IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 12, 2020, accepted October 23, 2020, date of publication November 5, 2020, date of current version November 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036074

A New Cluster Validity Index Based on the Adjustment of Within-Cluster Distance

QI LI, SHIHONG YUE, YARU WANG, MINGLIANG DING, AND JIA LI
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

Corresponding author: Shihong Yue (shyue1999@tju.edu.cn)

This work was supported by the National Science Foundation of China under Grant 61573251 and Grant 61973232.

ABSTRACT The evaluation on clustering results is an important component of clustering analysis, which can be conducted by the cluster validity index. However, the performances of most existing indices depend on not only the specific clustering algorithms but also the measurements of within- and between- cluster distances and data structures, resulting in limited applications in practice. In this paper, a new within-cluster distance under a general assumption is defined first. After adjusting within-cluster distances of each point according to the adjustment rule, a novel cluster validity index is proposed. Moreover, the notion of chain is introduced to eliminate the effects of sizes, densities, and shapes of clusters. This index does not need any prior information about clustering algorithms and is independent of data structures. Two groups of synthetic datasets with various characteristics and real-world datasets are used to validate this proposed validity index. Experimental results demonstrate that the evaluation accuracy of this index is higher than that of the existing typical indices and performs well on datasets with irregular-shaped clusters.

INDEX TERMS Cluster validity index, within-cluster distance (WD), between-cluster distance (BD).

I. INTRODUCTION

Clustering analysis [1]–[3] is one of the most used machine learning algorithms, which can reveal the hidden structures in a dataset and plays an important role in many domains such as image segmentation [4], [5], data analysis [6], and business applications [7]. There are two significant aspects of clustering analysis: clustering algorithm [8] and cluster validity index [9]–[11]. The number of clusters (c) is an essential parameter of a dataset, and most clustering algorithms must be initially provided with this parameter. The incorrect choice of this parameter can lead to very incorrect clustering results. Thus, it is vital to determine the correct number of clusters in any dataset, and this can be implemented by a cluster validity index. Generally, a cluster validity index is a function that takes various c as its variable, and the maximum or minimum values of this index can assess the correct number of clusters. In the past decades, a great number of cluster validity indices have been designed, such as Calinski-Harabasz (CH) [12] index, Davies-Bouldin (DB) measure [13], Xie-Beni's (XB) separation measure [14], Tibshirani Gap statistics (GS) [15], and Pakhira and Bandyopadhyay' (PB) index [16]. The above validity indices have their applicable ranges and have been widely used in practice.

With the fast development of information techniques in recent years, some challenging problems have been encountered in the process of designing an effective validity index. Firstly, most validity indices are designed by various similarity norms of within-cluster distance (WD) and between-cluster distance (BD), but different norms for measuring WDs and BDs may lead to very inconsistent evaluation results. This greatly decreases their robustness in practice. Recently, Yang *et al.* [17] proposed a novel validity index by optimizing the morphology similarity distance, which can enhance the consistency between WDs and BDs. But the parameters existed in this index play an important role in the evaluation process, and incorrect selection of parameter values may result in incorrect evaluation results. Besides, these parameters increase the processing burden and lead to low efficiency. Recently, Yue *et al.* [18] introduced the notion of dual center to evaluate clustering results. This new measure is effective in some cases, but it is based on the specified clustering algorithms to compute WDs and BDs. If the clustering algorithm chosen for the dataset is not suitable, the evaluation result is not guaranteed.

On the other hand, widespread big data all over the world have caused a great number of new clustering evaluation problems, such as density-different and shape-irregular clusters. Most validity indices only apply to datasets containing regular clusters (such as, spherical clusters). In reference

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

[19], Wang *et al.* proposed an effective cluster validity index of minimizing the seri-overlap(c) (MSO) by estimating the separation of WDs and BDs. This index is independent of clustering algorithms and initialization. Nevertheless, this index only assesses datasets containing spherical clusters. If the tested dataset contains arbitrary-shaped clusters, MSO may fail to suggest the correct cluster number. Besides, MSO assumes that the numbers of points of all clusters approximately satisfy an arithmetic series with a common difference. If the difference is very large or the total number of points is very small, the accuracy of MSO may not be guaranteed. In addition, Lee *et al.* [20] measured the compactness of clusters in the kernel space and proposed a new validity index based on support vector data description (SVDD). This index is suitable for datasets with irregular-shaped clusters. However, SVDD index cannot evaluate datasets with complex data structure and the parameter in the kernel function is hard to determine. Zhou and Xu [21] used the notions of cluster center and the nearest neighbor cluster to design a new internal cluster validity index, which can reflect the geometric distribution of objects. This index can suggest an optimal number of clusters for datasets with various features, but its evaluation results rely on the performance of the clustering algorithm used in the evaluation process. Wani and Riyaz [22] proposed a novel validity index by introducing a new compactness measure based on standard deviation and a penalty function for measuring separation among clusters. This index can evaluate datasets with complex structures, but there lacks a general criterion for selecting its parameter for any dataset.

In this paper, efforts have been made to solve the above problems. A novel validity index is proposed to generalize the measurement of WDs and BDs in a tested dataset. Owing to an adjustment rule and chain-based strategy, the new index can effectively evaluate datasets with density-different and shape-irregular clusters.

II. RELATED WORK

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset containing n points in a d-dimensional space, and $x_i \in R^d$. $S_1, S_2, \ldots, S_c$ are c disjoint subsets of X. Let U be a binary membership function to describe the relationship between points and subsets, satisfying,

$$u_{ij} = \begin{cases} 1, & x_j \in S_i \\ 0, & x_j \notin S_i, \end{cases} \quad i = 1, 2, \ldots, c;\ j = 1, 2, \ldots, n \tag{1}$$

Eq. (1) indicates that if point x_j belongs to ith subset S_i, $u_{ij} = 1$; otherwise, $u_{ij} = 0$. If each point only belongs to one subset, the partition of X is called a hard partition [23], satisfying,

$$X = S_1 \cup S_2 \cup \ldots \cup S_c,\ S_i \cap S_j = \phi, \quad i \neq j;\ i, j = 1,\ 2,\ \ldots,\ c. \tag{2}$$

On the contrary, a fuzzy partition [24] of X means that each point belongs to all subsets with its individual membership degrees, satisfying,

$$u_{ij} \in [0, 1],\ s.t.,\ \sum\nolimits_{i=1}^{c} u_{ij} = 1, \quad i = 1, 2, \ldots, c;\ j = 1, 2, \ldots, n. \tag{3}$$

In general, a validity index is a function $f(c)$ that takes c as its variable, and finds the correct number of clusters by optimizing the following objective function,

$$\min\ z = f(c)\ \ or\ \ \max\ z = f(c). \tag{4}$$

Most validity indices take the trial-and-error way to solve the optimum of Eq. (4) [25], [26]. Firstly, the possible range of c can be set among $[c_{\min}, c_{\max}]$. Generally, $c_{\min} = 2$ and $c_{\max} < \sqrt{n}$ [27] where n is the number of points in X. Then, the dataset X is partitioned into c clusters by using a clustering algorithm. Finally, computing the corresponding value of Eq. (4) at each value of c, and the maximum or minimum of Eq. (4) indicates the correct number of clusters.

Five typical cluster validity indices (PB, DB, CH, GS, and XB) are illustrated as follows. Besides, each index is accompanied by an upward (↑) or downward (↓) arrow. The upward arrow represents that the maximum of this index refers to the optimal partition, and the corresponding c denotes the optimal number of clusters. Inversely, the downward arrow represents the opposite meaning.

1) PAKHIRA AND BANDYOPADHYAY'S INDEX (PB ↑) [16]

The PB index is used for evaluating clustering results from both hard and fuzzy clustering algorithms. For a dataset X containing n points, PB can be defined as

$$\mathrm{PB}(c) = \left(\frac{1}{c} \times \frac{E_1}{J} \times \sum\nolimits_{i,j=1}^{c} |z_i - z_j|\right)^2, \quad s.t., \begin{cases} E_1 = \sum\nolimits_{j=1}^{n} |x_j - z| \\ J = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j=1}^{n} |x_j - z_i| \end{cases} \tag{5}$$

hereafter z_i and z denote the centroid of cluster i and the global centroid of X, respectively.

2) DAVIES-BOULDIN INDEX (DB ↓) [13]

Let Δ_i be the compactness of cluster i; δ_{ij} denotes the separation between clusters i and j. The DB index can be formulated as

$$\mathrm{DB}(c) = \sum\nolimits_{i=1}^{c} R_i / c, \quad s.t., \begin{cases} R_i = \max_{j, j \neq i} (\Delta_i + \Delta_j) / |z_i - z_j| \\ \Delta_i = \sum\nolimits_{x \in S_i} |x_i - z_i| / |S_i| \end{cases} \tag{6}$$

where $|S_i|$ is the number of points in cluster i.

