RESEARCH ARTICLE

*Journal of Medical Imaging and
Health Informatics*
Vol. 10, 1–6, 2020

Determining the Correct Number of Clusters in the CT Image Segmentation

Qi Li, Shihong Yue*, Mingliang Ding, Jia Li, and Zeying Wang

School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China

Clustering algorithm plays an essential role in CT image segmentation, and cluster validity index is an essential component in clustering analysis. There are a lot of validity indices used for assessing clustering results, that is, determine the optimal cluster number. But the existing validity indices are often ineffective for the datasets with irregular-shaped clusters and corrupted by noise. This study aims to define a novel validity index which cannot be affected by the shapes of clusters and corrupted by noise of the investigated datasets. Chain-based distance different from original Euclidean distance is defined first, then by a multidimensional scaling (MDS) transformation, all points are mapped into a new data space. After evaluation of compactness and separation twice in datasets, a novel validity index is proposed. A lot of synthetic datasets and several typical CT images were used for validating the proposed validity index. Experimental results validate the proposed index and this index is applicable to the datasets with arbitrary-shaped clusters and corrupted by noise, which is helpful in clustering analysis and computer-aided detection system.

Keywords: Cluster Validity Index, Chain-Based Distance, Multidimensional Scaling, CT Image.

1. INTRODUCTION

Many researches focus on supervised learning, but clustering in them is the most important way. Clustering and its validity evaluation have played a very important role in various fields, such as image segmentation [1], data analysis [2], business applications [3], and so on. Especially, X-ray Computed Tomography (CT) is a widely-used imaging technique for the detection of lung diseases [4]. With the improvement of CT imaging technology, the thickness of the scanning layer is becoming smaller and smaller. A large number of CT images can be generated by CT scanning for each patient, and the size of CT scans depends on many kinds of factors such as the length of the part of the body scanned and the z resolution. Manual interpretation of CT images has greatly increased the workload of radiologists. Thus, Computer-aided Detection (CAD) systems [5] are proposed. The CAD systems can help the radiologists detect pulmonary nodules.

Suggesting the optimal cluster number in CT images is crucial, which is the basis of precise segmentation of tissues and organs [6]. Cluster validity indices [7] is capable of solving this problem. Various validity indices (CVI) have been proposed, such as Davies-Bouldin (DB) measure (DB) [8], Xie–Beni's (XB) separation measure [9], Pakhira and Bandyopadhyay' (PB) index [10], Calinski-Harabasz (CH) [11] index, and so on [12].

2. RELATED RESEARCH

Assume dataset $X=\{x_1, x_2, \ldots, x_n\}$ contains n points. Generally, a CVI is the function of compactness (ϕ_c) and separation (δ_c), i.e.,

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

This section illustrates five classical indices.

(1) *Calinski-Harabasz* (CH) *index* [11]. Denote n_i be the point number in C_i, z_i and z be the center of cluster i and global center, respectively. CH index is described as

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

(2) *Davies-Bouldin* (DB) *index* [8]. Assume δ_{ij} be the inter-cluster distance between C_i and C_j. DB is defined as

$$\mathrm{DB}(c) = \sum_{i=1}^{c} R_i/c, \quad \text{s.t.,} \quad \begin{cases} R_i = \max_{j,\, j\neq i}(\phi_i + \phi_j)/\delta_{ij} \\ \delta_{ij} = \|z_i - z_j\| \\ \phi_i = \sum_{x\in C_i}\|x - z_i\|/|C_i| \end{cases} \tag{3}$$

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

(3) *Tibshirani's gap statistic* (GS) *index* [13].

$$\mathrm{Gap}\,(c) = E^*\,[\log(W(c))] - \log(W(c)),$$

$$\text{s.t.,} \quad \begin{cases} W(c) = \sum_{i=1}^{c} D_i/(2|C_i|) \\ D_i = 2|C_i|\sum_{j\in C_i}\|x_j - \bar{x}\|, \quad \bar{x} = \sum_{i=1}^{|C_i|} x_i/|C_i| \end{cases} \tag{4}$$

[*] Author to whom correspondence should be addressed.

