Hindawi
Journal of Electrical and Computer Engineering
Volume 2019, Article ID 2719617, 9 pages
https://doi.org/10.1155/2019/2719617

Research Article

A Validity Index for Fuzzy Clustering Based on Bipartite Modularity

Yongli Liu, Xiaoyang Zhang, Jingli Chen, and Hao Chao

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, Henan, China

Correspondence should be addressed to Yongli Liu; yongli.buaa@gmail.com

Received 4 January 2019; Revised 18 July 2019; Accepted 18 July 2019; Published 8 August 2019

Academic Editor: Yagang Zhang

Because traditional fuzzy clustering validity indices need to specify the number of clusters and are sensitive to noise data, we propose a validity index for fuzzy clustering, named CSBM (compactness separateness bipartite modularity), based on bipartite modularity. CSBM enhances the robustness by combining intraclass compactness and interclass separateness and can automatically determine the optimal number of clusters. In order to estimate the performance of CSBM, we carried out experiments on six real datasets and compared CSBM with other six prominent indices. Experimental results show that the CSBM index performs the best in terms of robustness while accurately detecting the number of clusters.

1. Introduction

Recently, in order to reveal valuable knowledge and patterns behind data, data mining has become increasingly crucial in many fields such as automatic categorization of text documents [1, 2], grouping search engine results [3, 4], analyzing time series data [5, 6], and others [7–10]. As one of the vital techniques of data mining, clustering can divide a group of samples into multiple clusters, so that elements in the same cluster are as similar as possible and elements in different clusters are as dissimilar as possible.

In fuzzy clustering, represented by the FCM (fuzzy C means) [11] algorithm, the value of membership degree is fuzzy, which means that a sample is allowed to belong to multiple clusters with different probabilities. It is more consistent with the rule of sample distribution than the hard clustering logic; therefore, the fuzzy clustering research has constantly been ongoing and innovative. As yet, a large number of fuzzy clustering algorithms have been increasingly improved in accuracy, efficiency, robustness, and other aspects, which significantly boosts the development of data mining. At the same time, the validity index used to measure the clustering quality of fuzzy clustering, as an indispensable part of algorithm research, plays a growing important role in fuzzy clustering.

Recently, the achievement on clustering validity index is pretty fruitful. Hu et al. [12] proposed a clustering validity index by combining intraclass compactness with intercluster separateness, which reduces the impact of noise data well. Chen and Pi [13] proposed a nondistance validity index based on fuzzy membership degree by mixing with compactness and separateness, which improves the identification of overlapping clusters while weakening the sensitivity to noise data. Although these indices enhance the robustness by combining intraclass compactness with interclass separateness, it is still necessary to manually specify the number of clusters owing to the restrictions of the FCM algorithm. Zhang et al. [14], regarding the fuzzy membership degree and the bipartite modularity as the global and local attributes, respectively, proposed a weighted global-local validity based index (WGLI). The WGLI can automatically determine the number of clusters but is vulnerable to noise data because of its higher dependency on the fuzzy membership degree of the FCM algorithm.

Motivated by the above analysis, in this paper, we apply the bipartite modularity to the constructed bipartite network based on the clustering result of the FCM algorithm and evaluate the final partition result combining intraclass compactness with interclass separateness at the same time. The proposed validity index fuses the bipartite modularity

with intraclass compactness and interclass separateness, which is not only able to enhance the robustness of clustering results but also able to determine the optimal number of clusters automatically.

2. Related Work

2.1. FCM Algorithm. The FCM algorithm divides N data samples $x_i (i = 1, 2, \ldots, N)$ into C fuzzy clusters by means of the fuzzy partition method and then calculates the centre of each cluster. Its objective function shown in formula (1) is minimized through the process of iteration:

$$J_m(U, V) = \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci}^m \left\| x_i - v_c \right\|^2, \tag{1}$$

where v_c represents the center of c-th cluster, m is the fuzzy parameter and $m \in (1, \infty)$, and u_{ci} indicates the membership degree that the sample i belonging to the cluster c and

$$\sum_{c=1}^{C} u_{ci} = 1, \quad u_{ci} \in [0, 1]. \tag{2}$$

Then, the expressions of u_{ci} and v_c can be obtained using the Lagrange multiplier method as follows:

$$u_{ci} = \frac{1}{\sum_{k=1}^{C} \left(d_{ci}/d_{ki}\right)^{2/(m-1)}}, \tag{3}$$

$$v_c = \frac{\sum_{i=1}^{N} u_{ci}^m x_i}{\sum_{i=1}^{N} u_{ci}^m}, \tag{4}$$

where d_{ci} and d_{ki}, respectively, represent the Euclidean distances from the clustering centers c and k to the sample i.