3) CALINSKI-HARABASZ (CH ↑) INDEX [12]

In the CH index, the compactness of ith cluster is computed by the distances between each point and z_i, $i = 1, 2, \ldots, c$, and the separation among clusters is measured in terms of the distances from all centroids to the global centroid z.

$$\mathrm{CH}(c) = \frac{n-c}{c-1} \cdot \frac{\sum_{i=1}^{c} n_i |z_i - z|^2}{\sum_{i=1}^{c} \sum_{k=1}^{n_i} |x_k - z_i|^2} \tag{7}$$

where n_i denotes the number of points of cluster i.

4) TIBSHIRANI'S GAP STATISTIC (GS ↑) INDEX [15]

The GS index can be expressed as,

$$\begin{aligned} &\mathrm{Gap}(c) = E * \left[\log(W(c))\right] - \log(W(c)),\\ &s.t., \begin{cases} W(c) = \sum_{i=1}^{c} D_i/(2|S_i|) \\ D_i = 2|S_i| \sum_{j \in S_i} |x_j - \sum_{i=1}^{|S_i|} x_i/|S_i|| \end{cases} \end{aligned} \tag{8}$$

where $E*$ denotes the expectation under a null reference distribution.

5) XIE–BENI'S SEPARATION INDEX (XB ↓) [14]

The XB index is proposed for fuzzy clustering algorithms, which is the ratio of compactness to separation of a dataset.

$$\mathrm{XB}(c) = (\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} |x_j - z_i|^2)/(n \cdot \min_{i \neq j} |z_j - z_i|^2) \tag{9}$$

where m refers to the fuzzy exponential.

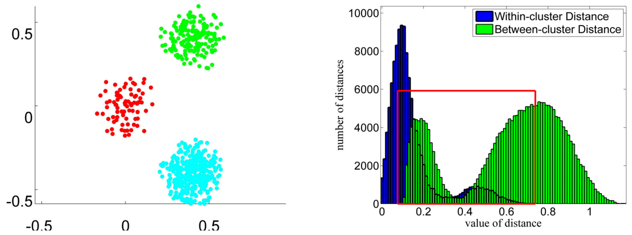
6) MSO INDEX (MSO ↓) [19]

The MSO index can find the real number of clusters in a tested dataset without any prior information (such as clustering algorithm and initialization process). This index assumes the numbers of points of all clusters in X, $|S_1|, |S_2|, \ldots, |S_c|$, approximately satisfy an arithmetic series with common difference d. Denote $d_{\mathrm{within}}(c)$ be the number of WDs, satisfying,

$$\begin{aligned} d_{\mathrm{within}}(c) = {} & (n^2/2c - n/2) \\ & + [(c-1)c(2c-1)/12 - (c-1)^2 c/8] d^2 \end{aligned} \tag{10}$$

Denote the distance from any point to its k-nearest neighbors be k-NN distances, where k is taken as $(1/c)(n\text{-}1)$. Thus, the k-NN distances of any point in X are defined as WDs, whereas the other $((n\text{-}1)\text{-}k)$ distances of the point are denoted as BDs. The distribution of all distances in X can be represented by a statistical histogram.

Fig. 1 shows a dataset and the distribution of distances in the form of statistical histogram under $c = 3$. Firstly, we compute the minimum and maximum of all distances: 0.0004 and 1.18. And then, the interval [0.0004, 1.18] is equally divided into 100 subintervals. Finally, all WDs and BDs are assigned into the above 100 subintervals according to their values, respectively. Each subinterval composes a bar. All bars compose the statistical histogram (see Fig. 1 (b)). The y-axis of Fig. 1 (b) denotes the number of distances in each bar. These bars composed of WDs and BDs are in blue and green, respectively. Bars that contain both WDs and BDs are regarded as set-overlap ones, framed up by a red line.



(a) Original dataset.

(b)Statistical histogram of distance distribution in (a).

FIGURE 1. Dataset and its statistical histogram.

Denote | within (c, q)| and | between (c, q)| as the numbers of WDs and BDs in the qth bar in set-overlap(c), respectively, $q = 1, 2, \ldots, Q$, where seri-overlap(c) is the area in which WDs and BDs overlap seriously, satisfying,

$$1/3 < |\mathrm{withinin}(c, q)|/(|\mathrm{within}(c, q)| + |\mathrm{between}(c, q)|) < 2/3 \tag{11}$$

Let | seri-within(c, p)| and | seri-between(c, p)| be the numbers of WDs and BDs in the pth bar in seri-overlap(c), respectively, $p = 1,2,\ldots P$. MSO index is formulated as

$$\mathrm{MSO}(c) = \frac{\sum_{p=1}^{P} \min\{|\mathrm{seri}-\mathrm{within}(c, p)|,\ |\mathrm{seri}-\mathrm{between}(c, p)|\}}{d_{\mathrm{within}}(c)} \tag{12}$$

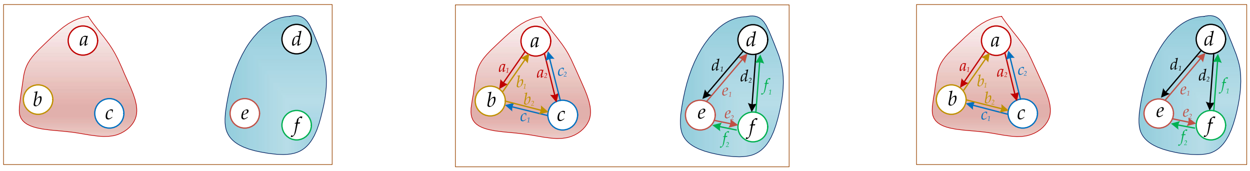
In sum, each index above is based on the clustering result from a specific algorithm (such as C-means and Fuzzy C-means). If the clustering algorithm selected for the dataset is not suitable, the evaluation results will not be guaranteed. Moreover, the above indices cannot give a criterion to choose the correct similarity norm. Therefore, a general and efficient method is necessary. Various combinations of WDs and BDs have constructed most existing validity indices [28], resulting in different evaluation results. This is very undesired in practice. In this paper, our proposed validity index is independent of clustering algorithms and data distributions, illustrating a novel solution to overcome the above problems.

III. NOVEL VALIDITY INDEX

In this section, the initial WDs of each point is defined. And then the adjustment operation is proposed to correct the deviation caused by various sizes of clusters. Finally, the notion of chain is introduced to eliminate the effects of densities and shapes of clusters.

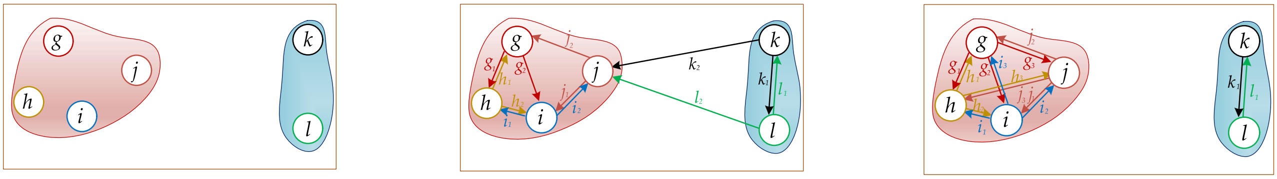
A. THE FUNDAMENTAL OF NEW VALIDITY INDEX

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset consisting of n points belonging to c clusters in a d-dimensional space, and $x_k \in R^d$. For any point $x_k \in X$, its m nearest neighbors are denoted as $x_{k,1}$, $x_{k,2}, \ldots, x_{k,m}$, with distances $dis(x_k, x_{k,1})$, $dis(x_k, x_{k,2}), \ldots, dis(x_k, x_{k,m})$, $m = 1, 2, \ldots, n-1$. Formally, all



(a) Original dataset. (b) Distribution of actual WDs. (c) Distribution of initial WDs.

FIGURE 2. WDs in datasets with evenly distributed clusters. Note: the arrow beginning form each point denotes WD of this point.



(a) Original dataset. (b) Distribution of initial WDs. (c) Distribution of actual WDs.

FIGURE 3. WDs in datasets with unevenly distributed clusters. Note: the arrow beginning form each point denotes WD of this point.

distances in X can be partitioned into two groups: WDs and BDs, as illustrated in Definition 1.

Definition 1: WD and BD. If two points in X are assigned to the same cluster by any clustering algorithm, their distance is called WD; otherwise is called BD.