2156-7018/2020/10/001/006 doi:10.1166/jmihi.2020.3199

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

(4) *Pakhira and Bandyopadhyay'* (PB) *index* [10]. PB is designed as

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

(5) *Xie–Beni's separation* (XB) *index* [9]. The clustering results obtained from fuzzy clustering algorithms can be evaluated by XB index, i.e.,

$$\mathrm{XB}(c) = \left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^{m} ||x_j - z_i||^2\right) \Big/ \left(n \cdot \min_{i \neq j} ||z_j - z_i||^2\right) \tag{6}$$

where m is the fuzzy exponential and u_{ij} is a partitioning matrix, satisfying

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

(6) *SH index* [14]. The SH index determines the optimal cluster number by maximizing intra-cluster similarity and inter-cluster differences, i.e.,

$$\mathrm{SH}(c) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^2 ||x_j - v_i||^2 + 1/c \sum_{i=1}^{c} ||v_i - \bar{v}||^2}{\min_{i \neq k}(||v_i - v_k||^2)} \tag{7}$$

where $\bar{v} = (\sum_{j=1}^{n} x_j)/n$.

However, these indices above are all difficult to evaluate the datasets with arbitrary-shaped clusters and corrupted by noise.

3. METHODOLOGY

3.1. The Chain-Based Space

Let dist(x_i, x_j) be the *Euclidean* distance between x_i and x_j, i, $j = 1, 2, \ldots, n$. We use the sign KNN$_p$ (x_i) to denote the set of p-nearest neighbors of x_i.

Definition 3.1. *Density*. The density of any point x_i can be computed by the distances between x_i and its p-nearest neighbors.

$$\rho_i = \Big(\sum_{j \in KNN_p(x_i)} \mathrm{dist}(x_i, x_j)\Big)^{-1} \tag{8}$$

where the parameter p is generally taken as $[2d\pi]$, and $[\bullet]$ is an integerizing operator.

With regard to x_i, we define a new distance based on density, formulated as Eq. (9), as the density-based distance σ_i [15]. And φ_i is denoted as the nearest density-based neighbor of x_i.

$$\sigma_i = \min_{j:\rho_i < \rho_j} \mathrm{dist}(x_i, x_j) \tag{9}$$

$$\varphi_i = \arg\min_{j:\rho_i < \rho_j} \mathrm{dist}(x_i, x_j) \tag{10}$$

Definition 3.2. *Key points* (KP). Points with higher value of Eq. (11) are denoted as key points.

$$\gamma_i = \rho_i \sigma_i \tag{11}$$

Dataset X can be connected into a group of chains according to the following connecting rule. In terms of x_i, the next point x_j is φ_i. Repeating connecting until visiting a key point.

Definition 3.3. *Chain*. A chain in X starts with any point x_i and stops at γ_i based on the above connecting rule.

Definition 3.4. *Chain-based distance* (d_{chain}). Suppose Cha$_i$ is the i-th chain. Cha$_i$contains x_i, and Cha$_j$ contains x_j. The chain-based distance d_{chain} (x_i, x_j) between x_i and x_j can be defined as Eq. (12).

$$d_{\text{chain}}(x_i, x_j) = \begin{cases} 0, & \text{if} \quad \mathrm{Cha}_i = \mathrm{Cha}_j \\ \mathrm{dist}(\mathrm{KP}_i, \mathrm{KP}_j), & \text{if} \quad \mathrm{Cha}_i \neq \mathrm{Cha}_j \end{cases} \tag{12}$$

where KP$_i$ and KP$_j$ denote the key points of Cha$_i$ and Cha$_j$, respectively.

The chain-based distance satisfies three general conditions, which can be shown as follows.