2.2. Fuzzy Clustering Validity Index. The fuzzy clustering validity index can measure the clustering performance and is fairly significant in the fuzzy clustering research. The prevalent fuzzy clustering validity indices are shown in Table 1.

2.3. Bipartite Modularity. Newman and Girvan [23] introduced modularity to measure the strength of community structure in a single network. However, networks in the real world not only exist in this simple form, which means, for example, in the metabolic networks [24], pathological networks [25], and the World Wide Web; there may be one-to-many and even many-to-many relationships between the vertices and the divided communities instead of the simple one-to-one relationship. For complicated networks, by extending and concretizing the matrix-based method that Newman used for bipartite network, Barber [26] defined a zero model and then proposed the bipartite modularity applied in bipartite networks with special constraints which means that the vertices of bipartite network are divided into two disjoint sets and each edge connects two vertices from these two sets, respectively. However, the disadvantages of the bipartite modularity include that the number of communities must be determined in advance, and the number of communities of the two types of vertices in the bipartite network must be equal. Guimerà et al. [27] proposed the bipartite modularity by adopting the idea of modularity maximization and transformed the module identification problem into the combinatorial optimization problem. Unfortunately, the module structure of a type of vertex should be manually specified, which will affect the accuracy of partition.

Then, Murata [22] defined a modified bipartite modularity, which allows random connections between two types of vertices, and a community containing a type of vertex can correspond to one or more communities of the other type of vertex. When there is a complete one-to-one relationship between communities of different types of vertices, the bipartite modularity reaches its maximum.

Let G be a bipartite network, in which M denotes the total number of edges and V is a set including all vertices. The bipartite network is divided into communities of X-vertex and Y-vertex, and the numbers of these communities are L^X and L^Y. V^X and V^Y are defined as sets of communities of X-vertex and Y-vertex and represented as $V^X = \{V_1^X, V_2^X, \ldots, V_{L^X}^X\}$ and $V^Y = \{V_1^Y, V_2^Y, \ldots, V_{L^Y}^Y\}$, respectively, in which the elements V_l^X and V_m^Y represent a community of X-vertex and Y-vertex, respectively. Let A be an adjacency matrix, and one of its elements can be expressed as $A(i, j)$. Moreover, if the vertexes i and j are connected, $A(i, j) = 1$, otherwise $A(i, j) = 0$.

Assuming that the two communities V_l and V_p are different from each other, which means $(V_l \in V^X \wedge V_p \in V^Y) \vee (V_l \in V^Y \wedge V_p \in V^X)$; the number of all the edges, denoted by e_{lp}, that connect the vertices from V_l and V_p and its row sum, denoted by a_l, can be respectively expressed as

$$e_{lp} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V_p} A(i, j), \tag{5}$$

$$a_l = \sum_{p} e_{lp} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V} A(i, j). \tag{6}$$

The bipartite modularity Q_B is defined as

$$\begin{aligned} Q_B = Q_{B^{XY}} + Q_{B^{YX}} &= \sum_{l_1 \in V^X} \left(e_{l_1 p_1} - a_{l_1} a_{p_1}\right) \\ &+ \sum_{l_2 \in V^Y} \left(e_{l_2 p_2} - a_{l_2} a_{p_2}\right), \end{aligned} \tag{7}$$

where

$$p_1 = \arg\max_{k_1 \in V^Y} \left(e_{l_1 k_1}\right), \tag{8}$$

$$p_2 = \arg\max_{k_2 \in V^X} \left(e_{l_2 k_2}\right). \tag{9}$$

Q_B indicates the sum of bipartite modularity in two different directions, $V^X \longrightarrow V^Y$ and $V^Y \longrightarrow V^X$, $Q_{B^{XY}}$ and $Q_{B^{YX}}$, respectively, denote the expected values in two directions, which means the number of edges connecting the corresponding vertices from communities of X-vertex and

Table 1: The commonly used fuzzy clustering validity indices.

