Compactness and Separateness Driven Fuzzy Clustering Validity Index Called TLW

Yiming Tang[1,2](✉), Xiang Wang[1], Bing Li[1], Xianghui Hu[1], and Wenjun Xie[1]

[1] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine,
School of Computer and Information, Hefei University of Technology,
Hefei 230601, Anhui, China
tym608@163.com

[2] Engineering Research Center of Safety Critical Industry Measure and Control Technology,
Ministry of Education, Hefei University of Technology, Hefei 230601, Anhui, China

Abstract. The design of validity index of fuzzy clustering has always been a historical problem in fuzzy clustering field. When the distribution of cluster centers is very close, it is difficult for the existing fuzzy clustering validity indexes to obtain a reasonable cluster number, and the separation mechanism of these indexes is too simple. In order to solve the above problems, we propose a novel fuzzy clustering validity index called TLW (Tang-Li-Wang) index. Firstly, compactness is expressed as the ratio of the membership weighted distance value to the sample variance of the dataset. Secondly, the sum of the maximum distance between cluster centers and the mean distance is used in separateness, and the sample variance of cluster centers is introduced, and the two are multiplied to describe the separateness. Thirdly, on the basis of considering compactness and separateness, the introduction of cluster number can alleviate the phenomenon that the index value may change monotonically with the increase of cluster number. Finally, the classical FCM (Fuzzy C-Mean) algorithm is used to conduct experiments on indexes. Comparative experiments and analyses were carried out on 17 typical datasets and 12 clustering validity indexes. From the experimental results of normal simple datasets and high-dimensional difficult datasets, the proposed index shows some advantages. All in all, these results verify that the proposed TLW index has better accuracy and stronger stability.

Keywords: Fuzzy clustering · clustering validity index · compactness · separateness · fuzzy c-means algorithm

1 Introduction

Clustering is an unsupervised machine learning method [1–4], which has been widely applied and studied in various fields. Cluster verification is an important method to evaluate the quality of clustering [5, 6] and plays a vital role in clustering analysis.

Y. Sun et al. (Eds.): ChineseCSCW 2023, CCIS 2013, pp. 177–190, 2024.
https://doi.org/10.1007/978-981-99-9640-7_13

Clustering validation is a quantitative evaluation of clustering results, which is called clustering validity index (CVI) [7]. The function of CVI is to judge the optimal number of clusters and the optimal partition result. For the results divided by clustering algorithm, we calculate the corresponding index, and analyze the quality of clustering results according to the CVI.

Nowadays, many CVIs [8] have been proposed for various clustering algorithms, but after a lot of research, it is found that there is no effective index for all types of datasets. CVIs mentioned here are mainly internal validity indexes, most of which are based on the fuzzy c-means (FCM) algorithm. For the optimal number of clusters determined by FCM, the Calinski-Harabasz index (CH) [9], partition coefficient (PC) [10], Dunn index (Dunn) [11], standard separation coefficient (NPC) [12], Fuyama-Sugeno index (FSI) [13], Xie-Beni index (XBI) [14], Davis-Bouldin index (DBI) [15], WLI index [16], IMI index [17], and Mittal Saraswat index (SMI) [18] were proposed. The VCVI index [19] was based on the initial center selection method of the density parameter, rather than randomly selecting the initial center. Secondly, Maulik and Bandyopadhyay [20] proposed MB index after comparing the hard clustering algorithm (K-means), the single link algorithm and the simulated annealing algorithm. Some complex CVIs considered the distance between the data objects and the cluster centers, or the distance between the cluster centers, and calculated the average of these distances, such as the CH index and the FSI index, and used them as key components of the CVI formula. However, using these factors as key components leaded to CVIs that only produced better results for specific datasets. There were also some indexes that referred to some combinations of the maximum, median, average and minimum values of the above distances, such as the WLI and IMI indexes. It improved the robustness of the index, but the computation is more complex.

The existing indexes still have problems that cannot be ignored [21]:

1) The basic principle of clustering is that data objects with homogeneous characteristics are grouped into the same cluster, while data objects in different clusters are heterogeneous. This is usually achieved by considering compactness and separateness between clusters. Compactness measures the concentration of data objects within a cluster. The distance between objects with homogeneous features should be relatively small. However, this processing strategy goes wrong when dealing with intensive datasets.
2) The separateness is used to measure the degree of separation between clusters, so that the data with heterogeneous characteristics should be as different as possible. This can be evaluated by calculating the distance between each pair of cluster centers, or by the distance between two heterogeneous objects from two different clusters. The greater the distance, the better the separation effect of clusters. Most of the existing indexes express the degree of separation too simply, leading to inadequate characterization of the dataset. Hence we need a more comprehensive CVI.

Therefore, in this study, a new clustering validity index called the Tang-Li-Wang (TLW) index is proposed based on the above factors.

2 The TLW Index for Fuzzy Clustering

2.1 Compactness Analysis

The TLW index consists of three factors. The first part is the number of clusters K, which can alleviate the phenomenon that the value of the index may change monotonically with the increase of K. In addition, separateness and compactness are described in the other two factors of the index.

