IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 21, 2019, accepted January 13, 2020, date of publication January 23, 2020, date of current version February 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968938

Volume and Surface Area-Based Cluster Validity Index

QI LI, SHIHONG YUE, AND MINGLIANG DING
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 Cluster validity index plays an important role in assessing the quality of clustering results. However, most of the existing validity indices take a trial-and-error strategy, and their correctness depend on not only the measurements of intra- and inter-cluster distances but also the specific clustering algorithms and data structures. Consequently, the applications of these indices are limited in practice. In this paper, we firstly define the total surface area and volume of all clusters in a 2-dimensinal data space, thereby recovering their natural interrelation among various numbers of clusters. On this basis, a novel validity index is proposed to directly assess the clustering results of any dataset, which does not require any trail-and-error process, clustering algorithms, data structures, or the measurements of intra- and inter-cluster distances. In the case of a high-dimensional data space, all clusters are transformed into spherical clusters of normalized size in a 2-dimensinal data space through a multidimensional scaling transformation. Two groups of typical synthetic datasets and real datasets with various characteristics are used to validate the novel validity index.

INDEX TERMS Cluster validity index, multidimensional scaling transformation, volume and surface area.

I. INTRODUCTION

Cluster analysis, with which one can find the hidden structures inside the investigated datasets, playing an important role in the domain of data mining [1], [2]. Clustering algorithms and cluster validity indices are two most important tasks in cluster analysis [3],[4]. Determining the optimal number of clusters is usually completed based on one cluster validity index or several [5], [6]. A great number of validity indices have been proposed, ranging from the typical Davies-Bouldin measure (DB) [7] to the latest unsupervised cluster validity index [8]. Various validity indices have played a very important role in evaluating results from any clustering algorithm. For example, in a novel way both internal and external validity indices were used to evaluate clustering task when clustering was carried on a large high dimensional dataset [9]. The symmetry validity in [10] was used to determine the number of clusters so that the human precuneus can effectively be subdivided to six connected parcels by using an eigen clustering approach. In [11], the notation of cluster has led to scaling hierarchical power efficient clustering with energy aware routing, and so on.

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

Nowadays, under the background of big data [12], [13], the efficiency of cluster validity index has become a focus beside clustering algorithm [14]. Various similarity norms may greatly affect the accuracy of the cluster validity [15], [16]. The computational complexity of each iteration when using Bregman clustering algorithms [17] is linear with respect to the number of data points. Therefore, these related algorithms are scalable and appropriate to largescale machine learning tasks. Recently, graph-based algorithms have effectively been applied to express clustering structure and natural relation hidden in investigated objects. For instance, the DB index was used as fitness function to evaluate the quality of the clusters in a largescale dataset [18], several indices commonly evaluated the clustering results from large uncertain graphs [19] or the minimum spanning tree [20] that were used to present the clustering structure.

The above methods have their own applicable and efficient ranges. However, they cannot solve the classical clustering assessment problems such as measurements of intra- and inter-cluster distances [21], or multiple time repeated computation of clustering algorithms [22]. In this paper, efforts have been made to solve the above problems. After uncovering the interrelation between surface area and volume of all clusters in a dataset, a novel validity index is proposed.

II. RELATED WORK

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset consisting of n points in a d-dimensional space. When X is partitioned into c subsets, i.e., $C_1, C_2, \ldots C_c$, a binary membership function U can be used to describe the relationship between points and subsets [23]. If point x_j belongs to the ith subset C_i, then u_{ij} is equal to 1; otherwise, 0. The binary membership function can be expressed as

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

A hard partitioning of X means that each point in X only belongs to one subset [24]. In this situation, all points are divided into c disjoint subsets, i.e.,

$$X = C_1 \cup C_2 \cup \ldots \cup C_c,\ C_i \cap C_j = \emptyset, \quad i, j = 1, 2, \ldots, c \tag{2}$$

On the contrary, if each point in X belongs to all subsets with its individual membership degrees, then this kind of partition is called fuzzy partitioning [25]. Here, the partitioning matrix satisfies

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

Generally, a validity index is a function of the number of clusters (c), which combines the intra- and inter-cluster distances [26], [27]. A validity index is usually denoted as

$$\min(\max)\ z = f(\phi_c,\ \delta_c), \quad c = 1, 2, \ldots, C \tag{4}$$

where ϕ_c and δ_c denote the intra- and inter-cluster distances, respectively.

By minimizing ϕ_c and maximizing δ_c, (4) can reach its maximum or minimum. Most validity indices take a trial-and-error way to find the optimum of (4) [28]. Various combinations of δ_c and ϕ_c can result in different kinds of indices [29]. Five typical cluster validity indices will be described as below. Here, the upward arrow indicates that the maximum of the corresponding index refers to the optimal partition, and the corresponding number of clusters denotes the optimal number of clusters. In contrast, the downward arrow indicates the opposite meaning.

A. CALINSKI-HARABASZ INDEX (CH ↑) [30]

The compactness of this index is computed in terms of the distances between each point and the centroid of a cluster, and the separation is estimated by the distances between centroid of each cluster and the global centroid. Thus, for a dataset containing n points, CH can be defined as

$$CH(c) = \frac{n-c}{c-1} \cdot \frac{\sum_{i=1}^{c} n_i ||z_i - z||^2}{\sum_{i=1}^{c} \sum_{k=1}^{n_i} ||x_k - z_i||^2} \tag{5}$$

where n_i and z_i denote the number of points and the centroid of cluster i, respectively; and z is the centroid of the whole dataset.

B. DAVIES-BOULDIN INDEX (DB ↓) [13]

Let Δ_i and z_i denote the compactness and centroid of cluster i, respectively; δ_{ij} represents the separation between clusters i and j. DB can be expressed as

$$DB(c) = \sum_{i=1}^{c} R_i/c, \quad s.t., \ \begin{cases} R_i = \max_{j, j\neq i}(\Delta_i + \Delta_j)/\delta_{ij} \\ \delta_{ij} = ||Z_i - Z_j|| \\ \Delta_i = \sum_{x\in C_i} ||x_i - z_i||/|C_i| \end{cases} \tag{6}$$

where $|C_i|$ denotes the number of points in cluster i.

C. TIBSHIRANI'S GAP STATISTIC INDEX (GS ↑) [31]

GS can be defined as

$$W(c) = \sum_{i=1}^{c} D_i/(2|C_i|), \quad s.t., \ D_i = 2|C_i| \sum_{j\in C_i} ||x_j - \overline{x}||, \ \overline{x} = \sum_{i=1}^{|C_i|} x_i/|C_i| \tag{7}$$

The optimal number of clusters appears at the inflection point on the curve computed by (7). Owing to the subjectivity of the detection of inflection point, the gap statistics can be formulated as,

$$gap(c) = E^*[\log(W(c))] - \log(W(c)), \quad s.t., \ W(c) = \sum_{i=1}^{c} D_i/(2|C_i|) \tag{8}$$

where E^* refers to the expectation under a null reference distribution.