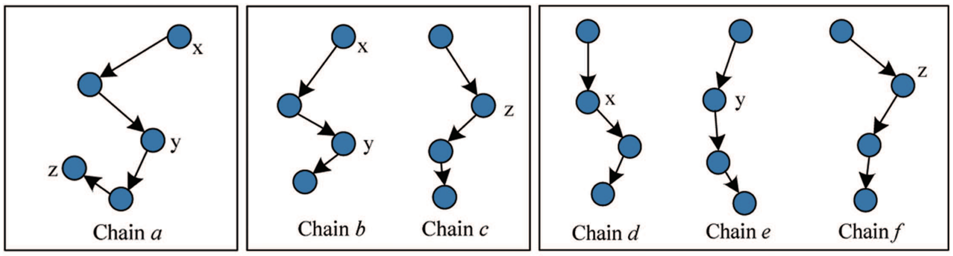
(1) $d_{\text{chain}}(x, y) \geq 0$;
(2) $d_{\text{chain}}(x, y) = d_{\text{chain}}(y, x)$;
(3) $d_{\text{chain}}(x, z) \leq d_{\text{chain}}(x, y) + d_{\text{chain}}(y, z)$;

The conditions (1) and (2) are obviously true. Condition (3) can be proven in three cases.

Case 1. Points x, y, and z are located on one same chain (see Fig. 1(a)). The distances among points x, y, and z are all equal to 0, which satisfies the triangle inequality.

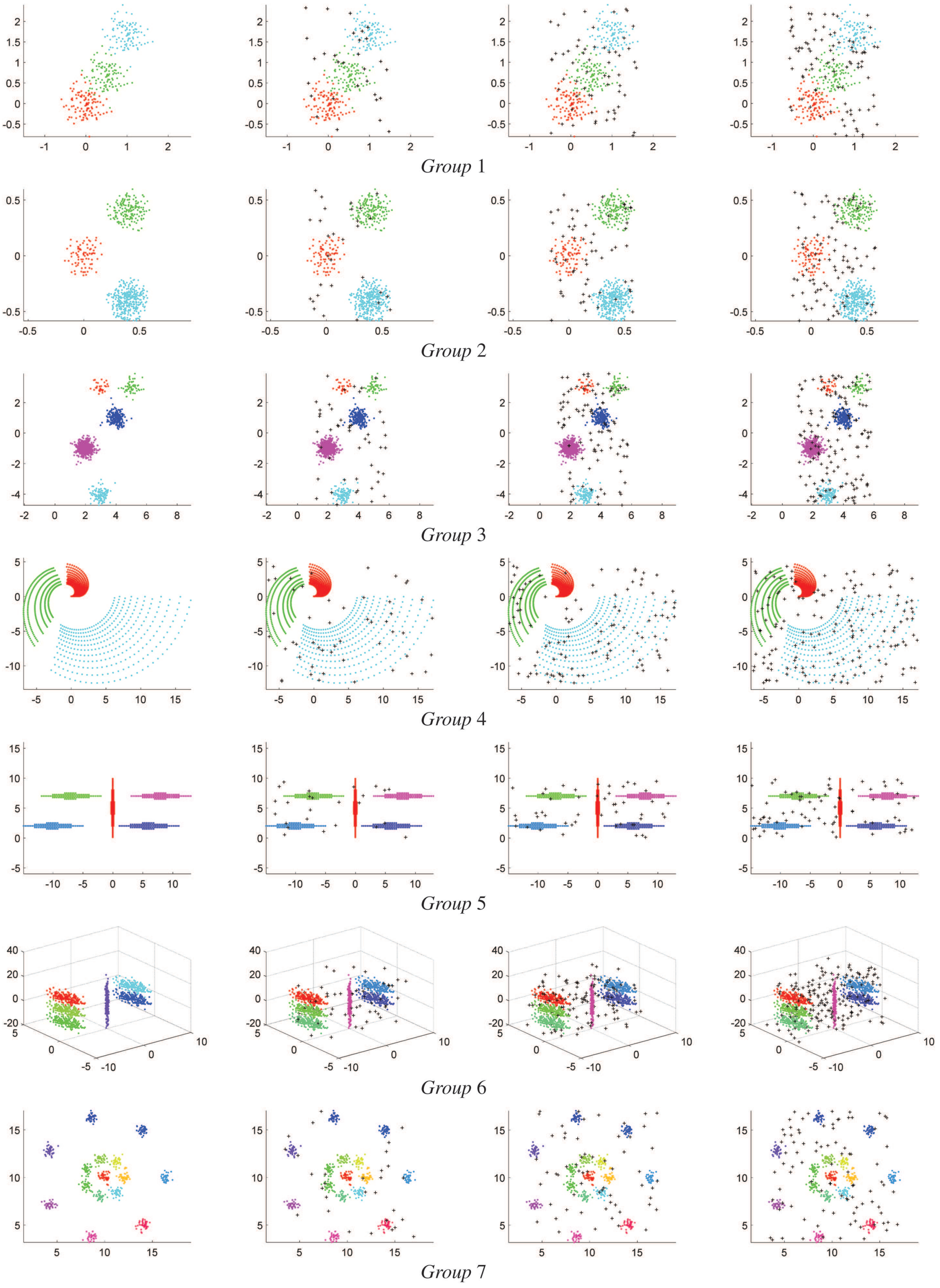
Case 2. Two points locate on the same chain while another on the other chain (see Fig. 1(b)). $d_{\text{chain}}(x, y) \leq d_{\text{chain}}(x, z) + d_{\text{chain}}(y, z)$ is true because $d_{\text{chain}}(x, y) = 0$. And $d_{\text{chain}}(x, z) = d_{\text{chain}}(y, z)$, which proves $d_{\text{chain}}(x, \mathrm{z}) \leq d_{\text{chain}}(x, y) + d_{\text{chain}}(y, z)$ and $d_{\text{chain}}(y, \mathrm{z}) \leq d_{\text{chain}}(x, y) + d_{\text{chain}}(x, z)$.

