, vol. 5, pp. 4991–5000, 2017.

[19] S. Ben-David and M. Ackerman, "Measures of clustering quality: A working set of axioms for clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 121–128.

[20] A. Adolfsson, M. Ackerman, and N. C. Brownstein, "To cluster, or not to cluster: An analysis of clusterability methods," *Pattern Recognit.*, vol. 88, pp. 13–26, Apr. 2019.

[21] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, U.K.: Chapman & Hall, 2012.

[22] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 155–169, Jan. 2015.

[23] J. N. Myhre, K. O. Mikalsen, S. Løkse, and R. Jenssen, "Robust clustering using a knn mode seeking ensemble," *Pattern Recognit.*, vol. 76, pp. 491–505, Apr. 2018.

[24] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Trans. Evol. Comput.*, vol. 11, no. 1, pp. 56–76, Feb. 2007.

[25] M. Garza-Fabre, J. Handl, and J. Knowles, "An improved and more scalable evolutionary approach to multiobjective clustering," *IEEE Trans. Evol. Comput.*, vol. 22, no. 4, pp. 515–535, Aug. 2018.

[26] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.

[27] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, Dec. 2001.

[28] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, and J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.

[29] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.-Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[30] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[31] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov./Dec. 2001, pp. 187–194.

[32] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[33] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 2008.

[34] M. Halkidi and M. Vazirgiannis, "A density-based cluster validity approach using multi-representatives," *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 773–786, Apr. 2008.

[35] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[36] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, "Density-based clustering validation," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2014, pp. 839–847.

[37] A. Ben Said, R. Hadjidj, and S. Foufou, "Cluster validity index based on Jeffrey divergence," *Pattern Anal. Appl.*, vol. 20, no. 1, pp. 21–31, 2017.

[38] Q. Zhao and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," *Data Knowl. Eng.*, vol. 92, pp. 77–89, Jul. 2014.

[39] J. C. Rojas-Thomas, M. Santos, and M. Mora, "New internal index for clustering validation based on graphs," *Expert Syst. Appl.*, vol. 86, pp. 334–349, Nov. 2017.

[40] S. Zhou, Z. Xu, and F. Liu, "Method for determining the optimal number of clusters based on agglomerative hierarchical clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 3007–3017, Dec. 2017.

[41] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang, "A novel cluster validity index based on local cores," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 985–999, Apr. 2018.

[42] S. Saha and S. Bandyopadhyay, "Application of a new symmetry-based cluster validity index for satellite image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 166–170, Apr. 2008.

[43] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based validation of clustering solutions," *Neural Comput.*, vol. 16, no. 6, pp. 1299–1323, 2004.

[44] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 28, no. 3, pp. 301–315, Jun. 1998.

[45] S. Zhou and Z. Xu, "A novel internal validity index based on the cluster centre and the nearest neighbour cluster," *Appl. Soft Comput.*, vol. 71, pp. 78–88, Jun. 2018.

[46] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Trans. Comput.*, vol. C-20, no. 1, pp. 68–86, Jan. 1971.

[47] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Appl. Intell.*, vol. 48, no. 12, pp. 4743–4759, 2018. [Online]. Available: http://cs.uef.fi/sipu/datasets/

[48] A. Hinneburg and D. A. Keim, "A general approach to clustering in large databases with noise," *Knowl. Inf. Syst.*, vol. 5, no. 4, pp. 387–415, 2003.

[49] E. M. Knox and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in *Proc. Int. Conf. Very Large Data Bases*, 1998, pp. 392–403.

[50] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[51] P. Fränti and S. Sieranoja, "Dimensionally distributed density estimation," in *Proc. Int. Conf. Artif. Intell. Soft Comput.* Cham, Switzerland: Springer, 2018, pp. 343–353.

[52] V. Hautamaki, I. Karkkainen, and P. Fränti, "Outlier detection using k-nearest neighbour graph," in *Proc. IEEE 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2004, pp. 430–433.