Considering an extreme case that all clusters in X have the same number of points $[n/c]$, where and hereafter $[\bullet]$ stands for a rounding operator, taking the integer part of the number in a bracket. For any point $x_k \in X$, the sum of all WDs from x_k can be formulated as

$$\mathrm{WCD}(x_k) = \sum\nolimits_{x_m \in KNN(x_k)} dis(x_k, x_m)\ ,\ k = 1, 2, \ldots, n \tag{13}$$

where $KNN(x_k)$ is the set of $[(n/c)\text{-}1]$ nearest neighbors of x_k. Then, the sum of all WDs in X can be computed as

$$\mathrm{WCD}(X) = \sum\nolimits_{k=1}^{n} \sum\nolimits_{x_m \in KNN(x_k)}^{dis(x_k, x_m)} \tag{14}$$

In reference [19], it has been proven that in the statistical sense WD of X satisfies the following properties:

Property 1: Assume that all points in X are partitioned to c clusters. The number of WDs attains the low bound $c \cdot C^2_{[n/c]}$ if any cluster evenly contains $[n/c]$ points.

The numbers of points in various clusters in X are usually different, and thus according to *Property* 1 the total number of WDs is larger than $c \cdot C^2_{[n/c]}$. Only if the initial WDs computed by Eq. (13) are all real WDs, Eq. (14) reaches the minimum. Inversely, if the initial WDs computed by Eq. (13) contain BDs, the value of Eq. (14) will increase.

Property 2: Assume that X contains c clusters, the maximal number of WDs is $C^2_{n-(c-1)}$.

Property 2 refers to an extreme case that there are (c-1) points in which each individually constructs a cluster and other {n-(c-1)} points forms one cluster. Hence, the number of WDs ranges in $[c \cdot C^2_{[n/c]}, C^2_{n-(c-1)}]$. As various cluster sizes, densities, and shapes cause uneven distributions of points among clusters, the number of WDs tends the value of $C^2_{n-(c-1)}$, as explained below.

If the sizes of clusters in a dataset are different, the number of WDs of a point in a large-sized cluster is much larger than that in a small-sized cluster. Hence, the initial $[n/c\text{-}1]$ WDs of a point in a small-sized cluster contain BDs (see Figs. 3 (b)).

Assume points in Figs. 2 and 3 are distributed into 2 clusters, i.e., $c = 2$ in Eq. (13). Fig. 2 (a) shows a dataset that contains 6 points in two clusters, and each cluster evenly has 3 points. Figs. 2 (b) and (c) show the two actual and initial WDs of each point, respectively. Specially, each point has two nearest neighbors along with their relative WDs. Consequently, the sum of total WDs computed by Eq. (14) is minimum since the initial WDs are all the real WDs. Alternatively, the two clusters in Fig. 3 (a) contain different numbers of points, with four and two points, respectively. Fig. 3 (b) shows the distribution of initial WDs of points by using Eq. (13), i.e., two distances of each point to its nearest neighbors when all points evenly have two initial WDs. But the sum of these distances is not minimum since some BDs (l_2, and k_2 in Fig. 3 (b)) are incorrectly regarded as WDs. Thus, these long BDs should be removed, and those short WDs (g_3, h_3, i_3, and j_3) that are incorrectly regarded as initial BDs should be added. In this case, the sum of WDs decreases. And thereby those added distances are more consistent with actual WDs, as shown in Fig. 3 (c).

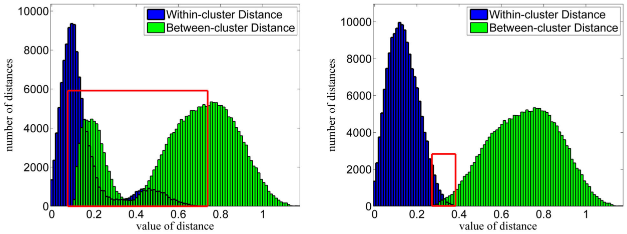
Under the properties above, when clusters have different sizes in a dataset, not all initial WDs of any point refer to actual WDs. To solve this problem, the notion of adjustment is introduced as follows.

Let the points with the largest and smallest values of WCD(x) in X be x_{max} and x_{min}, respectively. The value of WCD(X) can be minimized to attain its minimum by the following adjustment process. Let the final WCD(X) with the minimum value under one certain c be WCD*(c).

When the clusters in any dataset have different numbers of clusters, the adjustment algorithm can help to find the real WDs, and replace BDs with real WDs, minimizing Eq. (13). It can minimize the value of WCD(X) to attain its minimum. The above adjustment rule can correct the deviation caused

Algorithm 1 Adjustment Process

```
Input: A dataset $X \in R^d$ containing $n$ points.
Output: WCD*($c$).
Steps:
  1. Compute WCD($X$) by using Eq. (14) and denote it as WCD$_t$($X$), $t = 1$;
  2. Select points with the largest and smallest values of WCD($x$) be $x_{max}$ and $x_{min}$;
  2. Remove the longest WD of $x_{max}$ from WCD($x_{max}$), and add the closest BD of $x_{min}$ to WCD($x_{min}$);
  3. Compute the current WCD($X$) and denote it as WCD$_{t+1}$($X$);
  4. If WCD$_{t+1}$($X$) < WCD$_t$($X$), WCD$_t$($X$) = WCD$_{t+1}$($X$), $t = t + 1$, and turn to step 2; or else, turn to step 5;
  5. Stop and obtain WCD*($c$) = WCD$_t$($X$).
```



(a) Original statistical histogram.

(b) Statistical histogram after adjustment process.

FIGURE 4. Statistical histograms before and after adjustment.

by the assumption through removing BDs and adding WDs, changing the value of KNN of each point.

Fig. 4 shows the statistical histograms before and after adjustment of dataset in Fig. 1 (a), and c is fixed at 3. Fig. 4 (a) shows that many BDs are regarded as initial WDs, whereas WDs are identified as initial BDs. Fig. 4 (b) shows that more shorter distances are regarded as WDs after the adjustment process, and the longer ones are identified as BDs, illustrating that BDs are effectively replaced with WDs by using the adjustment process. Thus, the sum of WDs can be decreased by the adjustment process.

WCD*(c) is the smallest value under one certain c and can be used to suggest the correct c.

Figs. 5 and 6 show the distributions of WDs before and after the adjustment operation, respectively. Fig. 5 shows initial WDs of a dataset containing 12 points belonging to 3 clusters at c are smaller, equal to, and larger than 3, respectively. When $c = 2$, the number of initial WDs computed by Eq. (14) is [(n/c)-1] $\cdot$ n = [(12/2) $-$ 1] $\times$ 12 = 60, and the distributions of these distances are illustrated in Fig. 5 (a). The numbers of initial WDs are 36 (see Fig. 5 (b)) and 24 (see Fig. 5 (c)) when $c = 3$ and 4, respectively, and the actual number of WDs is $5 \times 4 + 3 \times 2 + 4 \times 3 = 38$. In Fig. 6 (a), the adjustment operation is executed by removing distances of points f, g, j, k, and l, and adding the distances of points a, b, c, d, and e. However, distances after adjustment still contain BDs because the computed number (60) by Eq. (14) is greater than the actual number of WDs (38). When $c = 3$, BDs can be removed and WDs can be added by using the adjustment operation (removing distances of points j, k, and l, and adding distances of points c, d, and e). When $c = 4$, the distances before and after adjustment are all WDs. It can be concluded that when c is smaller than the real number of clusters, the final distances after adjustment contain both WDs and BDs when c is equal to or great than the real cluster number, the final distances contain only WDs.

Reference [19] illustrates that for any dataset, the number of WDs attains the low bound $c \cdot C^2_{[n/c]}$ when all points in X are evenly partitioned into c clusters compared with the situation where points are unevenly partitioned. In this case, the initial number of WDs under one certain number of c is the smallest compared with other distributions.

1) **Case 1**. *Larger number of clusters.* If the given c is equal to or greater than the real number of clusters, the number of initial WDs is no greater than the real number. If points are distributed unevenly, the initial WDs contain both real WDs and BDs (see Fig. 5 (b)). The adjustment process can replace all BDs by WDs. Finally, the final distances contain only WDs (see Fig. 6(b)).

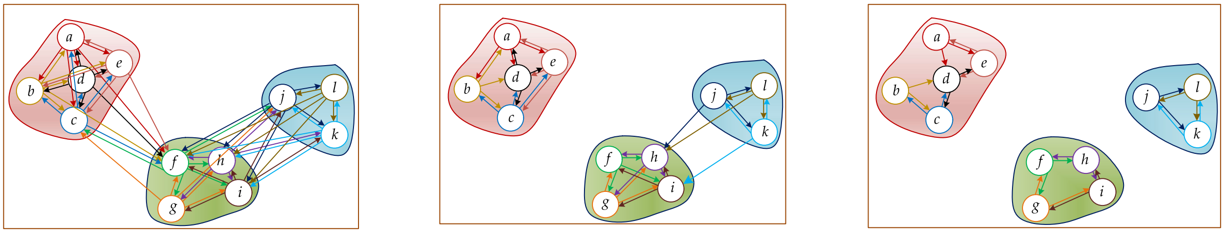
2) **Case 2.** *Smaller number of clusters.* If the given c is smaller than the real number of clusters, the number of initial WDs will larger than the real one. Although some BDs can be replaced by the corresponding WDs, some BDs cannot be replaced due to the rule of keeping the total number of WDs unchanged. Finally, the final distances contain both WDs and BDs (see Fig. 6 (a)).