Indices	Definition	Description
PC (partition coefficient)	$\mathrm{PC} = (1/N)\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^2$	The partition coefficient (PC) [15] measures the fuzzy degree of final divided clusters by means of the fuzzy partition matrix, and the larger its value, the better the partition result
PE (partition entropy)	$\mathrm{PE} = -(1/N)\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}\log(u_{ci})$	The partition entropy (PE) [16] measures the fuzzy degree of final divided clusters by means of the fuzzy partition matrix, and the smaller its value, the better the partition result
MPC (modified partition coefficient)	$\mathrm{MPC} = (C \times I_{\mathrm{PC}} - 1)/(C-1)$	Because the value of PC merely depends on the membership degree u_{ci}, Dave [17] proposed the modified PC index MPC, and the larger its value, the better the partition result
MPE (modified partition entropy)	$\mathrm{MPE} = (N \times I_{\mathrm{PE}})/(N-C)$	Because the value of PE merely depends on the membership degree u_{ci}, Dave [17] proposed the modified PE index MPE, and the smaller its value, the better the partition result
XB (Xie–Beni index)	$\mathrm{XB} = J_m/(N \times \min_{i,j=1,\ldots,C,i\neq j}\|v_i - v_j\|^2)$	Considering the membership degree and the structure of datasets, Xie and Beni [18] proposed the XB index to measure the overall average compactness and separateness, and the smaller its value, the better the partition result
FS (Fukuyama–Sugeno index)	$\mathrm{FS} = J_m(U,V) - K_m(U,V) = J_m - \sum_{i=1}^{N}\sum_{c=1}^{C}u_{ci}^m\|v_c - \bar{v}\|^2$, where $\bar{v} = (1/N)\sum_{i=1}^{N}x_i$	Fukuyama and Sugeno [19] also proposed the index FS considering the compactness and separateness, and when its value reaches the minimum, the partition result is the best
PCAES (partition coefficient and exponential separation)	$\mathrm{PCAES} = \sum_{c=1}^{C}\sum_{i=1}^{N}(u_{ci}^2/u_M) - \sum_{c=1}^{C}\exp(-\min_{k\neq c}\{\|v_c - v_k\|^2\}/\beta_T)$, where $u_M = \min_{1\le c\le C}\{\sum_{i=1}^{N}u_{ci}^2\}, \beta_T = (\sum_{l=1}^{C}\|v_l - \bar{v}\|^2)/C, \bar{v} = (1/N)\sum_{i=1}^{N}x_i$	Wu and Yang [20] proposed the index PCAES by combining the normalized partition coefficient with the exponential separateness degree of each cluster, and the larger its value, the better the partition result
CO (compactness and overlap measures)	$\mathrm{CO} = C(c,U) - O(c,U) = (1/n)\sum_{j=1}^{n}(\sum_{i=1}^{c}C_{ij}(c,U) - \sum_{a=1}^{c-1}\sum_{b=a+1}^{c}O_{abj}(c,U))$, where $C_{ij}(c,U) = \begin{cases} u_{ij} \text{ if } (u_{ij} - u_{ik}) \ge T_c, k = 1,\ldots,c, k\neq j \\ 0 \quad \text{otherwise} \end{cases}$ $O_{abj}(c,U) = \begin{cases} 1 - (u_{aj} - u_{bj}) \text{ if } (u_{aj} - u_{bj}) \ge T_o, a \neq b \\ 0 \quad \text{otherwise} \end{cases}$	Žalik [21] proposed the index CO based on the compactness and separateness, and the larger its value, the higher the compactness degree, and the lower the coverage degree between clusters, the better the partition result
WGLI (weighted global-local index)	$\mathrm{WGLI} = (2\mathrm{MMD} + Q_B')/3$, where $\mathrm{MMD} = (1/n)\sum_{j=1}^{n}\max_{1\le i\le C}u_{ij}$	Zhang et al. [14], based on the membership degree obtained from the FCM algorithm, proposed the index WGLI combining the bipartite modularity; Q_B' represents the bipartite modularity proposed by Murata [22], and the larger the value of WGLI, the better the partition result

Y-vertex minus the number of edges that are connected randomly between X-vertex and Y-vertex in the same divided communities. The larger the value of Q_B, the stronger the community structure of bipartite network and the better the result of community detection.

2.4. Constructing Bipartite Network. According to the membership degree matrix and C clusters obtained from the FCM algorithm, a weighted bipartite network can be constructed. The X-vertex is represented by all the cluster centers, the Y-vertex is denoted by all the sample points, and the weighted edges are indicated by membership degrees. Applying the bipartite modularity to the constructed bipartite network, we have $L^X = L^Y = C$, $M = \sum_{i=1}^{C}\sum_{j=1}^{N}A(i,j)$, and the adjacency matrix $A(i,j)$ can be defined as follows:

$$A(i,j) = \begin{cases} 1.0, & u_{ci} > \alpha, \\ u_{ci}, & (1-\alpha) \le u_{ci} \le \alpha, \\ 0.0, & u_{ci} < (1-\alpha), \end{cases} \tag{10}$$

where $\alpha = 0.7$, u_{ci} denotes the fuzzy membership degree of the FCM algorithm.