D. PAKHIRA AND BANDYOPADHYAY' INDEX (PB ↑) [32]

PB is proposed by Pakhira and Bandyopadhyay to evaluate the clustering results from both hard and fuzzy algorithms, i.e.,

$$PB(c) = \left(\frac{1}{c} \times \frac{E_1}{J} \times D_c\right)^2, \ s.t., \ \begin{cases} E_1 = \sum_{j=1}^{n} ||x_j - z|| \\ D_c = \sum_{i,j=1}^{c} ||z_i - z_j|| \\ J = \sum_{i=1}^{c} \sum_{j=1}^{n} ||x_j - z_i|| \end{cases} \tag{9}$$

E. XIE–BENI'S SEPARATION INDEX(XB ↓) [33]

XB is designed for fuzzy clustering algorithms, which is the ratio of compactness to separation of a dataset, i.e.,

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m ||x_j - z_i||^2}{n \cdot \min_{i\neq j} ||z_j - z_i||^2} \tag{10}$$

where m denotes the fuzzy exponential.

However, there are at least three problems with the existing validity indices.

1) Measurement of intra- and inter-cluster distances. Various measurements may lead to different assessment results based on distances such as Euclidean and Hausdorff [34]. Recently, the Bregman divergence has been applied in the process of assessing clustering results [35]. In addition, the line symmetry distance measures [36] can enhance the efficacy of existing widely used validity indices

and this method can deal with clusters of any shape or size in a given dataset. Nevertheless, there is no fixed rule for choosing the optimal measurement, and how to combine these measurements is still also a challenge.

2) Dependence on clustering algorithms. The existing validity indices depend on specific clustering algorithms, e.g., C-means algorithm [37]. In the case of clustering results obtained using other clustering algorithms, these indices will not be applicable.

3) Trial-and-error way. The existing validity indices take the trial-and-error way to find the optimal number of clusters [38]. However, in the case of a highly scalable dataset, the time consumption is intolerable since the clustering algorithm has to be performed repeatedly.

Recently, some advanced clustering techniques have been proposed to deal with datasets and evaluate the clustering results in complicated situations. Tong *et al* proposed a Scalable Clustering Using Boundary Information (SCUBI) algorithm [39], which can obtain almost the same clustering results as those obtained using the existing clustering algorithms when dealing with some typical datasets. Dunn's [40] cluster validity index has quadratic time complexity $O(pn^2)$, where p denotes the dimension of the dataset. As a result, its computation is impractical for datasets with large values of n.

The typical validity indices cannot solve the above problems. For example, the improved Dunn index [41] relies on the trail-and-error strategy and specific clustering algorithms. As for SCUBI, the intra- and inter-cluster distances cannot be defined well. In this paper, our novel validity index aims to take advantage the interrelation between surface area and volume of a dataset in an unsupervised manner.

III. NOVEL VALIDITY INDEX

In a d-dimensional data space, any cluster occupies a distributed space position. If the cluster is assumedly spherical, then there will be a natural relation between its hyper-surface area and hyper-volume, which can be used to construct a novel validity index as follows.

In any 2-dimensional (2-D) data space, the hyper-surface area and hyper-volume can be reduced to area and perimeter of all clusters. Firstly, we define the area and perimeter in a 2-D data space, where all clusters are assumedly spherical. In the case of a high-dimensional data space and arbitrary-shaped clusters, we use the Multidimensional Scaling (MDS) [42], [43] method to map all clusters to a 2-D data space, and transform the arbitrary-shaped clusters to approximately spherical clusters of normalized size using the notation of chain. By revealing the interrelation between area and perimeter, a novel cluster index can be formulated.

A. AREA AND PERIMETER OF CLUSTER IN 2-D DATA SPACE

In this section, we will introduce the notations of area and perimeter of all clusters of a dataset X in any 2-D data space. For any point $x_i \in X$, we approximate its neighborhood by a rectangle, whose side length and area can be characterized by its k-nearest neighbors.

Definition 1 (The Side Length and Area of a Point): Let $KNN_4(x_i)$ be the set of 4-nearest neighbors of x_i in X (see Fig. 1), then the side length of x_i can be defined as

$$l_i = \frac{1}{8}\sum_{j\in KNN_4(x_i)} dist(x_i, x_j) \tag{11}$$

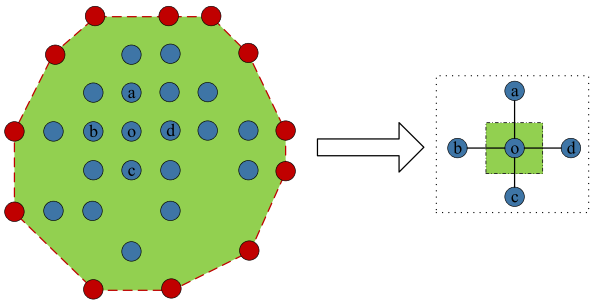


FIGURE 1. Occupied area in a 2-D data space.

And the area occupied by x_i is computed as

$$s_i = l_i^2 \tag{12}$$

where $dist(x_i, x_j)$ denotes the Euclidean distance between points x_i and x_j. Since $dist(x_i, x_j)$ is shared by x_i and its neighbor x_j, l_i can be represented by half of the average 4-nearest neighbors' distances. Consequently, the total occupied area S of all points in X is

$$S = \sum_{i=1}^{n} s_i \tag{13}$$

where n is the number of points in X.

Definition 2 (Density): For any data point $x_k \in X$, its m nearest neighbors are denoted as, $x_{k,1}$, $x_{k,2}$,..., $x_{k,m}$, with distances $dist\,(x_k, x_{k,1})$, $dist(x_k, x_{k,2})$,..., $dist(x_k, x_{k,m})$, where $m = 2d$. Thus, in a 2-D space, $m = 4$.

$$density(x_k) = \{\sum_{j=1}^{m} dis(x_k, x_{k,j})\}^{-1}, \quad k = 1, 2, \ldots, n \tag{14}$$

Different from the existing density notations [44], [45], the proposed density is nonparametric, which does not need any prior information and can reduce uncertainties in practice.

Definition 3 (Boundary and Interior Points): Point x_i in X is called a boundary point if half of its 4-nearest neighbors have higher density than its own; otherwise, it is called an interior point.

Hereafter, let BX and IX denote the set of boundary and interior points in X, respectively.

Definition 4 (Perimeter): The perimeter of all clusters in X can be defined as

$$P = \sum_{j\in BX} l_j \tag{15}$$

Fig. 1 shows a cluster in a 2-D data space, whose boundary points are determined after densities of all points are computed. Note the boundary and interior points are marked

in red and blue, respectively; the area in green denotes the occupied area of this cluster, and the perimeter represented by red dotted lines is computed according to the determined boundary points. The right part of Fig. 1 also shows the occupied area of point o and the length of this rectangle denotes the side length of point o, demonstrating that any point can be measured by its 4-nearest neighbors.