The first factor is K, as shown in (1):

$$C_1 = K. \tag{1}$$

In TLW index, we retain the advantages of the WLI index, but also increase the sample variance of the dataset. Therefore, we can have a macro grasp from the compactness of individual clusters to the compactness of the whole dataset. We add new elements to the compactness of TLW index, and also consider the possible drawbacks of traditional indexes in the past. The TLW index reduces the complexity of E_k and effectively reduces the operation time. Finally, the TLW index can get satisfactory results when dealing with complex datasets of different data distributions.

To sum up, we give a new fuzzy compactness expression, as follows:

$$C_2 = \frac{E_k}{E_1}. \tag{2}$$

$$E_k = \sum_{k=1}^{K} \sum\nolimits_{i=1}^{N} \mu_{ik}^{m} ||x_i - v_k||^2. \tag{3}$$

Here K is the number of data clustering categories. N is the sum of the samples of the dataset. m is the index of membership matrix and represents the fuzzy weighting index. x_i is a data object in all datasets and its subscript represents the i-th data object. v_k is a cluster center and the subscript represents the k-th cluster center. u_{ik} represents data membership. $\sum_{i=1}^{N} \mu_{ik}^{m} ||x_i - v_k||^2$ is called the sum of squared errors within the class.

$$E_1 = \frac{\sum_{i=1}^{N} ||x_i - v_0||^2}{N}. \tag{4}$$

$$v_0 = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{5}$$

2.2 Separateness Analysis

The TLW index uses the maximum distance of cluster centers, the mean distance of cluster centers and the sample variance of cluster centers to represent the separateness. Therefore, the TLW index can better control the separateness of each pair of cluster centers and the separateness of the whole cluster centers, and can achieve almost unbiased evaluation. Adding the mean distance of cluster centers to the separateness can make the index have a better result in evaluating the data cluster with dense cluster centers. The third factor D_k, which measures the maximum degree of separation between the two clusters in all possible cluster pairs, will increase as the value of K increases. This value is the upper bound on the maximum separation between two points in the dataset. This ensures that we do not over-partition data that belongs to a cluster. The experiment also proves that the combination of the minimum distance of cluster centers and the mean distance of cluster centers is worse than the combination of the maximum distance of cluster centers and the mean distance of cluster centers.

Secondly, the index adds the sample variance of the cluster centers. The new separateness measure is obtained as follows:

$$\begin{aligned} C_3 = {} & (\mathrm{max}_{\mathrm{i}\neq \mathrm{j}}||\mathrm{v}_\mathrm{i} - \mathrm{v}_\mathrm{j}||^2 + \mathrm{mean}_{\mathrm{i}\neq \mathrm{j}}||\mathrm{v}_\mathrm{i} - \mathrm{v}_\mathrm{j}||^2) \\ & \times \frac{1}{K-1}\sum\nolimits_{k=1}^{K} ||v_k - \bar{v}\,||^2. \end{aligned} \tag{6}$$

According to the above formula, it can be seen that the third factor of the index considers the maximum distance between cluster centers and the mean distance between cluster centers. It makes TLW index have better results in the evaluation of the dataset with intensive structure, and avoids excessive division of clusters that originally belong to one class.

2.3 Function Expressions

Based on the above, we can get the TLW index, and its formula is as follows:

$$TLW^{(-)} = C_1 \times C_2 \times C_3 \tag{7}$$

The TLW index consists of three factors. The first factor is used to avoid the monotonicity of the index as the number of clusters increases. The second factor is a characterization of compactness and the third factor is the characterization of separateness. The third one restricts and balances each other. Meanwhile, the smaller the index, the better the clustering result.

The evaluation of an CVI is to see whether it can adapt to more kinds of datasets, and try to exclude the influence of clustering algorithms. The TLW index adds the variance of the whole sample dataset and the sample variance of the cluster centers for compactness and separateness, respectively. This increases the grasp of the overall separateness of the data and can better adapt to different structured datasets. At the same time, the TLW index not only considers the compactness and separateness of the whole dataset, but also considers the compactness and separateness of each cluster. The evaluation of the effect of clustering division is more refined.

The following is the calculation process of Table 1, which is the clustering validity index TLW.

Next, we describe the reasons why the TLW index is better than previous indexes. The composition of the index is divided into three factors. The first factor is the number of clusters, which can alleviate the monotonicity of the index caused by the increase of the number of clusters. The second factor is compactness, which consists of the ratio of the weighted distance of cluster membership to the variance of the sample points in the datasets. The advantage of this is that the index can handle datasets with different structures well, and the overall separateness of the dataset can also be included in the consideration of index compactness. The third factor is separateness, which is the sum of the maximum distance of the cluster centers and mean distance of the cluster centers, then multiplied by the sample variance of the cluster centers. The mean distance allows the TLW index to handle intensive datasets and the maximum distance prevents the index from over-partitioning when dealing with datasets that already belong to the same cluster. The sample variance of cluster centers refers to the deviation degree of cluster centers. Therefore, when classifying clusters, the TLW index not only considers the overall separation degree of cluster centers, but also involves the relationship between each pair of cluster centers. Finally, the TLW index not only considers the compactness and separateness of the whole dataset, but also the compactness and separateness of each cluster. All these make the evaluation of the TLW index on the effect of data agglomeration analysis more refined.