Case 3. Three points are located on different chains, respectively (see Fig. 1(c)). The distances among x, y, and z can be calculated by the *Euclidean* distances among the three key points of chains d, e, and f, which satisfies the triangle inequality.



(a) Case 1 (b) Case 2 (c) Case 3

Fig. 1. Distributions of three points on chains.



Group 1

Group 2

Group 3

Group 4

Group 5

Group 6

Group 7

Fig. 2. Seven groups of synthetic datasets.

Table I. Evaluation results of different indices for synthetic datasets.

Datasets	CH	DB	GS	PB	XB	CMI
Group 1	2	3	3	30	2	3
	2	3	3	26	2	3
	2	2	3	25	2	3
	2	2	3	26	2	3
Group 2	2	3	3	30	3	3
	2	3	3	25	3	3
	2	3	3	23	3	3
	2	3	3	30	3	3
Group 3	3	5	6	30	3	5
	2	5	10	17	3	5
	2	5	4	30	5	5
	2	3	4	28	3	5
Group 4	2	27	7	30	2	3
	2	5	4	29	2	3
	2	13	10	22	8	3
	2	26	9	21	2	3
Group 5	2	30	30	29	2	5
	2	20	28	28	2	5
	2	17	29	27	2	5
	2	16	29	28	2	5
Group 6	6	10	28	23	5	6
	2	6	30	25	9	6
	2	5	30	15	9	6
	2	13	11	29	9	6
Group 7	2	8	15	9	15	15
	2	8	16	8	15	15
	2	8	23	8	10	15
	2	8	25	8	10	15

DEFINITION 3.5. *Chain-coordinate*. The chain-coordinate of any point x_i is a representation of the coordinate of x_i in chain-based space. It can be computed by the MDS algorithm [16].