B. MULTIDIMENSIONAL SCALING

In the case of a data space whose dimensionality is larger than 2, the occupied positions of all points have to be measured in terms of the hyper-volume and hyper-surface area in principle. However, they are difficult to measure. In this paper, MDS is used to project high-dimensional datasets into a 2-D data space, where the hyper-volume and hyper-surface area are reduced to area and perimeter, respectively.

MDS can reveal the structure of any dataset in a two/three-dimensional data space by constructing a low-dimensional configuration [46], which aims to preserve distances between points so that the structure of the dataset is unchanged [47]. The typical process of MDS is illustrated as follows. For any dataset X, the distance matrix $M \in R^{n\times n}$ can be defined as

$$M_{ij} = \left(dist^2\left(x_1, x_j\right) + dist^2\left(x_i, x_1\right) - dist^2\left(x_i, x_j\right)\right)/2 \quad (16)$$

Then, M can be decomposed by eigenvalue decomposition as

$$M = USU^T \quad (17)$$

where U is an $n \times n$ matrix of eigenvectors, and S is an $n \times n$ diagonal matrix whose diagonal elements are the corresponding eigenvalues.

Finally, the mapping coordinates $Y \in R^{n\times n}$ can be computed as follows.

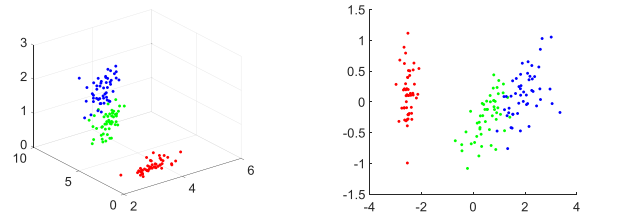
$$Y = U\sqrt{S} \quad (18)$$

Generally, the first two columns of Y represent the whole matrix Y with a small deviation [48], which is further chosen as the mapping coordinates in the corresponding 2-D data space in view of the tradeoff between accuracy and complexity.

Fig. 2 shows the well-known IRIS and Helix datasets [49], and the contained clusters from a 3-D distribution are transformed into a 2-D data space. In view of the unchangeable density characteristic of MDS, the transformed clusters keep both their mutual positions and the positions of all points unchangeable. Consequently, in the 2-D data space, there are identical clustering structures and points distribution. As a result, the correct number of clusters in any dataset can be estimated in the corresponding 2-D data space.

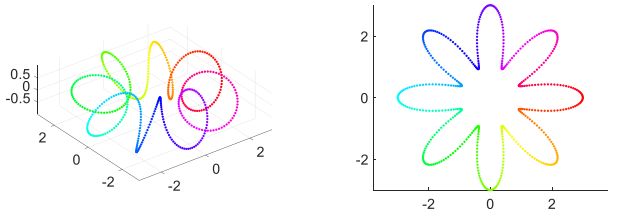
C. PROPOSED VALIDITY INDEX

Assume that a dataset X in a d-dimensional data space contains c clusters which are nearly spherical with the same radius r, and then the hyper-volume V_d[50] and hyper-surface area S_d of any cluster can be computed as follows.



(a) Three 3-D clusters of IRIS.

(b) Three transformed 2-D clusters of IRIS by MDS.



(c) 3-D cluster of Helix.

(d) Transformed 2-D cluster of Helix by MDS.

FIGURE 2. Two transformed datasets by MDS from a 3-D to 2-D data space.

$$V_d(r) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d \text{ and } S_d(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1} \quad (19)$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ $(x > 0)$ is the gamma function [51], satisfying the recurrence equation as below

$$\Gamma(x+1) = x \cdot \Gamma(x) \quad (20)$$

In the case of c clusters, their total hyper-volume and hyper-surface area are

$$V = c \cdot V_d(r) \text{ and } S = c \cdot S_d(r) \quad (21)$$

When V and S are known, the unknown variables of c and r can be computed by

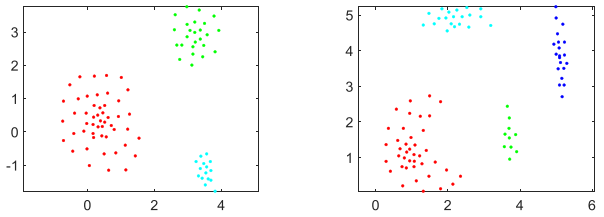
$$c = \frac{\pi^{-d/2}\Gamma(d/2)}{2d^{d-1}} \cdot \frac{S^d}{V^{d-1}} \quad (22)$$

Considering that the computations of V and S in a high-dimensional data space are difficult, we apply MDS to transform these clusters into the corresponding 2-D data space, and whereby V and S can be computed by (13) and (15).

However, all clusters in any dataset are not often nearly spherical with the same radius r, i.e., (22) cannot directly recognize the correct number of clusters when a dataset contains different-sized and arbitrary-shaped clusters (see Fig. 3).

To solve this problem, the notation of chain is introduced, based on which the arbitrary-shaped clusters can be transformed into spherical clusters and the size of each cluster can be normalized at the same time.

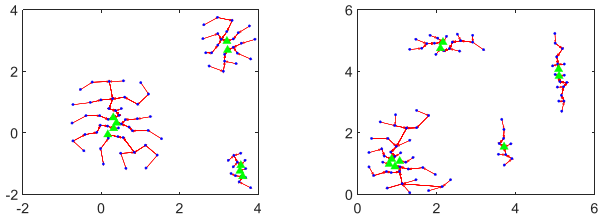
For each point x_i, we define a density-based distance σ_i, i.e., the minimum distance from x_i to other points with a higher density than x_i. The corresponding point is called the



(a) Dataset with different-sized clusters.

(b) Dataset with arbitrary-shaped clusters.

FIGURE 3. Datasets containing different-sized and arbitrary-shaped clusters.



(a) Key points and chains in different-sized clusters.

(b) Key points and chains in arbitrary-shaped clusters.

FIGURE 4. Distributions of key points and chains in the two datasets in Fig. 3.

nearest density-based neighbor φ_i. Here, σ_i and φ_i can be formalized as

$$\sigma_i = \min_{j:\rho_i>\rho_j} dist\left(x_i, x_j\right) \tag{23}$$

$$\varphi_i = \arg\min_{j:\rho_i>\rho_j} dist\left(x_i, x_j\right) \tag{24}$$

Definition 5 (Key Points): Points with relatively higher density and larger value of σ_i are regarded as key points, which can be determined by the value of kp_i, i.e.,

$$kp_i = \rho_i \cdot \sigma_i \tag{25}$$

In general, the maximal number of clusters is less than $|\sqrt{n}|$ [52], where $|\cdot|$ is an integer operator. Thus, the number of key points can be set as $|\sqrt{n}|$.

The adjacent points in X can be connected following the connecting rule. For any point x_i, the next point x_j is the nearest density-based neighbor of x_i. The above steps are repeated till a key point is visited.