Table 1. The calculation process of TLW.

```
Algorithm 1. The calculation process of TLW.
Input: Input hyperparameter: Maximum iterations Iter, threshold for stopping iteration $\varepsilon$, membership matrix $U=[u_{ij}]$, the minimum number of clusters $K_{\min}$ and the maximum number of clusters $K_{\max}$.
Output: The index value of TLW under different cluster number.
procedure
    P1: Set Iter, $\varepsilon$ and $K_{\max}$; initializes the membership matrix $U=[u_{ij}]$ (satisfy $\sum_{i=1}^{c} u_{ij}=1$); let the number of initial iterations be 0, $K_{\min}=2$, $K=K_{\min}$, $m=2$;
    P2: Update membership matrix U;
    P3: Update cluster center V;
    P4: $k=k+1$;
    P5: If $\| V^{(k+1)} - V^{(k)} \| \geq \varepsilon$ and $k < Iter$, return Step2; else, continue;
    P6: Calculate $C_2$ using (2);
    P7: Compute $C_3$ using (6);
    P8: Figure out the value of TLW using (7);
    P9: $K=K+1$;
    P10: If $K \leq K_{\max}$, return P2; else, continue;
    P11: Find the minimum value of TLW(K) and the corresponding number of clusters K (optimization fraction);
end procedure
```

3 Experiments and Analysis

We adopted the FCM algorithm for verification. We used Intel(R) Core(TM)i7-8700 CPU @ 3.20 GHz 3.19 GHz, and Windows 10OS. Employed programming software was MATLAB 2018b.

A total of 17 datasets are used in the comparative experiment. Firstly, we run the FCM algorithm with different K values, and get the clustering results on some datasets. Then we calculate the value of the index in each round to find the best value. When the index value obtains the optimal result, the corresponding cluster number is the obtained cluster number. In this experiment, 12 indexes are used for comparative experiment. The specific experimental results and analyses are as follows.

3.1 Datasets and Comparison Indexes

Three types of datasets are used in this experiment, namely the UCI datasets [17], the artificial datasets and the Olivetti face dataset.

Eight UCI datasets are used in the experiment, including six normal datasets and two high-dimensional datasets. Normal UCI datasets include SPECTF-heart, Monk, Hayes-Roth, Seeds, Glass and Zoo. The SPECTF-heart dataset collects diagnostic information from patients' heart scan (SPECT) images, which are divided into two categories: normal and abnormal. The SPECTF-heart dataset contains 267 patients' diagnostic information samples with 44 dimensions, and the number of clusters is 2. The Monk dataset has 432 data samples with a 6-dimensional data dimension, and the number of clusters is 2. The Hayes-Roth dataset has 132 data samples with a 5-dimensional data dimension, and the number of clusters is 3. The Seeds dataset has 210 samples with a 7-dimensional data dimension, and the number of clusters is 3. The Glass dataset has 214 data samples with 9 dimensions, and the number of clusters is 6. The Zoo dataset has 101 data samples with a 16-dimensional data dimension, and the number of clusters is 7. The other two high-dimensional datasets are Libras and Letter. The Libras dataset has 360 data samples with a 90-dimensional data dimension, and the number of clusters is 15. The Letter dataset has 20000 data samples with 16 dimensions, and the number of clusters is 26.

While real datasets can make our conclusions more convincing, real datasets are too homogeneous. In order to avoid this problem, we should select other types of datasets to enhance the strength of experimental persuasion. Eight artificial datasets are used in the experiment, including six normal datasets and two high-dimensional datasets. Normal artificial datasets include Data_60, Data_11, E6, X8D5K, Fc1 and Sn. The Data_60 dataset has 60 data samples with a 2-dimensional data dimension, and the number of clusters is 3. The Data_11 dataset has 150 data samples with a 2-dimensional structure, and the number of clusters is 3. The E6 dataset has 8537 data samples with 2 dimensions, and the number of clusters is 4. The X8D5K dataset has 1000 data samples with a 8-dimensional data dimension, and the number of clusters is 5. The Fc1 dataset has 1035 data samples with 2 dimensions, and the number of clusters is 5. The Sn dataset has 513 data samples with 2 dimensions, and the number of clusters is 5. The other two high-dimensional datasets are Dim_128 and Dim_256. The Dim_128 dataset has 1024 data samples with 128 dimensions, and the number of clusters is 16. The Dim_256 dataset has

1024 data samples with a 256-dimensional data dimension, and the number of clusters is 16.

In addition, 12 indexes are selected for comparative experiment. Some indexes have the best clustering result with maximum values, such as CH, PC, Dunn, NPC and MB. While others have the best clustering result with minimum values, such as FS, XBI, DB, WLI, VCVI, IMI and SMI.