Figs. 7 (a) and (b) show a dataset containing 400 points belonging to four clusters, and the curve of WCD*(c), respectively. Fig. 7 (b) illustrates that the values of WCD*(c) decrease fast when $c < 4$ but nearly unchanged when $c > 4$. When c is smaller than the actual one 3, the number of initial WDs by Eq. (14) is [((n/c)-1)] $\cdot$ n = [((400/3)-1)] $\times$ 400 = 52933, which is greater than that computed at $c = 4$(39600). Thus, when $c < 4$, distances after adjustment operation still contain BDs. Distances after the adjustment operation only contain WDs when c4. Thus, the values of WCD*(c) at $c <$ 4 are much larger than that at $c \geq 4$. And there is little difference between these values of WCD*(c) at $c \geq 4$. Consequently, there is an elbow point on the curve of WCD*(c), indicating the correct number of clusters.

Considering the variances of WCD*(c) can be calculated by curvature radius mathematically, we define a novel validity index based on the adjustment of WDs as follows.

$$F\ (c) = \ |\Delta(c)|^2/(1 + (\nabla(c))^2)^{3/2},$$
$$s.t., \begin{cases} \Delta(c) = \mathrm{WCD} * (c+1) + \mathrm{WCD} * (c-1) - \mathrm{WCD} * (c) \\ \nabla(c) = \mathrm{WCD} * (c+1) - \mathrm{WCD} * (c-1) \end{cases} \tag{15}$$

where the symboldenotes a second-order derivative operator, aiming to locate the elbow point on the curve of WCD*(c).

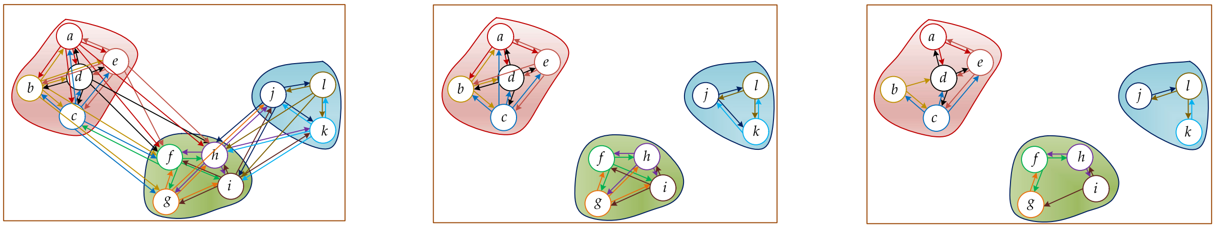


(a) $c = 2$.

(b) $c = 3$.

(c) $c = 4$.

FIGURE 5. **(a)-(c) show the distributions of WDs under the assumption above when $c = 2, 3$, and 4, respectively. Note: the arrow beginning form each point denotes WD of this point.**

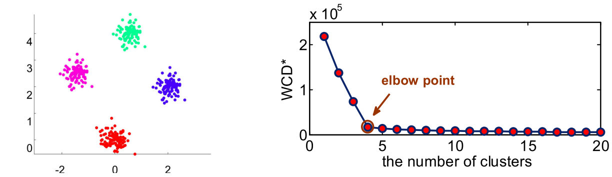


(a) $c = 2$.

(b) $c = 3$.

(c) $c = 4$.

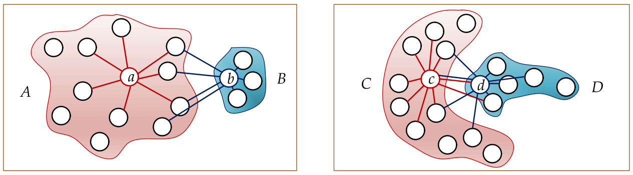
FIGURE 6. **(a)-(c) show the distributions of WDs after adjustment operation when $c = 2, 3$, and 4, respectively. Note: the arrow beginning form each point denotes WD of this point.**



(a) The original dataset.

(b) The curve of WCD*(c).

FIGURE 7. **The curve of WCD*(c).**



(a) Dataset with clusters of different sizes and densities.

(b) Dataset with irregular-shaped clusters.

FIGURE 8. **The distributions of initial WDs computed by Eq. (13) for datasets containing clusters of different sizes and densities, and irregular-shaped clusters. Note: The red and blue lines denote initial WDs of points *a*, *b*, *c*, and *d* by using Eq. (13), respectively.**

The optimal number of clusters $c*$ is computed as,

$$c* = \arg\max_c F(c) \tag{16}$$

B. ELIMINATION OF EFFECTS CAUSED BY SIZES, DENSITIES, AND SHAPES OF CLUSTERS

WCD∗ can easily be affected by the distribution of clusters (such as density, size, and shape). When the tested dataset contains clusters of different densities and sizes, the initial WD of a point in a sparse and large-sized cluster may be much greater than that in a dense and small-sized cluster (see Fig. 8 (a)). In this case, the adjustment process may incorrectly remove the larger WD from the former and incorrectly add the smaller BD to the latter. When the tested dataset contains irregular-shaped clusters (see Fig. 8 (b)), distances computed by Eq. (14) may contain BDs, affecting the accuracy of adjustment operation and leading to incorrect evaluation result.

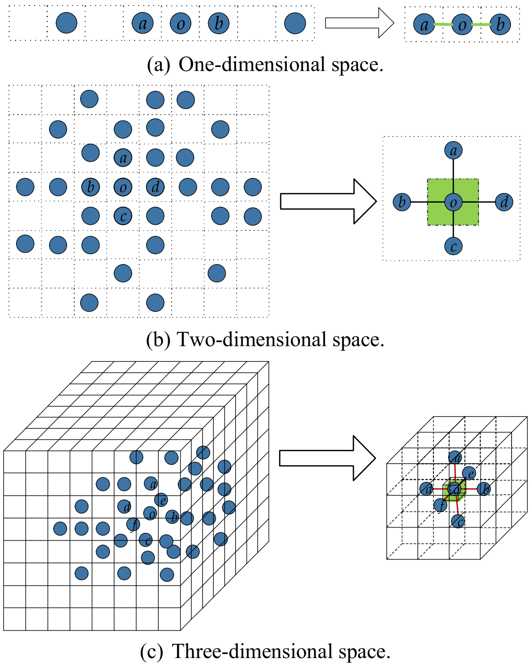
In Fig. 8 (a), cluster A has a relatively larger size containing 12 points compared with cluster B containing 4 points. The number of initial WDs for each point is 7 ($[((12 + 4)/2) - 1]$). Fig. 8 (a) shows that the sum of initial WD of points in cluster A is much larger than that in cluster B, i.e., WCD(a) > WCD(b). According to the adjustment rule above, the longest initial WD of point a should be removed from WCD(a) whereas the closest initial BD of point b should be added to WCD(b). However, the distribution in Fig. 8 (a) indicates that distances of point b contain BDs, which should be removed, and distances of point a are all WDs, which should not be removed. The dataset in Fig. 8 (b) has irregular-shaped clusters, containing 12 and 6 points, respectively. Fig. 8 (b) illustrates that both WCD(c) and WCD(d) contain WDs and BDs, which cannot be corrected by the above adjustment rule and result in incorrect evaluation results.

To solve this problem, the notion of chain is introduced, based on which the clusters of different sizes and densities, and irregular-shaped clusters can be transformed into spherical clusters. And the deviation caused by size and density can be eliminated.

Definition 2 (Density): The density of point x_k can be represented by its m nearest neighbors, satisfying,

$$\rho_k = \{\sum\nolimits_{j=1}^{m} dis(x_k, x_{k,j})\}^{-1}, \quad k = 1, 2, \ldots, n \tag{17}$$

where m is generally fixed at $2d$ (see Fig. 9).



(a) One-dimensional space.

(b) Two-dimensional space.

(c) Three-dimensional space.

FIGURE 9. The relationship between hypervolume and neighbors in different spaces.

A cluster consists of a group of points that occupies the spatial position; thus, it has the corresponding volume or hypervolume measure. Fig. 9 shows the hypervolume occupied by points in the one-/two-/three- dimensional space, respectively. In one-dimensional space, the occupied hypervolume of a point can be measured by its two nearest neighbors (see Fig. 9 (a)). Similarly, in the two-/three- dimensional space, the occupied hypervolume of any point can be computed by its 4/6 nearest neighbors (see Figs. 9 (b) and (c)). Thus, the occupied hypervolume of a point in a d-dimensional space can be measured by its $2d$ nearest neighbors, which considers all directions of this space.