Definition 6 (Chain): A chain is a subset of points in X, i.e., $x_{i1}, x_{i2}, \ldots, x_{ik}$, which starts with x_{i1} and stops at a key point x_{ik} according to the above connecting rule. The length of chain T_i is defined as

$$T_i = \sum_{i=1}^{k-1} dist\left(x_{ik}, x_{ik+1}\right), \quad i = 1, 2, \ldots, \sqrt{n} \tag{26}$$

where $dist(x_{ik}, x_{ik+1})$ is the distance between adjacent points on the i th chain.

Considering that the two datasets in Fig. 3 both have 90 points, then all points in each dataset can be divided into $\sqrt{90}$ chains, i.e., 9 chains (see Fig. 4), where the key points are marked by green triangles and the red lines with arrows denote the directions of chains. Fig. 4 shows that each chain contains a group of points. In most cases, due to the different-sized and arbitrary-shaped clusters, different chains have different numbers of points and lengths.

To normalize the size of each cluster and make the shape of each cluster spherical, the j th line segment $dist(x_{kj}, x_{kj+1})$ on any k th chain is transformed into a new one, i.e.,

$$dist^*\left(x_{kj}, x_{kj+1}\right) = dist\left(x_{kj}, x_{kj+1}\right)/T_k, \quad k = 1, 2, \ldots, \sqrt{n} \tag{27}$$

By using (27), the lengths of long chains will be shortened whereas those of the short chains will be relatively enlarged. 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 after all chains are transformed according to (27). Fig. 5 shows a detailed transformation process, with the smallest cluster in Fig. 3 (a) as an example.

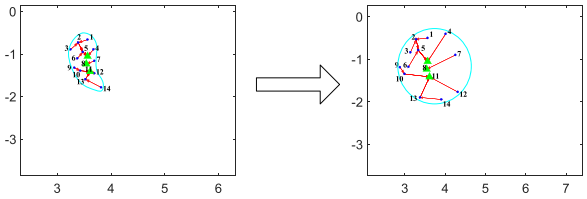


FIGURE 5. Original and transformed points in the smallest cluster in Fig. 3(a).

In Fig. 5, the numbers accompanying these points are ranked in order, and the corresponding points after transformation are marked with the numbers unchanged. The area occupied by each cluster is circled by a curve. The three chains (9-10-11, 12-11, 14-13-11) in the left of Fig. 5 have different lengths. And after transformation, the lengths of the three chains are similar, which can be illustrated in the right of Fig. 5. Fig. 5 shows that the shape of the cluster can be normalized after transformation. In addition, the noise points can be assigned to the nearest clusters by using the notation of chain, which has no effect on the evaluation process.

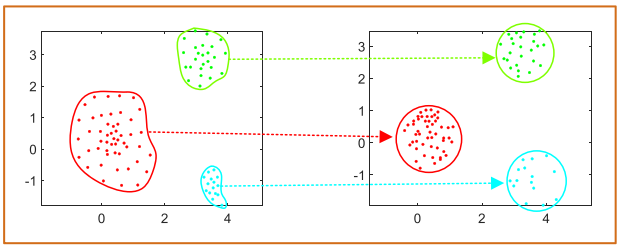
Fig. 6 illustrates the transformation results of datasets in Fig. 3. And different colors represent different clusters and the occupied area of each cluster is circled by a corresponding curve. The dotted lines with arrows connect the original cluster and the corresponding transformed cluster. Fig. 6 shows that the transformed datasets have spherical clusters of normalized size, which is consistent with the assumption above.

Hereafter, the proposed volume and area-based index is called VAI. The evaluation process of VAI is listed in **Algorithm 1**.

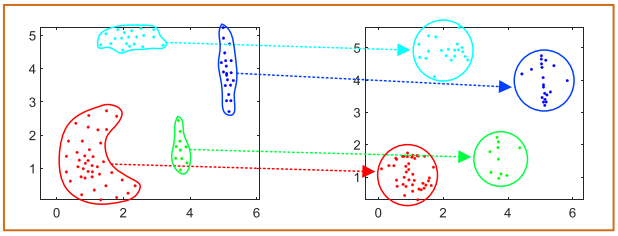
Compared with the existing indices, VAI has the following characteristics:

1) UNSUPERVISION

Since VAI is nonparametric and does not need any prior information (e.g., the clustering results from a specific clustering algorithm), the entire evaluation process is independent



(a) Different-sized clusters before and after transformation.



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

FIGURE 6. **Original and transformed clusters in two datasets.**

Algorithm 1 Evaluation Process of VAI

Input: A dataset $X \in R^d$ containing n points.
Output: Number of clusters.
Steps:

1) Map X into a 2-D data space using MDS.
2) Transform the mapping dataset into a normalized dataset with spherical clusters and normalized size.
3) Compute the density of each point in 2) and partition the points into boundary and interior points.
4) Compute the side length of each point in the transformed dataset using (11).
5) Compute the area of the transformed dataset using (13).
6) Compute the perimeter of the transformed dataset using (15).
7) Compute the number of clusters c using (22).
8) Stop and give the number of clusters.

of underlying clustering. In contrast, most of the existing indices take a trial-and-error way, which relies on clustering algorithms and can only work for spherical clusters.

2) GENERALITY

VAI can reflect the structure hidden in any dataset no matter what distributions it has. In addition, it can evaluate datasets containing noise points.

3) NO DIMENSION CONSTRAINTS

VAI has no limitation in dimensions, and it can suggest the number of clusters in high-dimensional datasets.

For any dataset containing n objects distributed in c clusters, the computation of VAI mainly consists of three parts: 1) computing all distances in X, 2) mapping all points in X into the corresponding 2-D data space using MDS, and 3) normalizing all distances in any chain. The computational complexity of the first part is $O(n^2)$. The runtime of the second part is the longest, since the computation of eigenvalues and eigenvectors leads to computational complexity $O(n^3)$. Note the runtime of the third part is much shorter than that of the second one. Therefore, the efficiency of the second part is key to reducing the runtime of VAI.

IV. EXPERIMENTAL RESULTS

To validate VAI, experiments are carried out on synthetic and real datasets. Four existing hard validity indices (i.e., CH, DB, GS, and PB) and one fuzzy validity index XB are used to make a comparison.

A. TESTS ON SYNTHETIC DATASETS

Fig. 7 shows seven groups of datasets with various characteristics. Different colors represent different clusters. Datasets in the first column denote the original datasets without noise points, and those in columns 2–4 are generated by adding 10%, 20%, and 30% uniformly distributed noises to the original ones, respectively, with "+" denoting a noise point. *Groups* 1–4 show regular datasets of spherical clusters, where *Groups* 1 and 2 contain clusters of different sizes and *Groups* 3 and 4 contain clusters of different densities. Datasets in *Group* 5 have 15 spherical clusters, and those in *Groups* 6 and 7 have arbitrary-shaped clusters.