The distance matrix $M \in R^{n\times n}$ is formulated as

$$M_{ij} = ({d_{\text{chain}}}^2(x_1, x_j) + {d_{\text{chain}}}^2(x_i, x_1) - {d_{\text{chain}}}^2(x_i, x_j))/2 \quad (13)$$

The eigenvalue decomposition can decompose M into matrixes U and S, where U is a eigenvector matrix, and S is a diagonal matrix, i.e.,

$$M = \text{USU}^T \quad (14)$$

The chain-coordinates $Y \in R^{n\times n}$ is computed as

$$Y = U\sqrt{S} \quad (15)$$

In generally, we select the first two columns of Y as the mapping coordinates.

3.2. The Validity Index

Let Cha_1, Cha_2, … Cha_c be c disjoint chains of X. The diameter of each chain $\text{diam}(\text{Cha}_k)$ and the distances between chains $\text{dist}(\text{Cha}_m, \text{Cha}_n)$ can be defined as

$$\text{diam}(\text{Cha}_k) = \max_{x,y\in\text{Cha}_k} \text{dist}(x, y), \quad k = 1, \dots c \quad (16)$$

$$\text{dist}(\text{Cha}_m, \text{Cha}_n) = \min_{x\in\text{Cha}_m, y\in\text{Cha}_n} \text{dist}(x, y), \quad m, n = 1, \dots c, \quad m \neq n \quad (17)$$

The maximum value of Eq. (18) can be used for suggesting the optimal cluster number c_{op} in a dataset.

$$\text{CVI}_c = \frac{\min_{1\leq m\leq c} \min_{1\leq n\leq c, m\neq n} \text{dist}(\text{Cha}_m, \text{Cha}_n)}{\max_{1\leq k\leq c} \text{diam}(\text{Cha}_k) + \varepsilon} \quad (18)$$

$$c_{op} = \arg\max_{c} \text{CVI}(c) \quad (19)$$

where ε is a small positive number which guarantees that the denominator cannot be equal to 0.

Eq. (18) are illustrated as follows:

(1) If the number of KP is smaller or equal to the real one, the nonredundant chain-coordinates are located far away from each other, thus c_{op} is equal to the number of KP.

(2) If the number of KP is larger than the real one, at least one cluster is partitioned into several clusters. These chain-coordinates from the same cluster are located close to each other. In this case, c_{op} calculated by Eq. (19) is equal to the real cluster number.

Considering the variances of c_{op} can be computed by curvature radius mathematically, we define a novel index based on the chain-based space.

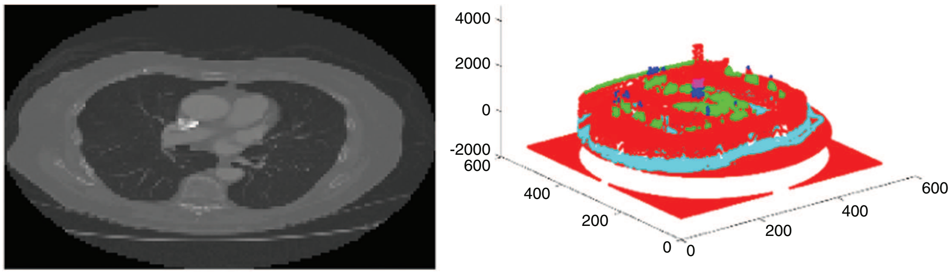
$$F(c_{\text{KP}}) = |\Delta_1(c_{\text{KP}})|^2/(1 + (\nabla_1(c_{\text{KP}}))^2)^{3/2}$$
$$\text{s.t.,} \quad \begin{cases} \Delta_1(c_{\text{KP}}) = c_{op}(c_{\text{KP}} + 1) + c_{op}(c_{\text{KP}} - 1) \\ \qquad - 2c_{op}(c_{\text{KP}}) \\ \nabla_1(c_{\text{KP}}) = c_{op}(c_{\text{KP}} + 1) - c_{op}(c_{\text{KP}}) \end{cases} \quad (20)$$

where symbol Δ denotes a two-order difference operator and c_{KP} denotes the number of key points.