Different from the existing density notions [29], [30], the proposed density does not need any prior information, and thus it is nonparametric, which can reduce the uncertainties in clustering process.

Definition 3 (Nearest Density-Based Neighbor): For any point $x_k \in X$, let the minimum distance from x_k to other points with a higher density than x_k be the nearest density-based distance σ_k, and the corresponding point is denoted as the nearest density-based neighbor φ_k [31].

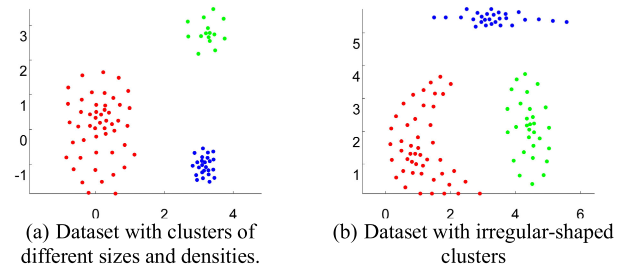
$$\sigma_k = \min_{j:\rho_k<\rho_j} dis(x_k, x_j) \tag{18}$$

$$\varphi_k = \arg \min_{j:\rho_k<\rho_j} dis(x_k, x_j) \tag{19}$$

Definition 4 (Key Points): Points with higher values of KP, defined as Eq. (20), are regarded as key points.

$$KP_k = \rho_k \cdot \sigma_k \tag{20}$$

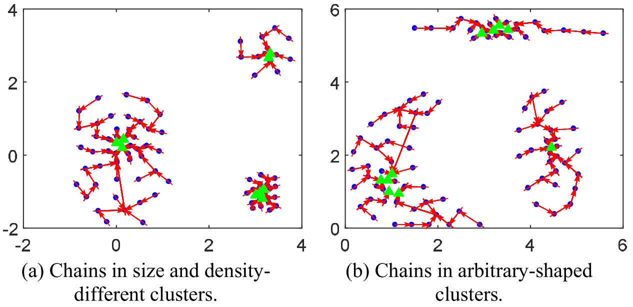
If there is no prior knowledge, c in any dataset is less than $\lfloor\sqrt{n}\rfloor$ [18].



(a) Dataset with clusters of different sizes and densities.

(b) Dataset with irregular-shaped clusters

FIGURE 10. Datasets containing clusters with different characteristics.



(a) Chains in size and density-different clusters.

(b) Chains in arbitrary-shaped clusters.

FIGURE 11. Distributions of chains in the two datasets in Fig. 10. Note: the green triangles and the red lines with arrows denote the key points and the directions of chains, respectively.

Definition 5 (Chain): For any point x_k, the next point x_j is the nearest density-based neighbor of x_k, i.e., φ_k. A chain is a group of points in X, i.e., $x_{i1}, \varphi_{i1}, \ldots, x_{ip}, \varphi_{ip}$, which starts with x_{i1} and stops at a key point φ_{ip}.

Definition 3 illustrates that for any point x_k in X, the nearest density-based neighbor φ_k is unique; and thus, according to Definition 5, the chain starting from this point is unique.

The adjacent points in X can be connected to form a chain following the connecting rule. And the direction of each arrow is from low to high density points in a chain. The above steps are repeated until each point is visited in X.

Let T_i be the length of chain $x_{i1}, \varphi_{i1}, \ldots, x_{ip}, \varphi_{ip}$,satisfying,

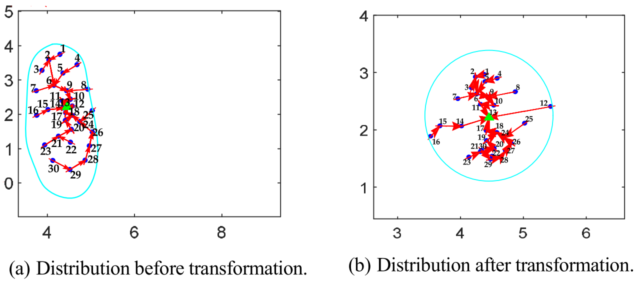
$$T_i = \sum\nolimits_{p=1}^{p} dis(x_{ip}, x_{ip+1}) \tag{21}$$

where $dis(x_{ip}, x_{ip+1})$ is the distance between adjacent points x_{ip} and x_{ip+1} on the ith chain, $i = 1, 2, \ldots, \sqrt{n}$.

Dataset in Fig. 10 (a) contains 90 points belonging to three clusters of different sizes and densities, and dataset in Fig. 10 (b) contains irregular-shaped clusters with 105 points in total. In general, the two datasets can be divided into 9 and 10 chains, i.e., $\sqrt{90}$ and $\sqrt{105}$, respectively (see Fig. 11). Fig. 11 shows that different chains contain different numbers of points due to the sizes, densities, and shapes of clusters, leading to the result that different chains have different values of length.

To normalize the size and density of each cluster and make its shape regular, the mth line segment $dis(x_{im}, x_{im+1})$ on the ith chain is transformed into a new one, satisfying,

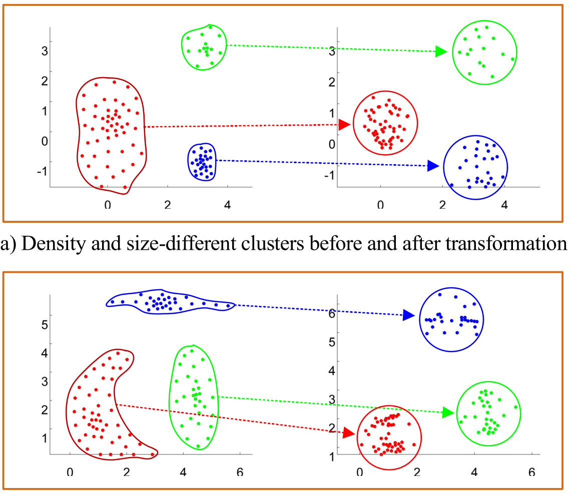
$$dis^*(x_{im}, x_{im+1}) = dis(x_{im}, x_{im+1})/T_i, \quad i=1, 2, \ldots, \sqrt{n} \tag{22}$$



(a) Distribution before transformation.

(b) Distribution after transformation.

FIGURE 12. **Distribution of points before and after transformation.**



(a) Density and size-different clusters before and after transformation.

(b) Arbitrary-shaped clusters before and after transformation.

FIGURE 13. **Datasets before and after transformation. Note: Different colors denote different clusters and the occupied area of each cluster is circled by a corresponding curve. The original and the corresponding transformed cluster are connected by a dotted line.**

Eq. (22) can shorten the lengths of long chains and enlarge those of the short chains. Consequently, centralizing at any key point, the points on a long chain move to the key points, and those on a short chain move far from the key points.

Fig. 12 shows a detailed transformation process, with the cluster in green in Fig. 10 (b) as an example. In Fig. 12 (a), the length of chain (16-15-14-13) is smaller than that of chain (1-2-6-9-10-11-13). Fig. 12 (b) shows that Eq. (22) can shorten the length of the first chain, whereas enlarge that of the second chain, making the lengths of the two chains similar. Fig. 12 illustrates that the shape of the cluster can be normalized after transformation.

Fig. 13 shows the transformation results of datasets in Fig. 10. Fig. 13 illustrates that the clusters after transformation are all spherical clusters. And the sparse and large-sized cluster is transformed into a dense and large-sized cluster, whereas the dense and small-sized cluster is transformed into a sparse and small-sized cluster. And the initial WDs in the transformed dense and large-sized cluster are much smaller than that in a sparse and small-sized cluster, which is consistent with the adjustment operation above.

Hereafter, the proposed validity index based on the adjustment of WDs is called AWCD (Adjustment of within-cluster distance) index. The evaluation process of AWCD index is listed in **Algorithm 2**.

Algorithm 2 Evaluation Process of AWCD Index

```
Input: A dataset $X \in R^d$ containing $n$ points.
Output: $c$.
Steps:
  1. Normalize clusters in $X$ based on chain;
  2. Compute the initial WCD for each point in $X$ according to Eq. (13);
  3. Adjust WDs following the adjustment rule;
  4. Compute the value of WCD*($c$) at $c = 1, 2, \ldots, c_{max}$;
  5. Solve the optimal value of Eq. (15);
  6. Suggest optimal $c*$;
  7. Stop.
```

Compared with the existing indices, AWCD index has the following characteristics:

1) The entire evaluation process of AWCD index does not need any prior information. In contrast, most of the existing cluster validity indices depend on clustering algorithms and can only perform well on spherical clusters.

2) AWCD index can reveal the hidden structure by using the transformation process regardless of the sizes, densities, and shapes of clusters.