3.2 Experiments

In our experiments, the optimal number of clusters in some datasets is less than 10. We repeat the datasets for 10 rounds, with the number of clusters in each round ranging from 2 to 10. For other datasets, the optimal cluster number is greater than 10. These datasets are run for 30 rounds, with the number of clusters per round ranging from 2 to 30. At the end of each round, each index will have a maximum or minimum value, and then the optimal index value of each round is calculated. In this case, the number of clusters corresponding to the most optimal index values is the number of optimal clusters we seek. We make use of CVI^{+} to denote a larger-the-better index, and CVI^{-} to stand for a smaller-the-better one. The experimental results on the high-dimensional and normal datasets of UCI datasets are shown in Table 2.

Table 2. Results of experiments on the UCI datasets.

	Spectf-heart	Monk	Hayes-Roth	Seeds	Glass	Zoo	Letter	Libras
CH^{+}	$2^{3}3^{6}4^{1}$	10^{10}	2^{10}	2^{10}	2^{10}	2^{10}	$2^{18}3^{11}9^{1}$	30^{30}
PC^{+}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{30}	2^{30}
Dunn^{+}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	$2^{2}3^{8}$	$2^{20}3^{5}4^{5}$	2^{30}
NPC^{+}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{30}	2^{30}
FSI^{-}	10^{10}	2^{10}	10^{10}	3^{10}	4^{10}	10^{10}	30^{30}	30^{30}
XBI^{-}	$2^{7}3^{3}$	$8^{2}9^{8}$	3^{10}	3^{10}	$2^{4}6^{6}$	$2^{9}8^{1}$	$2^{6}24^{5}26^{19}$	$5^{1}14^{2}15^{9}19^{12}28^{4}30^{2}$
DB^{-}	$2^{4}3^{4}4^{2}$	$2^{7}3^{3}$	2^{10}	2^{10}	2^{10}	4^{10}	$2^{12}3^{9}4^{6}5^{3}$	2^{30}
WLI^{-}	$2^{7}3^{3}$	$8^{2}9^{8}$	4^{10}	3^{10}	$3^{8}6^{2}$	$3^{4}7^{6}$	$2^{4}24^{2}26^{22}28^{2}$	$3^{3}15^{18}27^{4}20^{2}$
VCVI^{-}	$5^{2}6^{5}8^{3}$	2^{10}	$3^{6}4^{4}$	3^{10}	$2^{4}6^{6}$	2^{10}	$2^{6}7^{4}9^{5}14^{7}18^{8}$	$7^{6}9^{4}10^{5}11^{4}15^{7}18^{4}$
IMI^{-}	$2^{5}4^{5}$	$8^{8}9^{2}$	$2^{2}3^{7}4^{1}$	3^{10}	$2^{9}6^{1}$	$2^{4}3^{6}$	$8^{5}23^{13}26^{5}30^{7}$	$4^{5}15^{15}17^{5}18^{3}20^{1}21^{1}$
MB^{+}	$2^{9}9^{1}$	$2^{7}9^{3}$	$3^{9}10^{1}$	3^{10}	$6^{5}10^{5}$	$7^{6}10^{4}$	$26^{19}30^{11}$	$15^{21}29^{9}$
SMI^{-}	2^{10}	2^{10}	$3^{8}9^{2}$	3^{10}	2^{10}	$2^{2}3^{8}$	$2^{16}7^{7}8^{7}$	2^{30}
TLW^{-}	2^{10}	2^{10}	$3^{9}4^{1}$	3^{10}	$6^{8}7^{2}$	$6^{3}7^{7}$	$20^{11}26^{19}$	$11^{5}15^{25}$

In Table 2, for the Zoo dataset, the optimal cluster number of CH after 10 rounds is 2, so the optimal cluster number determined by CH is 2. In fact, the result is wrong. The WLI index produced two results after 10 rounds, which were 3 and 7. Among them, 3 appeared 4 times, and 7 appeared 6 times, so the index finally determined that the best cluster number was 7. It can be seen that the WLI index get the correct result. The meanings of other indexes in the table are the same as those described above.

K is taken as the optimal cluster number when the index obtains the most times, as shown in Table 3, in which * indicates that the obtained result is inconsistent with the correct result.

Table 3. The optimal value of the index on the UCI datasets.