[53] I. Borg, P. J. F. Groenen, and P. Mair, *Applied Multidimensional Scaling and Unfolding*. New York, NY, USA: Springer, 2018.

[54] B. Fischer and J. M. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 4, pp. 513–518, Apr. 2003.

[55] C. Zhong, D. Miao, and R. Wang, "A graph-theoretical clustering method based on two rounds of minimum spanning trees," *Pattern Recognit.*, vol. 43, no. 3, pp. 752–766, 2010.

[56] K.-H. Kim and S. Choi, "Neighbor search with global geometry: A minimax message passing algorithm," in *Proc. ACM 24th Int. Conf. Mach. Learn.*, 2007, pp. 401–408.

[57] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster kernels for semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 601–608.

[58] M. Seeger, "Learning with labeled and unlabeled data," Inst. Adapt. Neural Comput., Univ. Edinburg, Edinburgh, Scotland, Tech. Rep. TR.2001, 2001.

[59] A. Ultsch, "Clustering with SOM: U*C," in *Proc. Workshop Self-Organizing Maps*, 2005, pp. 75–82.

[60] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[61] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[62] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Trans. Knowl. Discovery From Data*, vol. 1, no. 1, p. 4, Mar. 2007.

[63] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1601–1608.

[64] J. Handl and J. Knowles, "Multi-objective clustering and cluster validation," in *Multi-Objective Machine Learning*. Berlin, Germany: Springer, 2006, pp. 21–47.

[65] P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," *Pattern Recognit.*, vol. 39, no. 5, pp. 761–775, 2006.

[66] P. Fränti, O. Virmajoki, and V. Hautamäki, "Fast agglomerative clustering using a k-nearest neighbor graph," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1875–1881, Nov. 2006.

[67] I. Kärkkäinen and P. Fränti, "Dynamic local search algorithm for the clustering problem," Dept. Comput. Sci., Univ. Joensuu, Joensuu, Finland, Tech. Rep. A-2002-6, 2002.

[68] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2173–2186, Aug. 2016.

[69] P. Fränti, M. Rezaei, and Q. Zhao, "Centroid index: Cluster level similarity measure," *Pattern Recognit.*, vol. 47, no. 9, pp. 3034–3045, 2014.

[70] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.

[71] N. Nguyen and R. Caruana, "Consensus clusterings," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 607–612.

[72] S. Huang, Z. Kang, I. W. Tsang, and Z. Xu, "Auto-weighted multi-view clustering via kernelized graph learning," *Pattern Recognit.*, vol. 88, pp. 174–184, Apr. 2019.

[73] K. Zhan, F. Nie, J. Wang, and Y. Yang, "Multiview consensus graph clustering," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1261–1270, Mar. 2019.

[74] S. Mukherjee *et al.* (2018). "ClusterGAN: Latent space clustering in generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1809.03627

[75] P. Fränti and M. Rezaei, "Generalizing centroid index to different clustering models," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*. Cham, Switzerland: Springer, 2016, pp. 285–296.

[76] D. Marin, M. Tang, I. Ben Ayed, and Y. Boykov, "Kernel clustering: Density biases and solutions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 136–147, Jan. 2019.

[77] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, and J. M. Pérez, and J. I. Martín, "Towards a standard methodology to evaluate internal cluster validity indices," *Pattern Recognit. Lett.*, vol. 32, no. 3, pp. 505–515, 2011.

[78] P. Fränti, "Efficiency of random swap clustering," *J. Big Data*, vol. 5, no. 1, p. 13, 2018.

[79] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396–2409, Dec. 2011.

LIANYU HU is currently pursuing the M.S. degree in computer science from Ningbo University. His current research interests include machine learning, data mining, and pattern recognition.

CAIMING ZHONG received the Ph.D. degree from Tongji University and the University of Eastern Finland, in 2013. He is currently a Professor with the College of Science and Technology, Ningbo University. His research interests include machine learning, data mining, and pattern recognition.