For any tested dataset containing n points, the computation complexity of AWCD index mainly consists of two parts: 1) normalizing all distances in any chain, and 2) adjusting the WDs according to the adjustment rule. The runtime of the first part mainly results from the computation of all distances in X, and thus the computational cost of this part is $O(n^2)$; the computational complexity of the second part is $O(tn^2)$, where t is the number of executions of adjustment process. Thus, the computation complexity of AWCD index is $O(tn^2)$.

IV. EXPERIMENTAL RESULTS

To validate AWCD index, experiments are conducted on two groups of typical datasets (i.e., synthetic and real datasets). Four most used hard validity indices (i.e., PB [16], DB [13], CH [12], and GS [15]), one fuzzy validity index XB [14], as well as MSO [19] index are used to make a comparison. Reference [27] has proven that the maximum c in a dataset is $\sqrt{n}$. Thus, considering the time complexity, if there is no prior knowledge, $c_{\min}$ is often taken as 2, and $c_{\max} \leq \sqrt{n}$. For datasets with small number of clusters (Sets 1-16), $c_{\max}$ is set at the maximal $\sqrt{n}$ (i.e., 30). For Sets 18 and 20, which has large number of points distributed in small number of clusters, we simply choose 30 as the maximal cluster number to decrease the time complexity. For datasets with large cluster number and relatively smaller number of points (Sets 17 and 19), we experientially fixed [$c_{\min}$, $c_{\max}$] at [30, 60] and [90, 120]. For the four hard indices, C-means is used as the clustering algorithm; for XB index, fuzzy C-means is used to cluster data points. And the fuzzy exponent in fuzzy C-means is set as 2.0. All experiments are carried out by a computer with MATLAB R2017, 2.80 GHz CPU, 8.00 GB RAM in Windows 10. The runtime is computed by functions of "tic" and "toc" in MATLAB toolbox.

A. TESTS ON SYNTHETIC DATASETS

Fig. 14 shows five groups of synthetic datasets with various characteristics generated by the Matlab®toolbox. Different colors denote different clusters. The original and transformation results of datasets indicate that the transformation rule can transform irregular clusters into spherical clusters and eliminate the effects of size and density. The characteristics of these datasets are described as follows:

- Datasets in Group 1 contain clusters of different densities. Clusters in each dataset are all spherical;
- Datasets in Group 2 contain spherical clusters of different sizes and densities;
- Group 3 contains datasets with irregular-shaped clusters;
- Group 4 contains datasets with overlapped clusters;
- Group 5 contains datasets with mid and large size clusters. Dataset in Set 17 contains 31 spherical clusters of total 3100 points, with each cluster containing 100 points. Dataset in Set 18 contains 15 arbitrary-shaped clusters of total 5000 points. Dataset in Set 19 contains 10000 points that are evenly distributed in 100 clusters. Dataset in Set 20 has 28000 points in total. Different clusters in this dataset have different sizes and densities. The diameter of the largest cluster is twice as large as that of the smallest cluster.

The detailed descriptions of these datasets are illustrated in Table 1. The first column of Table 1 shows the names of datasets; the second and third columns denote the number of clusters and dimension of each dataset; the fifth and fourth columns represent the number of points in each cluster and the whole dataset, respectively.

Table 2 shows the accuracy and processing time of seven different indices on five groups of synthetic datasets in Fig. 14. Firstly, we computed the value of each cluster validity index under [$c_{\min}$, $c_{\max}$]. The suggested c$*$ occurs when the value of this index obtains its maximum or minimum. Then, we select this value of c as the evaluation result of this index. The evaluation results are analyzed as follows.

1) For datasets containing clusters of different densities (Group 1), PB cannot find the real numbers of clusters; the suggested cluster numbers by CH are smaller than the real ones except Set 4; DB, GS, XB, and MSO indices have the similar performance on these four datasets; AWCD can find the correct numbers of clusters for all datasets in Group 1.

2) For datasets containing clusters of different sizes and densities (Group 2), the numbers of clusters suggested by PB and GS are larger than the real numbers; in contrast, the computed cluster numbers by XB and CH are smaller than the real ones, which is nearest to the real numbers compared with the indices above; MSO index can suggest the correct cluster numbers for Sets 5 and 7, but it fails on Sets 6 and 8, which is smaller than the real ones; both DB and AWCD can suggest the correct numbers of clusters for these four datasets.

TABLE 1. Characteristics of 16 Datasets in Fig. 14.

Name	Clusters	Dimension	Number of Points	Data in each cluster
Set 1	3	2	600	83/164/353
Set 2	4	3	450	30/60/120/240
Set 3	5	2	930	30/60/120/240/480
Set 4	7	2	1200	50/100/150/180/200/220/300
Set 5	3	2	210	30/60/120
Set 6	4	3	450	30/60/120/240
Set 7	5	2	500	100/100/100/100/100
Set 8	7	2	1050	50/100/120/150/180/200/250
Set 9	4	2	1000	250/250/250/250
Set 10	4	2	600	24/24/276/276
Set 11	4	2	90	10/20/20/40
Set 12	4	2	110	17/25/27/41
Set 13	3	2	240	60/60/120
Set 14	3	2	300	100/100/100
Set 15	4	3	400	100/100/100/100
Set 16	5	2	500	100/100/100/100/100
Set 17	31	2	3100	100 points for each cluster
Set 18	15	2	5000	314/327/352/351/341/335/349/319/ 346/351/339/325/314/340/297
Set 19	100	2	10000	100 points for each cluster
Set 20	5	2	28000	5600/5600/5600/5600/5600

3) For datasets containing irregular-shaped clusters (Group 3), the evaluation results of the other six indices are all incorrect; AWCD index can reveal the structures of these datasets and suggest the correct cluster numbers.

4) For datasets containing overlapped clusters (Group 4), the cluster numbers computed by PB and GS are larger than the real numbers; whereas the evaluation results by CH and XB are smaller than the real ones; MSO is capable of finding the correct numbers of clusters except Set 14; AWCD can suggest the more accurate numbers of clusters for datasets with overlapped clusters.

5) For Set 17, the suggested cluster numbers by GS and PB are much larger than the real one; XB computes a smaller number; the evaluation results suggested by DB, CH, and MSO are nearest to the real one compared with GS and PB; AWCD can find the correct cluster number for Set 17. For Set 18, the numbers of clusters suggested by the other six indices are all inaccurate due to the irregular-shaped clusters; the evaluation results suggested by DB, XB, and MSO are nearest to the real number; AWCD can find the hidden structure in this dataset and suggest the correct cluster number. For Set 19, PB and GS suggest larger numbers, whereas the cluster numbers suggested by CH and DB are relatively small; the numbers of clusters computed by XB, MSO, and AWCD are close to the real one. For Set 20, the result computed by CH is much smaller than the real number; whereas GS computes a much larger one; the suggested numbers of clusters by PB, DB are nearest to the real number; XB, MSO, and AWCD can find the correct number of clusters.