	Spectf-heart	Monk	Hayes-Roth	Seeds	Glass	Zoo	Letter	Libras
Best	2	2	3	3	6	7	26	15
CH^{+}	3*	10*	2*	2*	2*	2*	2*	30*
PC^{+}	2	2	2*	2*	2*	2*	2*	2*
$Dunn^{+}$	2	2	2*	2*	2*	3*	2*	2*
NPC^{+}	2	2	2*	2*	2*	2*	2*	2*
FSI^{-}	10*	2	10*	3	4*	10*	30*	30*
XBI^{-}	2	9*	3*	3	6	2*	26	19*
DB^{-}	2,3*	2	2*	2	2*	4*	2*	2*
WLI^{-}	2	9*	4*	3	3*	7	26	15
$VCVI^{-}$	5*	2	3	3	6	2*	18*	15
IMI^{-}	2,4*	8*	3	3	2*	3*	23*	15
MB^{+}	2	2	3	3	6,10*	7	26	15
SMI^{-}	2	2	3	3	2*	3*	2*	2*
TLW^{-}	2	2	3	3	6	7	26	15

In the Letter dataset, there are 20000 16-dimensional samples, and the correct number of categories is 26. In Table 2, we can see the results of each index on the Letter dataset. The optimal cluster number of CH obtained for 18 times is 2, and the optimal one for 11 times is 3, and the optimal one for 1 time is 9. The optimal cluster numbers obtained by the PC and NPC after 30 rounds of operation are all 2. The optimal cluster number of Dunn obtained for 20 times is 2, and the optimal one for 5 times is 3, and the optimal one for 5 time is 4. The optimal cluster numbers obtained by FSI after 30 rounds of operation are all 30. The optimal cluster number of XBI obtained for 19 time is 26, and the optimal one for 5 times is 24, and the optimal one for 6 times is 2. The optimal cluster number of DB obtained for 12 times is 2, and the optimal one for 9 times is 3, and the other outcomes are distributed in 4 and 5 classes. The optimal cluster number of WLI obtained for 22 times is 26, and the optimal one for 4 times is 2, and the other outcomes are distributed in 24 and 28 classes. The optimal cluster number of VCVI obtained for 8 times is 18, and the optimal one for 7 times is 14, and the other outcomes are distributed in 2, 7 and 9 classes. The optimal cluster number of IMI obtained for 13 times is 23, and the optimal one for 7 times is 30, and the other outcomes are distributed in 8 and 26 classes. The optimal cluster number of MB obtained for 19 times is 26, and the optimal one for 11 times is 30. The optimal cluster number of SMI obtained for 16 times is 2, and the optimal one for 7 times is 7, and the optimal one for 7 time is 8. The optimal cluster number of TLW index obtained for 19 times is 26, and the optimal one for 11 times is 20. We find that the result of TLW is correct, and the occurrence of correct optimal cluster number is the most frequent and relatively stable. Other indexes are inferior to TLW in terms of occurrence of correct cluster number and stability.

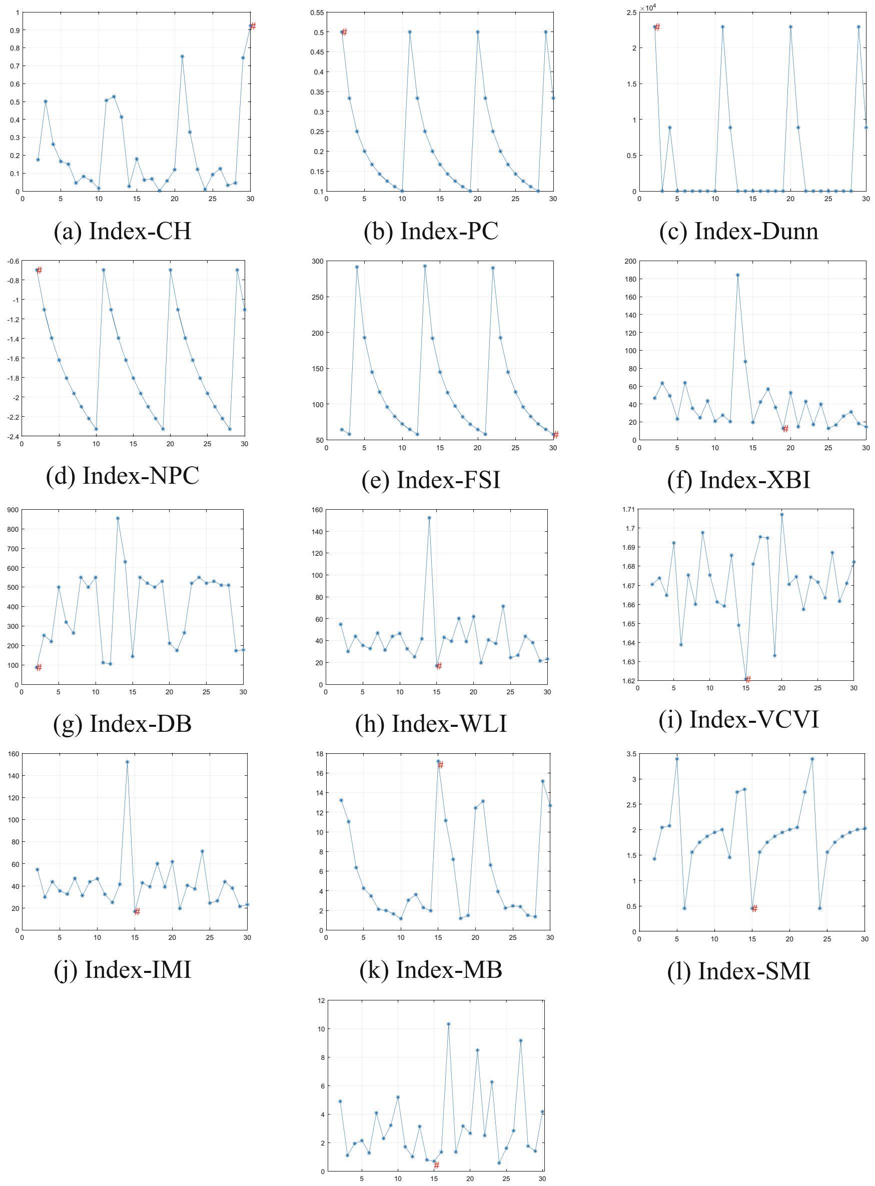
The Best row in the table is the correct number of clusters for the datasets. Taking the Libras dataset as an example, WLI, VCVI, IMI, MB and TLW all obtain correct clustering number. However, CH, PC, Dunn, NPC, FSI, XBI, DB, and SMI do not get clustering number. The changes of the results obtained from K of each index from 2 to 10 on the Libras dataset are shown in Fig. 1. Red marks the optimal number of clusters for each index.