Fig. 8 shows the corresponding transformed datasets of Fig. 7, indicating that the transformation rule can transform irregular clusters into spherical clusters and the occupied area of each cluster is normalized. Moreover, the noise points in datasets can be assigned to the nearest clusters and have no effect on the evaluation process, demonstrating the robustness of VAI.

Table 1 lists the evaluation results of datasets in Fig. 7. The number marked by "$\surd$" denotes that the evaluation result based on the corresponding index is true; otherwise, it is wrong.

The validity indices are analyzed as follows.

1) *Different Sizes:* Datasets in *Groups* 1 and 2 contain clusters of different sizes, on which CH, DB, and GS have similar evaluation results, i.e., they are capable of determining the correct numbers of clusters when datasets have fewer noise points (e.g., 10%). However, when the proportion of noise points is higher than 10% (e.g., 20% and 30%), the evaluation results of these indices are incorrect. XB suggests the correct cluster numbers for *Sets*1–6. PB cannot find the correct numbers of clusters in these eight datasets. On the contrary, VAI can give the correct numbers of clusters in all datasets.

2) *Different Densities:* Datasets in *Groups* 3 and 4 contain clusters of different densities. The evaluation results of CH are all 2, which is relatively smaller. On the contrary, PB gives relatively larger numbers. DB, GS, and XB all can find the correct cluster numbers for *Sets*9–11 that contain fewer noise points. When the proportion of noise points is larger (e.g., *Set* 12), DB, GS, and XB cannot give correct evaluation results. As for *Group* 4, the other five indices

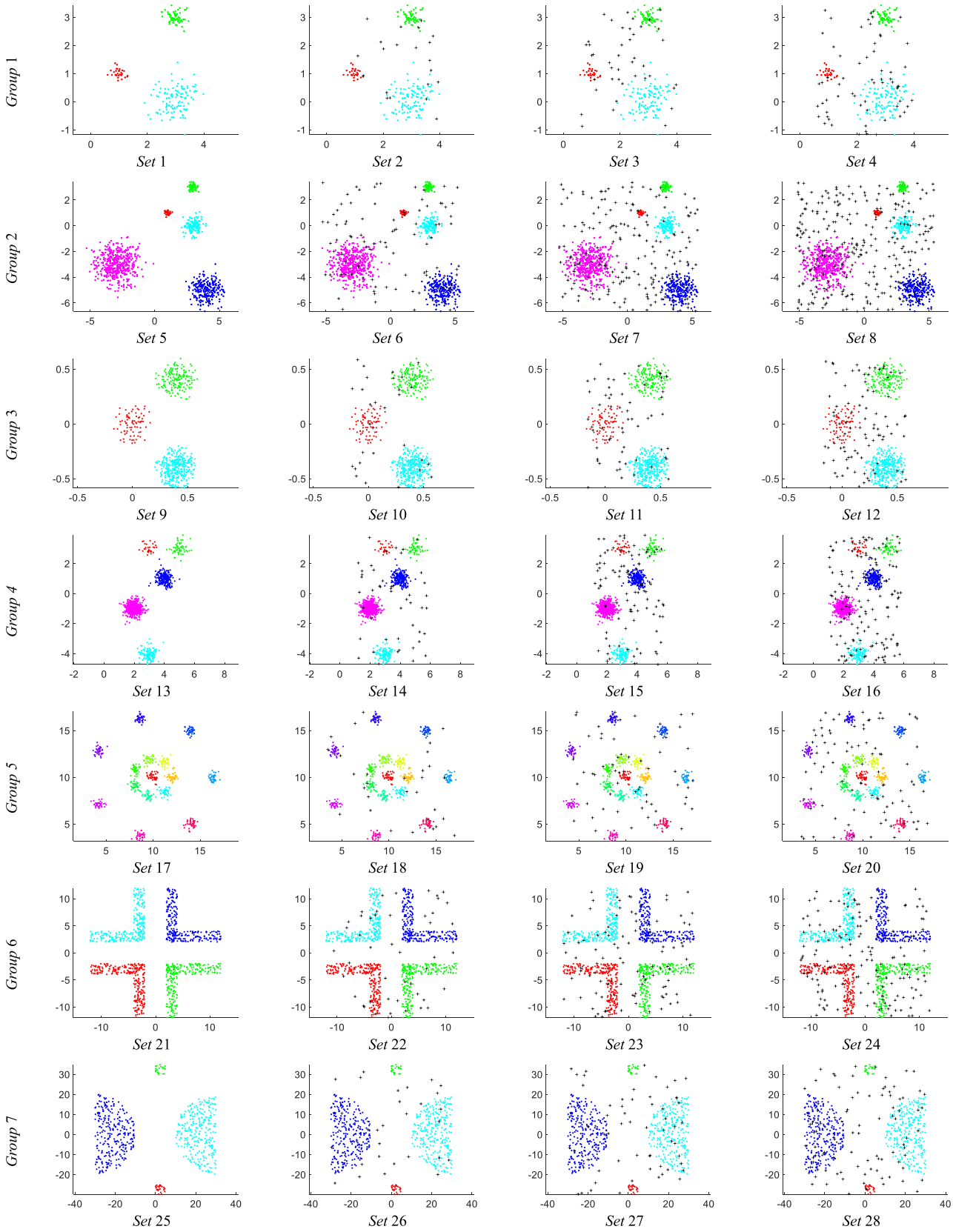


FIGURE 7. Seven groups of synthetic datasets.