Thus, the optimal cluster number c^* can be computed as

$$c^* = \arg\max_{c_{\text{KP}}} F(c_{\text{KP}}) \quad (21)$$

Hereafter, the index of Eq. (21) is called a CMI (Chain-Mapping Index).



(a) The original CT image.

(b) Representation in a 3-D space.

Fig. 3. The CT image and its representation in a 3-D data space.

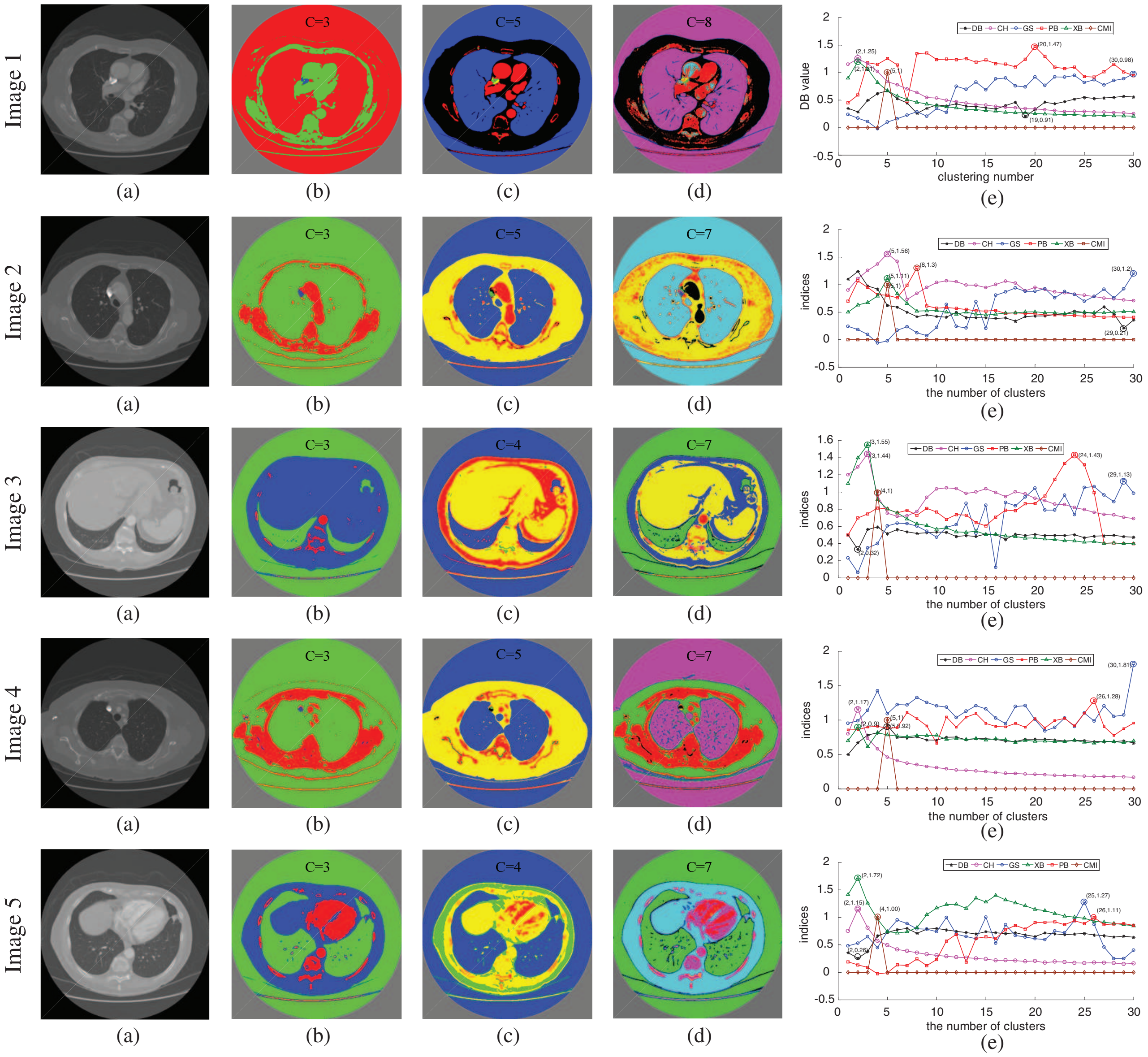


Fig. 4. Evaluation results of CMI index for five CT images.

4. RESULTS

4.1. Tests on Synthetic Datasets

Figure 2 illustrates the synthetic datasets with different characteristics. Columns 1–4 denote the original datasets and datasets with 10%, 20%, and 30% uniformly distributed noise, respectively. The marker '+' denotes noise.

Table I illustrates the evaluation results of six validity indices for datasets in Figure 2. The evaluation results shows that when datasets contain more noise, the performances of the five existing indices are easy to be affected, thus they cannot obtain the correct results; CMI index performs well for datasets corrupted by noise. In terms of datasets with arbitrary-shaped clusters, the performance of CMI also exceeds the five existing indices.

4.2. Tests on CT Images

Figure 3 shows a CT image and its representation in a 3-D space. Each image contains 512×512 pixels, and each pixel is identified by its CT value and location, indicating that the 3-D representation of the CT image is nonlinear.

Figure 4 shows the evaluation results of CMI index. Columns 2–4 show the chains under different numbers of key points. The fifth column denotes the evaluation results of CMI.