Figure 1 is a line chart of each index function, representing the different index function values corresponding to the change of cluster number K from 2 to 30. When each index in the figure obtains the optimal value, the corresponding number of clusters is not exactly the same. And the convergence direction of each index algorithm is not exactly the same when obtaining the optimal result. In the Fig. 1, we can see that the CH, PC, Dunn, NPC and MB indexes are all the best results when the value of the index reaches the maximum value, while the remaining indexes are the best clustering results when the value of the minimum index. By observing the function value graphs of 13 indexes and their performance on UCI datasets, the newly proposed TLW index has high accuracy and stability.

Table 4 shows the results of the indexes on artificial datasets. We can see the results of each index on the Dim_128 dataset. The optimal cluster number of CH obtained for 25 times is 2, and the optimal one for 3 times is 4, and the optimal one for 2 time is 6. The optimal cluster numbers obtained by the PC, Dunn and NPC indexes after 30 rounds of operation are all 2. The optimal cluster numbers obtained by FSI after 30 rounds of operation are all 30. The optimal cluster number of XBI index obtained for 18 times is 16, and the optimal one for 7 time is 30, and the optimal one for 5 times is 25. The optimal cluster number of DB obtained for 20 times is 2, and the optimal one for 5 time is 4, and the optimal one for 5 times is 7. The optimal cluster number of WLI obtained for 18 times is 16, and the one for 8 time is 15, and the one for 4 times is 24. The optimal cluster number of VCVI obtained for 28 times is 2, and the optimal one for 2 time is 3. The optimal cluster number of IMI obtained for 12 times is 16, and the optimal one for 10 time is 24, and the optimal one for 8 times is 15. The optimal cluster number of MB obtained for 22 times is 16, and the optimal one for 8 time is 9. The optimal one of SMI obtained for 14 times is 6, and the optimal one for 10 time is 2, and the optimal one for 6 times is 8. The optimal one of TLW obtained for 22 times is 16, and the optimal one for 8 time is 13.

We select the number of clusters with the most optimal values as the final result. The performance of all indexes is summarized in Table 5. K* indicates a case where the evaluation is incorrect.

In Table 5, the FS index gets the correct number of clusters in all normal artificial datasets, but not in high-dimensional artificial datasets. Another index that performed well is the XBI index, which gets correct results in all seven datasets. And our proposed TLW index gets correct results in all datasets. The remaining indexes, such as CH and Dunn, have many errors and even fail to evaluate a dataset correctly.



(a) Index-CH

(b) Index-PC

(c) Index-Dunn

(d) Index-NPC

(e) Index-FSI

(f) Index-XBI

(g) Index-DB

(h) Index-WLI

(i) Index-VCVI

(j) Index-IMI

(k) Index-MB

(l) Index-SMI

(m) Index-TLW

Fig. 1. Comparison of results for each index on the Libras dataset. (Color figure online)

Table 6 shows the clustering results of all indexes on the Olivetti face dataset. The Olivetti face dataset has a total of 40 sets of photos, and each set contains 10 images. These 10 images are from the same person's face information, which is the same person's different expressions and images.

Table 4. Results of experiments on the artificial datasets.