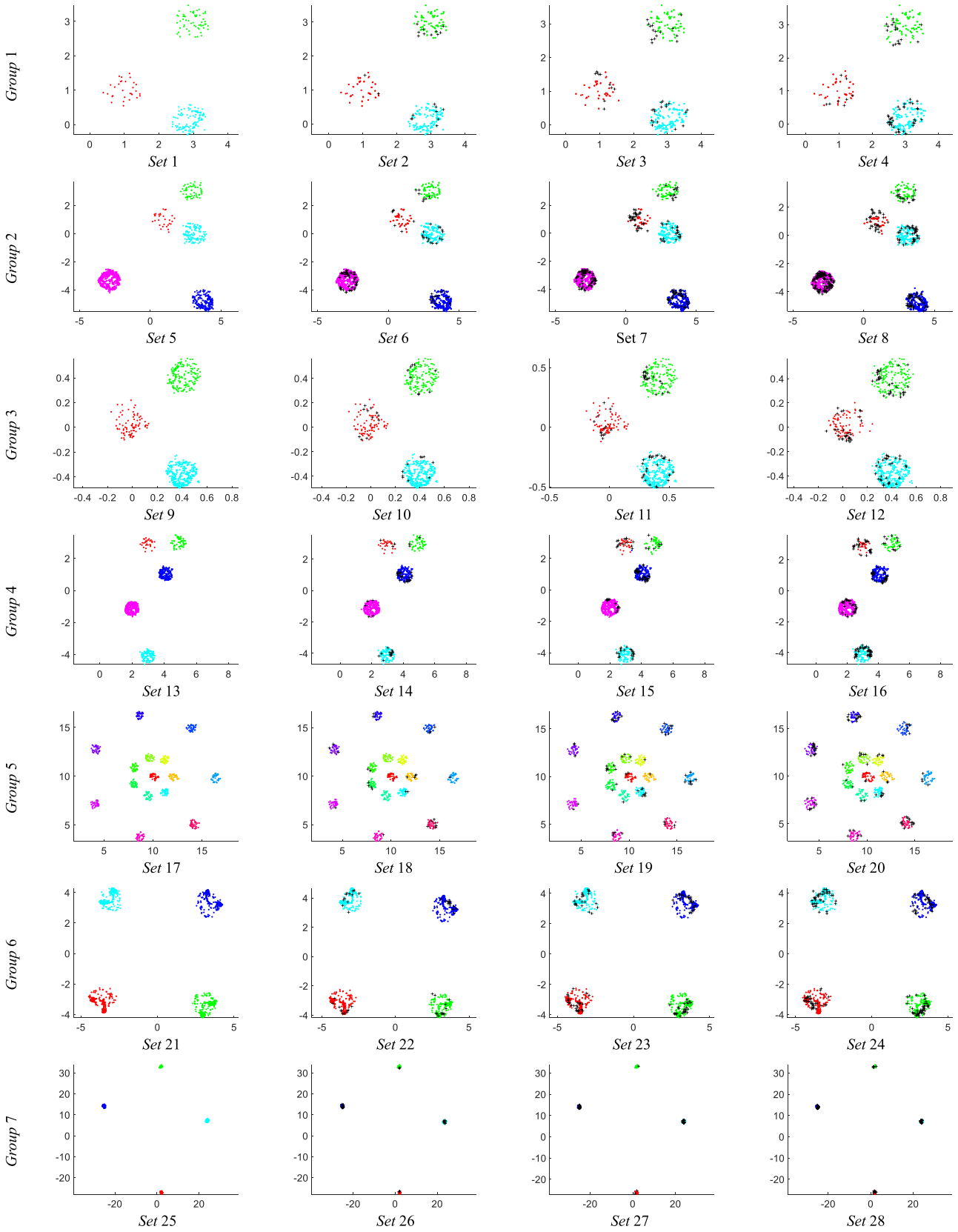


FIGURE 8. Transformed results of datasets in Fig. 7.

TABLE 1. Evaluation results of seven groups of synthetic datasets.

Datasets	CH	DB	GS	PB	XB	VAI
Set 1/3	3√	3√	3√	20	3√	3.0√
Set 2/3	3√	3√	3√	21	3√	3.0√
Set 3/3	4	5	5	20	3√	2.9√
Set 4/3	5	5	5	26	3√	3.1√
Set 5/5	5√	5√	5√	26	5√	5.0√
Set 6/5	5√	5√	5√	29	5√	4.9√
Set 7/5	3	3	3	24	3	5.1√
Set 8/5	3	3	3	30	3	5.2√
Set 9/3	2	3√	3√	28	3√	3.0√
Set 10/3	2	3√	3√	27	3√	3.0√
Set 11/3	2	3√	3√	26	3√	3.0√
Set 12/3	2	4	4	28	4	3.0√
Set 13/5	2	5√	8	23	3	5.0√
Set 14/5	2	5√	4	24	3	5.0√
Set 15/5	2	3	6	27	4	5.0√
Set 16/5	2	3	4	27	3	5.1√
Set 17/15	2	8	16	8	15√	15.0√
Set 18/15	2	8	16	8	11	15.0√
Set 19/15	2	8	22	8	10	15.0√
Set 20/15	2	8	21	8	10	15.1√
Set 21/4	2	18	24	25	11	4.0√
Set 22/4	2	18	18	25	11	4.0√
Set 23/4	2	12	26	26	12	4.0√
Set 24/4	2	12	25	24	12	4.0√
Set 25/4	2	6	22	8	2	4.0√
Set 26/4	2	6	12	24	2	4.0√
Set 27/4	2	7	20	23	2	4.0√
Set 28/4	2	14	8	23	2	4.0√

Note: for "*Set x/y*", *x* and *y* refer to the investigated dataset and the correct number of clusters, respectively.

cannot obtain the correct cluster numbers in most cases. VAI is capable of finding the correct cluster numbers for datasets in *Groups* 3 and 4.

3) *Large Numbers of Clusters:* When datasets have a large number of clusters (see *Group*5), CH, DB, PB, and XB will give relatively smaller numbers. On the contrary, GS gives the opposite evaluation results, which is the nearest to the correct cluster number. VAI can obtain the correct cluster numbers for datasets in *Group* 5.

4) *Arbitrary Shape:* Datasets in *Groups* 6 and 7 contain arbitrary-shaped clusters. The evaluation results of the other five indices are all incorrect. On the contrary, VAI can reveal the structures of datasets and determine the real numbers of clusters, regardless of noise points in datasets.

In summary, the evaluation results of the other five indices may be affected by the distributions and noise points in the investigated datasets. When the proportion of noise points becomes larger, the evaluation results will be worse. VAI can find the hidden structures in datasets and suggest the correct numbers of clusters for all these datasets.

B. TESTS ON REAL DATASETS

The UCI Machine Learning Repository [53] contains many kinds of databases, domain theories, and data generators. The datasets in UCI are from the real-world, covering a wide range of domains so that they are relevant and representative. The characteristics of these datasets are introduced in detail, such as the attribute types, number of instances, number of attributes and year published. UCI datasets are usually used to evaluate machine learning algorithms and provide a useful baseline for comparison.

In this paper, eight real datasets from UCI are selected to test the proposed VAI index. The characteristics of these datasets are listed in Table 2. The first column of Table 2 denotes the names of these datasets. The second and fourth columns represent the numbers of clusters and points of datasets, respectively. And the third column denotes the dimensions of these datasets. The last column shows the number of points of each cluster in datasets.

TABLE 2. Characteristics of eight real datasets from UCI.

Name	Number of clusters	Dimension	Number of points	Number of each cluster
Cancer	2	9	683	444/239
Seeds	3	7	210	70/70/70
Iris	3	4	150	50/50/50
Ecoli	8	7	336	143/77/52/35/20/5 /2/2
Satimage	6	36	2000	1533/1508/1358/ 707/703/626
Vertebral	3	6	310	60/150/100
Wholesale	2	7	440	298/142
Wine	3	13	178	71/59/48

1) *Cancer* has 699 points in total, and the number of features of each point is 9. In this paper, we remove 16 records which have missing features, so the number of remaining records is 683. One cluster has 444 records which represents the cluster *Benign*, and the other has 239 records denoting the cluster *Malignant*.

2)*Seeds* contains 210 points with 7 features. The number of clusters is 3 and each cluster has 70 instances.