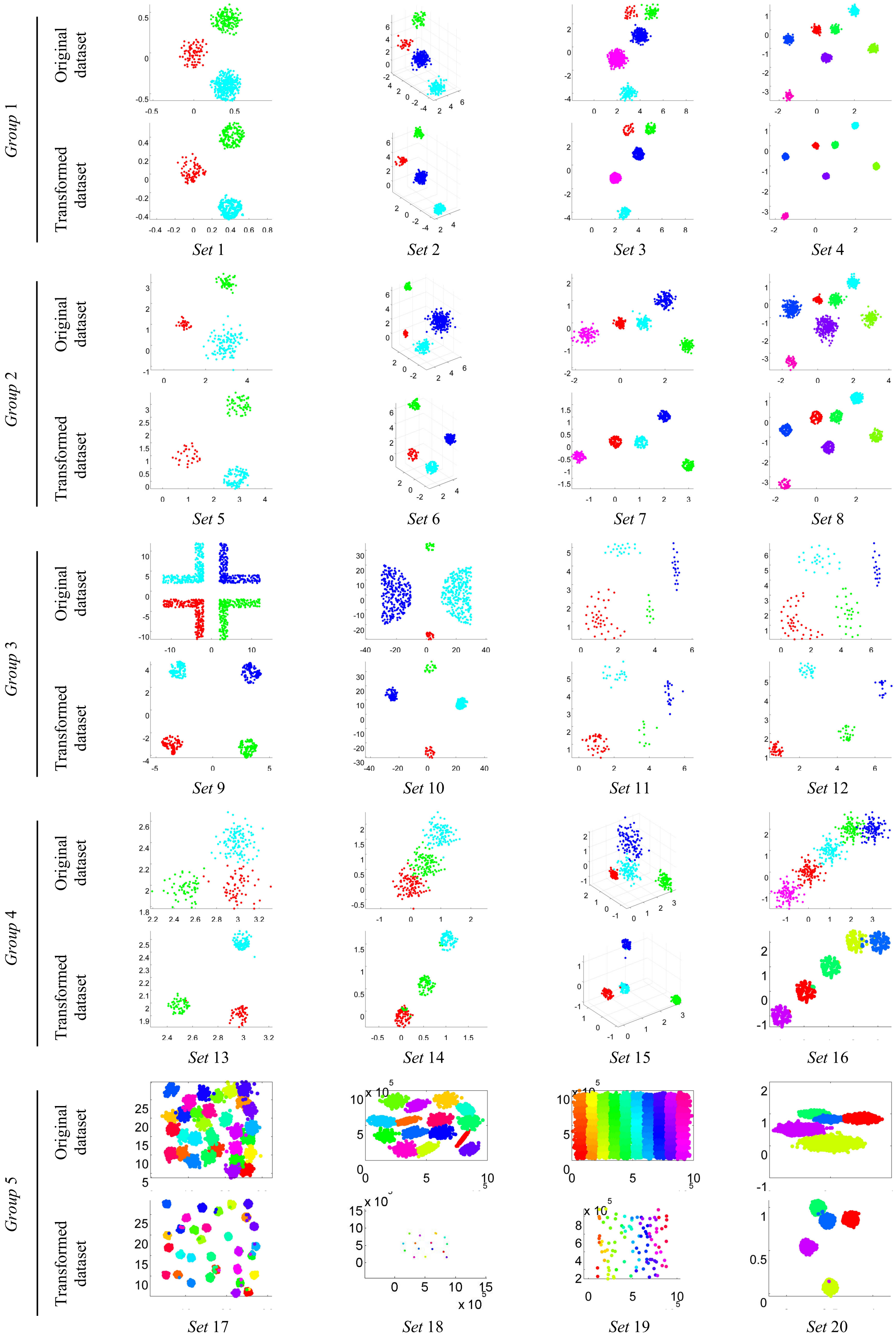


FIGURE 14. Five groups of synthetic datasets.

TABLE 2. Evaluation results of four groups of synthetic datasets.

Datasets	PB [16]		DB [13]		CH [12]		GS [15]		XB [14]		MSO [19]		AWCD	
	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT
Set 1/3	30	2.96	3√	1.50	2	2.98	3√	8.41	3√	3.55	3√	1.05	3√	2.12
Set 2/4	6	7.06	4√	6.02	2	5.95	24	51.55	4√	14.95	4√	0.50	4√	7.13
Set 3/5	8	6.86	3	6.18	2	6.09	5√	52.83	3	13.53	3	9.38	5√	6.10
Set 4/7	6	6.18	7√	5.49	7√	6.02	8	83.15	7√	11.72	7√	15.96	7√	5.98
Set 5/3	26	1.57	3√	1.48	3√	1.73	5	7.01	3√	2.92	3√	0.42	3√	2.20
Set 6/4	15	3.73	4√	3.50	3	3.98	7	20.07	3	6.97	2	2.20	4√	3.77
Set 7/5	9	2.90	5√	2.70	5√	2.77	5√	19.90	4	4.73	5√	2.64	5√	3.02
Set 8/7	6	5.50	7√	5.47	5	5.45	10	62.59	6	11.26	4	12.30	7√	5.77
Set 9/4	18	3.70	13	2.71	2	2.43	28	19.96	11	5.94	7	2.60	4√	3.05
Set 10/4	8	2.91	6	2.82	2	3.11	21	11.01	2	5.04	20	0.92	4√	2.99
Set 11/4	26	0.82	5	0.83	2	0.95	20	2.05	3	0.80	2	0.03	4√	0.88
Set 12/4	30	0.93	6	0.90	2	1.10	15	2.44	3	1.02	4√	0.04	4√	0.55
Set 13/3	29	1.76	4	1.64	2	1.77	6	4.28	3√	1.72	3√	0.12	3√	1.70
Set 14/3	30	2.39	3√	2.45	2	2.09	3√	10.93	2	4.12	2	0.87	3√	2.25
Set 15/4	30	2.46	3	2.38	3	2.66	27	7.59	3	2.87	4√	0.35	4√	2.43
Set 16/5	28	3.79	7	3.51	2	3.59	7	20.64	2	13.63	5√	2.45	5√	3.44
Set 17/31	60	17.59	35	18.30	30	21.05	43	230.27	20	33.41	32	15.85	31√	41.45
Set 18/15	26	34.13	14	26.27	2	31.11	23	530.76	14	64.18	14	35.23	15√	40.55
Set 19/100	115	61.51	95	53.38	93	54.64	120	940.32	98	105.5	99	60.13	99	77.12
Set 20/5	7	130.45	6	110.1	2	125.3	10	2068	5√	230.5	5√	113.2	5√	190

Note: NC and PT refer to the number of clusters and processing time, respectively. The unit of PT is second (s). The number marked by "√" indicates that the result evaluated by the corresponding index is correct; otherwise, it is incorrect. For the sign '*Set* x/y', x and y refer to the tested dataset and the real number of clusters, respectively. The range of cluster numbers is [2, 30] for Sets 1-16, 18, 20, [30, 60] for Set 17, and [90, 120] for Set 19.

TABLE 3. Characteristics of ten real-world datasets from UCI.

Name	Clusters	Dimension	Number of points	Data in each cluster
Banknote	2	4	1372	762/610
Cancer	2	9	683	444/239
Iris	3	4	150	50/50/50
Parkinsons	2	22	195	147/48
Pima	2	8	768	268/500
Satimage	6	36	2000	1533/1508/1358/707/703/626
Seeds	3	7	210	70/70/70
Segmentation	7	19	2310	330/330/330/330/330/330/330
Wholesale	2	7	440	298/142
Wine	3	13	178	71/59/48
Pendigits	10	16	3498	363/364/364/336/364/335/336/364/336/336
Letter	26	16	20000	789/766/736//805/768/775/773/734/755/747/739/761/792/783/753/803/783/758/748/796/813/764/752/787/786/734
GFE	2	300	27936	18059/9877

For the 20 synthetic datasets, PB index cannot suggest the real number of clusters for any tested dataset; DB, CH, GS, and XB indices can only obtain 8, 3, 4, and 6 correct results, respectively; MSO and the other five indices are not capable of computing datasets containing arbitrary-shaped clusters; our proposed AWCD method can find the correct cluster numbers for 19 datasets, outperforming the other six indices. Fig. 15 shows the error ratio between the proposed method and the other six indices based on the real numbers of clusters, illustrating the validity of AWCD.

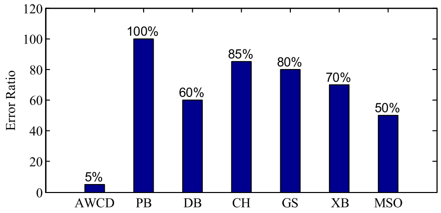


FIGURE 15. Error ratio between AWCD and the other six indices on synthetic datasets.

In sum, the evaluation results of the other six indices can be easily affected by the distributions of objects, such as the densities, sizes, shapes of clusters; in contract, the proposed validity index AWCD is effective for dealing with such a problem and shows better performance. From the processing time (PT) in Table 2, we can conclude that the

TABLE 4. Evaluation results of seven indices for ten UCI datasets.

Datasets	PB [16]		DB [13]		CH [12]		GS [15]		XB [14]		MSO [19]		AWCD	
	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT	NC	PT
Banknote	13	5.39	19	6.03	3	5.53	30	41.79	2√	9.06	3	5.06	2√	7.21
Cancer	25	5.91	2√	6.70	3	6.94	27	29.99	2√	6.23	5	12.72	2√	6.03
Iris	17	1.33	2	1.18	2	1.23	20	3.40	2	1.13	3√	0.57	3√	1.43
Parkinsons	18	1.78	11	2.03	3	1.74	14	4.84	8	2.47	2√	0.09	2√	1.89
Pima	14	7.49	3	6.93	3	7.17	27	18.70	3	6.47	2	1.44	3	7.02
Satimage	29	14.5	3	15.1	2	14.45	30	90.38	3	25.45	2	10.55	6√	15.88
Seeds	29	1.92	20	1.87	2	1.75	30	5.80	2	2.52	9	0.10	3√	2.09
Segmentation	12	15.38	12	17.23	2	14.39	22	95.44	2	30.50	4	14.31	4	15.01
Wholesale	22	6.53	8	6.43	4	6.22	28	12.42	3	4.44	4	0.48	2√	6.57
Wine	21	1.57	7	1.49	2	1.76	21	4.04	2	2.46	7	0.07	3√	1.02
Pendigits	29	49.91	24	44.37	2	38.37	30	740.32	3	64.03	2	40.99	10√	45.33
Letter	30	140.3	20	160.4	10	125.37	30	2577	24	270.1	22	140.37	24	155.3
GFE	7	210.2	5	243.2	2	205.3	15	3020	3	383.4	2√	233.3	2√	227.4

Note: NC and PT refer to the number of clusters and processing time, respectively. The unit of PT is second (s). The number marked by "√" indicates that the result evaluated by the corresponding index is correct; otherwise, it is incorrect. The range of cluster numbers is [2, 30].

time consumption depends on not only the size, but also the characteristic of the tested dataset, such as density and shape. The processing time of XB and GS is much longer than that of PB, DB, CH, and AWCD. Among the seven indices, the time consumed by GS is the longest. The runtime of MSO index is much shorter than that of the other indices since there is no need to execute clustering algorithm repeatedly.