	Data_60	Data_11	E6	Fc1	X8D5K	Sn	Dim_128	Dim_256
CH$^+$	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	$2^{25}4^{3}6^{2}$	$2^{24}3^{5}5^{1}$
PC$^+$	3^{10}	2^{10}	4^{10}	3^{10}	5^{10}	5^{10}	2^{30}	2^{30}
Dunn$^+$	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{10}	2^{30}	$2^{18}4^{5}6^{5}8^{2}$
NPC$^+$	3^{10}	2^{10}	4^{10}	3^{10}	5^{10}	5^{10}	2^{30}	2^{30}
FSI$^-$	$3^{7}7^{3}$	3^{10}	4^{10}	$4^{1}5^{7}6^{2}$	5^{10}	$5^{6}6^{4}$	30^{30}	30^{30}
XBI$^-$	3^{10}	3^{10}	4^{10}	3^{10}	5^{10}	5^{10}	$16^{18}25^{5}30^{7}$	$16^{19}25^{3}30^{8}$
DB$^-$	2^{10}	3^{10}	4^{10}	3^{10}	5^{10}	5^{10}	$2^{20}4^{5}7^{5}$	$2^{16}4^{5}6^{5}7^{2}9^{2}$
WLI$^-$	2^{10}	2^{10}	$3^{4}4^{6}$	3^{10}	$4^{2}5^{8}$	3^{10}	$15^{8}16^{18}24^{4}$	$9^{2}16^{18}29^{5}30^{5}$
VCVI$^-$	2^{10}	2^{10}	$2^{2}4^{8}$	2^{10}	$2^{3}5^{7}$	2^{10}	$2^{28}3^{2}$	2^{30}
IMI$^-$	2^{10}	3^{10}	$2^{7}4^{3}$	$2^{2}3^{3}5^{5}$	$3^{2}5^{8}$	$3^{6}5^{4}$	$15^{8}16^{12}24^{10}$	$15^{11}16^{6}29^{8}30^{5}$
MB$^+$	2^{10}	3^{10}	$2^{9}4^{1}$	$5^{6}8^{4}$	$5^{8}8^{2}$	$5^{7}8^{3}$	$9^{8}16^{22}$	$9^{10}16^{20}$
SMI$^-$	2^{10}	$3^{8}9^{1}10^{1}$	$4^{6}10^{4}$	$5^{7}8^{3}$	$2^{3}5^{7}$	$3^{3}5^{7}$	$2^{10}8^{6}16^{14}$	$2^{14}9^{5}16^{11}$
TLW$^-$	3^{10}	3^{10}	4^{10}	$5^{8}9^{2}$	5^{10}	5^{10}	$13^{8}16^{22}$	$9^{7}12^{2}16^{21}$

Table 5. The optimal value of each index on the artificial datasets.

	Data_60	Data_11	E6	Fc1	X8D5K	Sn	Dim_128	Dim_256
Best	3	3	4	5	5	5	16	16
CH$^+$	2*	2*	2*	2*	2*	2*	2*	2*
PC$^+$	3	2*	4	3*	5	5	2*	2*
Dunn$^+$	2*	2*	2*	2*	2*	2*	2*	2*
NPC$^+$	3	2*	4	3*	5	5	2*	2*
FSI$^-$	3	3	4	5	5	5	30*	30*
XBI$^-$	3	3	4	3*	5	5	16	16
DB$^-$	2*	3	4	3*	5	5	2*	2*
WLI$^-$	2*	2*	4	3*	5	3*	16	16
VCVI$^-$	2*	2*	4	2*	5	2*	2*	2*
IMI$^-$	2*	3	2*	5	5	3*	16	15*
MB$^+$	2*	3	2*	5	5	5	16	16
SMI$^-$	2*	3	4	5	5	5	16	2*
TLW$^-$	3	3	4	5	5	5	16	16

Table 6. Results of experiments on the Olivetti face data set.

	Olivetti face
Best	10
CH^{+}	30^{30}
PC^{+}	2^{30}
Dunn^{+}	2^{30}
NPC^{+}	2^{30}
FSI^{-}	30^{30}
XBI^{-}	$10^{20}30^{10}$
DB^{-}	$2^{13}3^{6}4^{8}12^{3}$
WLI^{-}	$10^{20}12^{9}27^{1}$
VCVI^{-}	$2^{9}5^{4}10^{17}$
IMI^{-}	$10^{12}30^{18}$
MB^{+}	$10^{15}24^{6}30^{9}$
SMI^{-}	$2^{13}4^{6}10^{11}$
TLW^{-}	$8^{2}10^{22}12^{6}$

Table 7 provides the number of indexes correctly classified on the three datasets. It can be seen that the number of datasets with correct classification of TLW is the largest. It shows that the TLW index has better stability and correctness.

Table 7. Clustering cases of all CVIs.

	UCI datasets	Artificial datasets	The Olivetti face dataset	Sum total
CH^{+}	0	0	0	0
PC^{+}	2	4	0	6
Dunn^{+}	2	0	0	2
NPC^{+}	2	4	0	6
FSI^{-}	2	6	0	8
XBI^{-}	4	7	1	12
DB^{-}	2	4	0	6
WLI^{-}	5	4	1	10
VCVI^{-}	5	2	1	8

(*continued*)

Table 7. (*continued*)

	UCI datasets	Artificial datasets	The Olivetti face dataset	Sum total
IMI^{-}	3	4	1	8
MB^{+}	7	6	1	14
SMI^{-}	4	6	1	11
TLW^{-}	8	8	1	17

4 Summary and Outlook

In this study, we propose a new fuzzy clustering validity index named TLW index. The TLW index takes the good aspects of previous index and improves them. The TLW index have three components. On the basis of separateness and compactness, cluster number K is added to alleviate the problem that the index may change monotonically with the increase of cluster number. The TLW index also improves separateness and compactness. The experiment adopts the classical FCM clustering algorithm, 12 comparative indexes and 17 datasets. The experimental results prove the feasibility and accuracy of the TLW index.