Table II shows the evaluation results, indicating that CMI index is capable of finding the correct cluster numbers towards all CT images shown in Figure 4.

Table II. The evaluation results of different indices for five CT images.

Datasets	CH	DB	GS	PB	XB	CMI
Image 1	2	19	30	20	2	5
Image 2	5	29	30	8	5	5
Image 3	3	2	29	24	3	4
Image 4	2	5	30	26	2	5
Image 5	2	2	25	26	2	4

Table III. The comparisons of the four methods to determine the number of neighbors for five CT images.

	Image 1	Image 2	Image 3	Image 4	Image 5
$\sqrt{n}$	512/4 (×)	512/3 (×)	512/4 (✓)	512/2 (×)	512/4 (✓)
$(1\sim 2\%)\times n$	2621∼5243/4 (×)	2621∼5243/5(✓)	2621∼5243/5(×)	2621∼5243/7(×)	2621∼5243/6(×)
Eq.(22)	17086/5 (✓)	17086/4(×)	17086/4 (✓)	17086/6(×)	17086/6(×)
$[2d\pi]$	19/5 (✓)	19/5 (✓)	19/4 (✓)	19/5(✓)	19/4 (✓)

Notes: Marks "✓" and "×" denote the correct and wrong number of clusters, respectively.

5. DISCUSSION

Although the number of neighbors for density computation has been taken as $[2d\pi]$ in Eq. (8), a general solution to determine an accurate number of neighbors remains unsolved so far. Currently, there are three noticeable methods used to choose the optimal number of neighbors for a dataset containing n points, such as $(1\sim 2\%)\times n$ [15], $\sqrt{n}$ [16] and the equation [17] as follows

$$\begin{aligned} k_{NS,1} &= v_0[4/(d+4)]^{d/(d+6)}n^{6/(d+6)} \\ \text{s.t.,}\quad v_0 &= \pi^{d/2}\Gamma((d+2)/d) \end{aligned} \tag{22}$$

where $k_{NS,1}$ denotes the optimal number of neighbors, and $\Gamma(\bullet)$ is the typical mathematical gamma function [18]. In order to compare effectiveness of these methods, the five tested CT images in Figure 4 are used for experiments. The experimental results are shown in Table III.