3) *Iris* has 150 points in total and each point has four attributes. It has three clusters and each cluster has 50 points. One cluster is separated from the other two clusters, whereas the latter two clusters are overlapped with each other.

4) *Ecoli* is a nonlinear dataset with 8 clusters. It contains 336 instances with 7 features. And the majority of clusters have different numbers of instances except the last two clusters. The numbers of points in the last three clusters are much less than the other clusters, which can be regarded as noise instances.

5) *Satimage* dataset has 2000 samples consisting of 6 clusters. Each sample has 36 attributes. And these clusters have various shapes and sizes.

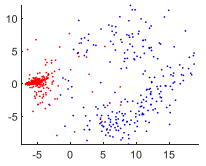
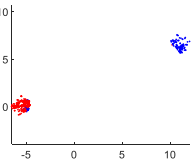
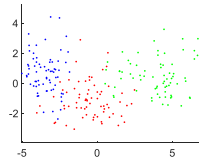
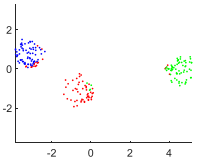
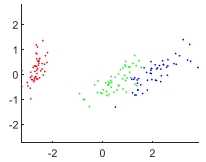
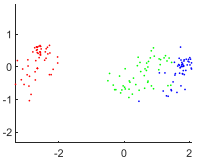
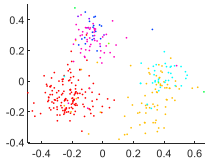
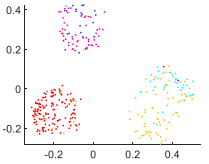
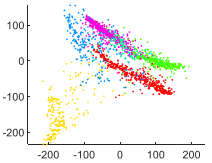
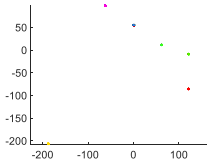
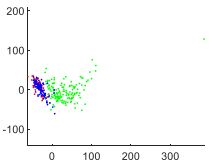
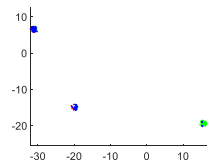
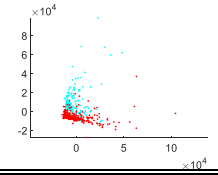
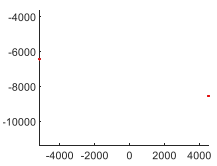
6)*Vertebral* has 310 instances in total, and each instance has 6 features. And the number of clusters is 3. Its characteristics are similar to those in *Iris*.

7) *Wholesale* dataset contains 440 points with 7 attributes. There are two clusters in this dataset, containing 298 and 142 points, respectively.

8) *Wine* is a dataset with a relatively higher dimension, and each point in *Wine* has 13 attributes. The three clusters in this dataset are nonlinear and mutually overlapped, which is similar to those in *Iris*.

When evaluating datasets in a high-dimensional data space, VAI maps these datasets into a 2-D data space at first. The mapping results of UCI datasets in Table 2 are listed in Table 3. Columns 2 denotes the mapping results by using MDS algorithm. Columns 3 denotes the transformation results by using the transformation rule above. Points in different colors represent different clusters, which is based on the true labels of these points.

TABLE 3. Mapping and transformation results of eight UCI datasets.

Name	Mapping results obtained using MDS algorithm	Transformation results
Cancer		
Seeds		
Iris		
Ecoli		
Satimage		
Vertebral		
Wholesale		

1) *Cancer* has relatively clear boundaries in a 2-D data space. However, the sizes of its two clusters are quite different. After transformation, clusters in *Cancer*have similar sizes.

2) Each cluster in the mapping result of *Seeds*is circularly distributed and overlapped with each other slightly.

The normalized clusters are well separated with several misclassified points.

3) The mapping result of *Iris*shows that the left cluster is linearly separable from the other two, which are partially overlapped. The boundary of different clusters is much clearer compared with the other datasets, which can help to calculate the volume and surface area occupied by the dataset precisely. Datasets *Vertebral*, *Wine*, and *Wholesale* has similar mapping results as *Iris*.

4) The clusters in the 2-D projection of *Ecoli*are a little overlapped. The number of data points in each cluster is different, and the four clusters with fewer objects do not have an obvious distribution structure, which are easy to be neglected.

5) The mapping clusters of *Satimage*are all nonspherical. Although the transformation rule can make these clusters spherical, there are some misclassified points due to the greatly overlapped clusters.

Table 4 shows the evaluation results of VAI and other five indices. CH suggests the real numbers of clusters for *Cancer* and *Wholesale*. XB can find the right evaluation result of *Cancer*. With regard to the other datasets, CH and XB give relatively small numbers, which are close to the real numbers. On the contrary, GS and PB give relatively larger numbers. DB suggests the right evaluation results for *Cancer* and *Vertebral*. It can be seen that VAI is capable of finding the hidden structures in datasets, and the corresponding evaluation results are the nearest to the real numbers of clusters.

TABLE 4. Evaluation results of eight real datasets.

Name	CH	DB	GS	PB	XB	VAI
Cancer	2√	2√	28	28	2√	1.99√
Seeds	2	30	29	23	2	3.17√
Iris	2	2	25	19	2	3.21√
Ecoli	2	29	29	30	3	5.12
Satimage	3	2	29	27	3	5.89√
Vertebral	2	3√	30	8	2	2.78√
Wholesale	2√	16	29	23	3	2.04√
Wine	2	12	28	26	2	2.96√

V. CONCLUSION

Finding the real number of clusters in a dataset is the first task in clustering analysis. The correct clustering results result from the correct identification of the number of clusters. In this paper, we map the original datasets into a 2-D data space firstly, and then transform the mapping clusters into spherical shapes with normalized sizes. Finally, we uncover the interrelation between hyper-volume and hyper-surface area of all clusters in a dataset, and originally propose a novel validity index. This index is unsupervised and independent of clustering algorithms and data distributions. Experimental results validate the accuracy of the novel index.

There are some opportunities for future research.

1) There are two possible ways to reduce the computational complexity of the proposed index. Firstly, using alternative algorithm to replace MDS but keeping its basic functions in our proposed method. Secondly, decomposing the MDS tasks into multiple units can greatly reduce the computational complexity, such as parallel algorithm, popular cloud computing, and so on.

2) The misclassified points in the overlapped area may affect the transformation process, which leads to inaccurate evaluation results. Therefore, how to identify the points in the overlapped area and rectify the deviation caused by the transformation process still remains as one of our research focuses in the future.

REFERENCES

[1] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering approach based on mini batch Kmeans for intrusion detection system over big data," *IEEE Access*, vol. 6, pp. 11897–11906, 2018.

[2] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, "Proposal-free network for instance-level object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2978–2991, Dec. 2018.

[3] M.-S. Yang, S.-J. Chang-Chien, and Y. Nataliani, "A fully-unsupervised possibilistic c-means clustering algorithm," *IEEE Access*, vol. 6, pp. 78308–78320, 2018.