In recent years, granular computing [22, 23] and logical reasoning [24, 25] have highlighted great research value in the field of artificial intelligence. We hope that the proposed TLW index combined with new research theories and directions may bring new breakthroughs.

Acknowledgment. It was subsidized from National Natural Science Foundation of China (62176083, 62176084) and Fundamental Research Funds for Central Universities of China (PA2023GDSK0061).

References

1. Tang, Y.M., Pan, Z.F., Pedrycz, W., Ren, F.J., Song, X.C.: Viewpoint-based kernel fuzzy clustering with weight information granules. IEEE Trans. Emerg. Top. Comput. Intell. **7**(2), 342–356 (2023)
2. Tang, Y.M., Ren, F.J., Pedrycz, W.: Fuzzy C-means clustering through SSIM and patch for image segmentation. Appl. Soft Comput. **87**, 105928, 1–16 (2020)
3. Tang, Y.M., Li, L., Liu, X.P.: State-of-the-art development of complex systems and their simulation methods. Complex Syst. Model. Simulat. **1**(4), 271–290 (2021)
4. Tang, Y.M., Huang, J.J., Pedrycz, W., et al.: A fuzzy cluster validity index induced by triple center relation. IEEE Trans. Cybernet. **53**(8), 5024–5036 (2023)
5. Wu, C.H., Ouyang, C.S., Chen, L.W., et al.: A new fuzzy clustering validity index with a median factor for centroid-based clustering. IEEE Trans. Fuzzy Syst. **23**(3), 701–718 (2014)
6. Wan, Y.T., Ma, A.L., Zhang, L.P., Zhong, Y.F.: Multiobjective sine cosine algorithm for remote sensing image spatial-spectral clustering. IEEE Trans. Cybernet. **52**(10), 11172–11186 (2022)
7. Rathore, P., Ghafoori, Z., Bezdek, J.C., Palaniswami, M., Leckie, C.: Approximating Dunn's cluster validity indices for partitions of big data. IEEE Trans. Cybernet. **49**(5), 1629–1641 (2019)

8. Salem, S.A., Nandi, A.K.: Development of assessment criteria for clustering algorithms. Pattern Anal. Appl. **12**(1), 79–98 (2009)
9. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
10. Bezdek, J.C.: Numerical taxonomy with fuzzy sets. J. Math. Biol. **7**(1), 57–71 (1974)
11. Dunn, J.C.: A fuzzy relative of the ISODA TA process and its use in detecting compact well-separated clusters. Cybern. Syst. **3**(3), 32–57 (1973)
12. Roubens, M.: Pattern classification problems and fuzzy sets. Fuzzy Sets Syst. **1**(4), 239–253 (1978)
13. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of cluster for the fuzzy c-means method. In: 5th Fuzzy Systems Symposium Kobe, pp. 247–250 (1989)
14. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**(8), 841–847 (1991)
15. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979)
16. Wu, C.H., Ouyang, C.S., Chen, L.W., et al.: A new fuzzy clustering validity index with a median factor for centroid-based clustering. IEEE Trans. Fuzzy Syst. **23**(3), 701–718 (2015)
17. Liu, Y., Jiang, Y., Hou, T., et al.: A new robust fuzzy clustering validity index for imbalanced data sets. Inf. Sci. **547**, 579–591 (2021)
18. Mittal, H., Saraswat, M.: A new fuzzy cluster validity index for hyperellipsoid or hyperspherical shape close clusters with distant centroids. IEEE Trans. Fuzzy Syst. **29**(11), 3249–3258 (2020)
19. Zhu, E., Ma, R.: An effective partitional clustering algorithm based on new clustering validity index. Appl. Soft Comput. **71**, 608–621 (2018)
20. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. **24**(12), 1650–1654 (2002)
21. Liang, J., Bai, L., Dang, C., et al.: The K-means-type algorithms versus imbalanced data distributions. IEEE Trans. Fuzzy Syst. **20**(4), 728–745 (2012)
22. Tang, Y.M., Ren, F.J.: Fuzzy systems based on universal triple I method and their response functions. Int. J. Inf. Technol. Decis. Mak. **16**(2), 443–471 (2017)
23. Tang, Y.M., Zhang, L., Bao, G.Q., Ren, F.J., Pedrycz, W.: Symmetric implicational algorithm derived from intuitionistic fuzzy entropy. Iranian J. Fuzzy Syst. **19**(4), 27–44 (2022)
24. Tang, Y.M., Pan, Z.H., Hu, X.H., Pedrycz, W., Chen, R.H.: Knowledge-induced multiple kernel fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. (2023). https://doi.org/10.1109/TPAMI.2023.3298629
25. Tang, Y.M., Pedrycz, W.: Oscillation bound estimation of perturbations under Bandler-Kohout subproduct. IEEE Trans. Cybernet. **52**(7), 6269–6282 (2022)