Table III shows that the proposed way in this paper is the most effective method to suggest the optimal number of neighbors for CT images. Nevertheless, it remains not breaking through the natural problem, i.e., how to generally determine the number of neighbors to compute density in any dataset.

6. CONCLUSION

In this study, a novel validity index is designed for datasets with arbitrary-shaped clusters and corrupted by noise. Firstly, the original dataset is mapped into a chain-based space. After evaluating the chain-coordinates twice, a novel index is proposed. Experimental results demonstrate higher accuracy of the proposed index than the existing typical validity indices for most tested datasets.

However, the proposed index is time-consuming. And, if different clusters overlap seriously in a dataset, the accuracy of the proposed index may be reduced. Thus, how to enhance the time resolution and correct the deviation caused by the overlapped clusters are our further concern.

References and Notes

1. Dhanachandra, N. and Chanu, Y.J., **2019**. A new image segmentation method using clustering and region merging techniques. *Applications of Artificial Intelligence Techniques in Engineering*, pp.603–614.
2. Mohebi, A., Aghabozorgi, S. and Ying, W.T., **2016**. Iterative big data clustering algorithms: A review. *Software: Practice and Experience, 46*(1), pp.107–129.
3. Shyamala, G. and Pooranam, N., **2016**. A Survey on Online Stock Forum Using Subspace Clustering. *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pp.1–6.
4. Yue, S., Wang, Y. and Wang, J., **2017**. Relationships between lung cancer incidences and air pollutants. *Technology and Health Care, 25*(S1), pp.411–422.
5. Zhu, H., Pak, C.H. and Song, C., **2017**. A novel lung cancer detection algorithm for CADs based on SSP and Level Set. *Technology and Health Care, 25*(S1), pp.345–355.
6. Ali, A.R., Couceiro, M.S. and Hassanien, A.E., **2016**. Fuzzy *C*-means based on minkowski distance for liver CT image segmentation. *Intelligent Decision Technologies, 10*(4), pp.393–406.
7. Hu, L. and Zhong, C., **2019**. An internal validity index based on density-involved distance. *IEEE Access*, 1.
8. Davies, D.L. and Bouldin, D.W., **1979**. A cluster separation measure. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 2*, pp.224–227.
9. Xie, X.L. and Beni, G., **1991**. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 8*, pp.841–847.
10. Pakhira, M.K., Bandyopadhyay, S. and Maulik, U., **2004**. Validity index for crisp and fuzzy clusters. *Pattern Recognition, 37*(3), pp.487–501.
11. Caliński, T. and Harabasz, J., **1974**. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods, 3*(1), pp.1–27.
12. Masud, M.A., Huang, J.Z. and Wei, C., **2018**. I-nice: A new approach for identifying the number of clusters and initial cluster centres. *Information Sciences, 466*, pp.129–151.
13. Tibshirani, R., Walther, G. and Hastie, T., **2001**. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), pp.411–423.
14. Kwon, S.H., **1998**. Cluster validity index for fuzzy clustering. *Electronics Letters, 34*(22), pp.2176–2177.
15. Rodriguez, A. and Laio, A., **2014**. Clustering by fast search and find of density peaks. *Science, 344*(6191), pp.1492–1496.
16. MacQueen, J., **1967**. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp.281–297.
17. Duong, T., Beck, G. and Azzag, H., **2016**. Nearest neighbour estimators of density derivatives, with application to mean shift clustering. *Pattern Recognition Letters, 80*, pp.224–230.
18. Paulsen, W., **2019**. Gamma triads. *The Ramanujan Journal, 4*, pp.1–11.

Received: xx Xxxx xxxx. Accepted: xx Xxxx xxxx.