B. TESTS ON REAL DATASETS

The UCI Machine Learning Repository contains various kinds of benchmark datasets, which is usually used for evaluating machine learning algorithms. The UCI datasets, collected from the real-world, cover a wide range of representative domains. And the characteristics of these datasets are described in detail, so that they are capable of providing a baseline for comparison.

In this paper, thirteen UCI datasets containing clusters of various sizes (*Parkinsons*, *and GFE*), density (*Wholesale*), shapes (*Satimage*), and overlapped clusters (*Seeds*, *Segmentation*, *Banknote*, *Iris*, *Cancer*, *Pima*, *Wine*, *Letter*, *and pendigits*) are selected for validating our proposed index AWCD. The detailed characteristics of these datasets are listed in Table 3. The first column denotes the names of datasets; the second and third columns represent the number of clusters and dimension of each dataset, respectively; the fifth and fourth columns denote the number of points in each cluster and the whole dataset, respectively.

Table 4 shows the accuracy and processing time of the seven indices on the thirteen UCI datasets. The numbers of clusters suggested by PB and GS are all incorrect, which is larger than the real ones; in contract, CH gives relatively smaller numbers; compared with PB and GS, the evaluation results of DB and XB are nearest to the real cluster numbers; MSO index can find the correct cluster numbers on 3 datasets, and the evaluation results are nearest to the real cluster numbers compared with the other five indices; AWCD index is capable of finding the correct numbers of clusters for ten datasets except *Pima*, *Segmentation*, and *Letter* whereas the evaluation results for the three datasets are also closest to the real ones compared with the other six indices. The comparison of error ratio is illustrated in Fig. 16.

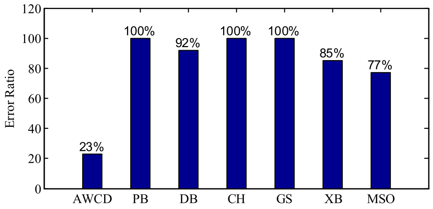


FIGURE 16. Error ratio between AWCD and the other six indices on UCI datasets.

Table 4 shows that the time consumed by PB, DB, CH, and AWCD indices is similar, which is shorter than that consumed by GS and XB indices. The processing time of GS index is the longest among the seven indices. The runtime of MSO index is the shortest among the seven indices.

V. CONCLUSION

The number of clusters in any dataset is an essential parameter, and the correct clustering evaluation results from the correct identification of cluster numbers. In this paper, we propose a novel cluster validity index based on the adjustment of within-cluster distances. This index is independent of clustering algorithms and data distributions. Thus, it does not need the users to provide prior information, which reduces the uncertainty in the evaluation process. And it performs well on datasets with density-different and shape-irregular clusters.

The effectiveness of the proposed method results from the adjustment of clusters. We believe that the adjustment process can perfectly be used to clustering process itself rather than only to the clustering evaluation process. In this direction, the clustering accuracy by typical clustering algorithms may be improved.

There are two possible opportunities for the future research. Firstly, the path finding algorithm is an important notion in clustering analysis and performs well in optimization problems. Our future work is to introduce path finding algorithm into the adjustment process to reduce the computational complexity of the proposed index. Secondly, how to identify the points in the overlapped area is still a problem in clustering analysis. The proposed transformation process may misclassify some points in the overlapped area and result in deviation. Therefore, how to correct the deviation caused by the overlapped area remains as one of our research focuses in the future.

REFERENCES

[1] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[2] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.

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

[4] H. Huang, F. Meng, S. Zhou, F. Jiang, and G. Manogaran, "Brain image segmentation based on FCM clustering algorithm and rough set," *IEEE Access*, vol. 7, pp. 12386–12396, 2019.

[5] N. Dhanachandra and Y. J. Chanu, "A new image segmentation method using clustering and region merging techniques," in *Applications of Artificial Intelligence Techniques in Engineering*. Singapore: Springer, 2019, pp. 603–614.

[6] A. Mohebi, S. Aghabozorgi, T. Ying Wah, T. Herawan, and R. Yahyapour, "Iterative big data clustering algorithms: A review," *Softw. Pract. Exper.*, vol. 46, no. 1, pp. 107–129, Jan. 2016.

[7] G. Shyamala and N. Pooranam, "A survey on online stock forum using subspace clustering," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2016, pp. 1–6.

[8] M. Du, S. Ding, and Y. Xue, "A robust density peaks clustering algorithm using fuzzy neighborhood," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 7, pp. 1131–1140, Jul. 2018.

[9] 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. 2019.

[10] M. A. Masud, J. Z. Huang, C. Wei, J. Wang, I. Khan, and M. Zhong, "I-nice: A new approach for identifying the number of clusters and initial cluster centres," *Inf. Sci.*, vol. 466, pp. 129–151, Oct. 2018.

[11] J.-S. Wang and J.-C. Chiang, "A cluster validity measure with a hybrid parameter search method for the support vector clustering algorithm," *Pattern Recognit.*, vol. 41, no. 2, pp. 506–520, Feb. 2008.

[12] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist. Simul. Comput.*, vol. 3, no. 1, pp. 1–27, 1974.

[13] 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.

[14] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

[15] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, May 2001.

[16] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognit.*, vol. 37, no. 3, pp. 487–501, Mar. 2004.

[17] S. Yang, K. Li, Z. Liang, W. Li, and Y. Xue, "A novel cluster validity index for fuzzy C-means algorithm," *Soft Comput.*, vol. 22, no. 6, pp. 1921–1931, Mar. 2018.

[18] S. Yue, J. Wang, J. Wang, and X. Bao, "A new validity index for evaluating the clustering results by partitional clustering algorithms," *Soft Comput.*, vol. 20, no. 3, pp. 1127–1138, Mar. 2016.

[19] Y. Wang, S. Yue, Z. Hao, M. Ding, and J. Li, "An unsupervised and robust validity index for clustering analysis," *Soft Comput.*, vol. 23, no. 20, pp. 10303–10319, Oct. 2019.

[20] S.-H. Lee, Y.-S. Jeong, J.-Y. Kim, and M. K. Jeong, "A new clustering validity index for arbitrary shape of clusters," *Pattern Recognit. Lett.*, vol. 112, pp. 263–269, Sep. 2018.

[21] 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, Oct. 2018.

[22] M. A. Wani and R. Riyaz, "A new cluster validity index using maximum cluster spread based compactness measure," *Int. J. Intell. Comput. Cybern.*, vol. 9, no. 2, pp. 179–204, Jun. 2016.

[23] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.

[24] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1275–1291, Jul. 2005.

[25] 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.

[26] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, 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.

[27] J. W. Han, M. Kamber, and J. Pei, "Cluster analysis: Basic concepts and methods," in *Data Mining: Concepts and Techniques* (The Morgan Kaufmann Series in Data Management Systems), 3rd ed. Boston, MA, USA, 2012, pp. 443–495.

[28] M. Kim and R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognit. Lett.*, vol. 26, no. 15, pp. 2353–2363, Nov. 2005.

[29] Q. Tong, X. Li, and B. Yuan, "Efficient distributed clustering using boundary information," *Neurocomputing*, vol. 275, pp. 2355–2366, Jan. 2018.

[30] 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.

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

[32] D. Dua and C. Graff, "UCI machine learning repository [http://archive.ics.uci.edu/ml]," Ph.D. dissertation, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019.

QI LI received the B.S. degree from Qufu Normal University, Rizhao, China, in 2017. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her current research interests include data mining, clustering algorithm, and information fusion.

SHIHONG YUE received the M.S. degree from the Xi'an University of Technology, in 1997, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2000. From 2000 to 2004, he was a Postdoctoral Researcher with the Institute of Industrial Process Control, Zhejiang University, Hangzhou, China. He is currently a Professor with Tianjin University, Tianjin, China. His current research interests include electrical tomography, medical image processing, and data mining.

YARU WANG is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her current research interests include data mining and data fusion.

MINGLIANG DING received the B.S. degree from the China University of Petroleum, Qingdao, China, in 2007. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His current research interests include electrical impedance tomography and digital signal processing.

JIA LI received the B.S. degree from North China Electric Power University, Baoding, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her current research interests include electrical impedance tomography and data fusion.