[4] L. Hu and C. Zhong, "An internal validity index based on density-involved distance," *IEEE Access*, vol. 7, pp. 40038–40051, 2019.

[5] 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.

[6] 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.

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

[8] 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.

[9] M. A. Maniar and A. R. Abhyankar, "Validity index based improvisation in reproducibility of load profiling outcome," *IET Smart Grid*, vol. 2, no. 1, pp. 131–139, Mar. 2019.

[10] Z. Luo, L.-L. Zeng, J. Qin, C. Hou, H. Shen, and D. Hu, "Functional parcellation of human brain precuneus using density-based clustering," *Cerebral Cortex*, pp. 1–14, May 2019.

[11] M. A. Shah, G. Abbas, A. B. Dogar, and Z. Halim, "Scaling hierarchical clustering and energy aware routing for sensor networks," *Complex Adapt. Syst. Model.*, vol. 3, no. 1, p. 5, 2015.

[12] S. Salloum, J. Z. Huang, Y. He, and X. Chen, "An asymptotic ensemble learning framework for big data analysis," *IEEE Access*, vol. 7, pp. 3675–3693, 2019.

[13] X. Chen, W. Hong, F. Nie, D. He, M. Yang, and J. Z. Huang, "Spectral clustering of large-scale data by directly solving normalized cut," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1206–1215.

[14] J.-S. Wang and J.-C. Chiang, "An efficient data preprocessing procedure for support vector clustering," *J. UCS*, vol. 15, no. 4, pp. 705–721, 2009.

[15] F. Nielsen and R. Nock, "Optimal interval clustering: Application to Bregman clustering and statistical mixture learning," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1289–1292, Oct. 2014.

[16] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, "Weakly supervised scene parsing with point-based distance metric learning," *Assoc. Adv. Artif. Intell.*, vol. 33, pp. 8843–8850, Aug. 2019.

[17] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Oct. 2005.

[18] Z. Halim and J. H. Khattak, "Density-based clustering of big probabilistic graphs," *Evolving Syst.*, vol. 10, no. 3, pp. 333–350, Sep. 2019.

[19] Z. Halim, M. Waqas, A. R. Baig, and A. Rashid, "Efficient clustering of large uncertain graphs using neighborhood information," *Int. J. Approx. Reasoning*, vol. 90, pp. 274–291, Nov. 2017.
[20] Z. Halim, "Optimizing the minimum spanning tree-based extracted clusters using evolution strategy," *Cluster Comput.*, vol. 21, no. 1, pp. 377–391, Mar. 2018.
[21] S. A. L. Mary, A. Sivagami, and M. U. Rani, "Cluster validity measures dynamic clustering algorithms," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 9, pp. 4009–4012, May 2015.
[22] K. R. Žalik, "Validity index for clusters of different sizes and densities," *Pattern Recognit. Lett.*, vol. 32, no. 2, pp. 221–234, Jan. 2011.
[23] 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.
[24] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using k-means clustering algorithm and subtractive clustering algorithm," *Procedia Comput. Sci.*, vol. 54, pp. 764–771, Jun. 2015.
[25] B. Rezaee, "A cluster validity index for fuzzy clustering," *Fuzzy Sets Syst.*, vol. 161, no. 23, pp. 3014–3025, Dec. 2010.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat. Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[31] 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*, vol. 63, no. 2, pp. 411–423, May 2001.
[32] 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.
[33] X. 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.
[34] W.-L. Hung and M.-S. Yang, "Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance," *Pattern Recognit. Lett.*, vol. 25, no. 14, pp. 1603–1611, Oct. 2004.
[35] V. Berikov, "Weighted ensemble of algorithms for complex data clustering," *Pattern Recognit. Lett.*, vol. 38, pp. 99–106, Mar. 2014.
[36] V. Kumar, J. K. Chhabra, and D. Kumar, "Performance evaluation of line symmetry-based validity indices on clustering algorithms," *J. Intell. Syst.*, vol. 26, no. 3, pp. 483–503, 2017.
[37] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Stat. Probab.*, Jun. 1965, pp. 281–297.
[38] M. P. Windham, "Cluster validity for the fuzzy c-means clustering algorithrm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–4, no. 4, pp. 357–363, Jul. 1982.
[39] Q. Tong, X. Li, and B. Yuan, "A highly scalable clustering scheme using boundary information," *Pattern Recognit. Lett.*, vol. 89, pp. 1–7, Apr. 2017.
[40] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.
[41] M. Misuraca, M. Spano, and S. Balbi, "BMS: An improved Dunn index for Document Clustering validation," *Commun. Stat. Theory Methods*, vol. 48, no. 20, pp. 5036–5049, Oct. 2019.
[42] N. Saeed, H. Nam, M. I. U. Haq, and D. B. M. Saqib, "A survey on multidimensional scaling," *ACM Comput. Surv.*, vol. 51, no. 3, p. 47, 2018.
[43] P. Mair, "Multidimensional scaling," in *Modern Psychometrics With R*. Cham, Switzerland: Springer, 2018, pp. 257–287.
[44] Q. Tong, X. Li, and B. Yuan, "Efficient distributed clustering using boundary information," *Neurocomputing*, vol. 275, pp. 2355–2366, Jan. 2018.
[45] 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.
[46] G. Pai, R. Talmon, and R. Kimmel, "Parametric manifold learning via sparse multidimensional scaling," Mar. 2017, *arXiv:1711.06011*.
[47] J. Tenreiro Machado, A. Lopes, and A. Galhano, "Multidimensional scaling visualization using parametric similarity indices," *Entropy*, vol. 17, no. 4, pp. 1775–1794, Mar. 2015.
[48] J.-J. Huang, G.-H. Tzeng, and C.-S. Ong, "Multidimensional data in multidimensional scaling using the analytic network process," *Pattern Recognit. Lett.*, vol. 26, no. 6, pp. 755–767, May 2005.
[49] R. Gupta and R. Kapoor, "Comparison of graph-based methods for nonlinear dimensionality reduction," *Int. J. Signal Imag. Syst. Eng.*, vol. 5, no. 2, p. 101, 2012.
[50] J. Emert and R. Nelson, "Volume and surface area for polyhedra and polytopes," *Math. Mag.*, vol. 70, no. 5, pp. 365–371, Dec. 1997.
[51] W. Paulsen, "Gamma triads," *Ramanujan J*, vol. 50, no. 1, pp. 123–133, Oct. 2019.
[52] S. Bandaru, A. H. Ng, and K. Deb, "Data mining methods for knowledge discovery in multi-objective optimization: Part A–Survey," *Expert Syst. Appl.*, vol. 70, pp. 139–159, Mar. 2017.
[53] M. M. R. Khan, R. B. Arif, M. A. B. Siddique, and M. R. Oishe, "Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository," in *Proc. 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. (iCEEiCT)*, Sep. 2018, pp. 124–129.

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 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.

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.