Suppose that the FCM algorithm runs on a dataset including 10 sample points and 4 clusters. According to formula (10), a bipartite network can be constructed by cluster centers and sample points and shown as Figure 1.

In Figure 1, the top nodes, which are also the X-vertex, represent different communities composed of cluster centers, and the bottom nodes, which are correspondingly the Y-vertex, represent different communities composed of divided datasets. Besides, the weighted values of the edges

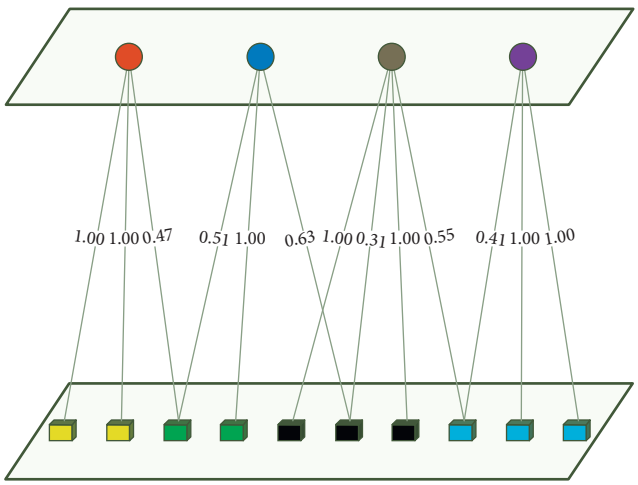

Figure 1: The example of bipartite network.

are the corresponding membership degrees, which are used to express the values of A (i, j). According to the formulas (5)–(10), the value of Q_B can be calculated.

3. Fuzzy Clustering Validity Index CSBM

In this paper, we proposed a fuzzy clustering validity index named CSBM. This index combines three components: (1) bipartite modularity, (2) intraclass compactness, and (3) interclass separateness. First, CSBM builds a bipartite network based on the clustering results of FCM and applies the bipartite modularity to this bipartite network. Second, CSBM evaluates the clustering results by combining intraclass compactness and interclass separation.

Compared with the conventional validity indices, the index CSBM can enhance the robustness on the one hand and automatically determine the optimal number of clusters on the other hand.

3.1. Intraclass Compactness. The intraclass compactness of the index CSBM is defined as

$$\mathrm{NC} = \sum_{i=1}^{N}\sum_{c=1}^{C}\frac{u_{ci}^2}{u_{\max}}, \tag{11}$$

where

$$u_{\max} = \max_{1\le i\le N}\left\{\sum_{c=1}^{C}u_{ci}^2\right\}. \tag{12}$$

NC (novel compactness) improves the performance of the partition coefficient PC, where the compactness of the cluster c is expressed as $u_{ci}^2/u_{\max}$. The larger the value of NC, the higher the intraclass compactness, and the better the result of fuzzy partition.

3.2. Interclass Separateness. In order to reduce the impact of noise data on the clustering result, the interclass separateness is measured by the distance between different fuzzy clusters, which is defined as follows:

$$\mathrm{SEP} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{a=1}^{C-1}\sum_{b=a+1}^{C}O_{abi}(C,U)\right), \tag{13}$$

where

$$O_{abi}(C,U) = \begin{cases} 1-\left|u_{ai}-u_{bi}\right|, & \text{if}\left|u_{ai}-u_{bi}\right|\ge T_o,\ a\ne b,\\ 0, & \text{otherwise.}\end{cases} \tag{14}$$

The threshold T_o is used to eliminate the noise points on the cluster boundary and represents the separateness between samples of two clusters a and b. u_{ai} and u_{bi} indicate the membership degrees that sample point i belonging to cluster a and b, respectively. The smaller the value of $O_{abi}(C, U)$, the lower the coverage degree between clusters a and b, and the higher the separation degree between the two clusters, whereas SEP (separateness) indicates the sum of the cluster separation degrees of all sample points in the fuzzy membership matrix, and the smaller the value of SEP, the better the result of fuzzy partition.

3.3. Index CSBM. The objective function used to calculate the index CSBM is defined as

$$\mathrm{CSBM} = (C-1)^{(1/C)}\times\frac{(\mathrm{NC}-\mathrm{SEP}+Q_B)}{2}. \tag{15}$$

The introduction of $(C-1)^{(1/C)}$ is to adjust the value of CSBM. The larger the value of NC, the smaller the value of SEP and the better the clustering result. At the same time, the larger the value of NC-SEP and Q_B, the larger the value of CSBM. The better the clustering quality of FCM algorithm, the more accurate the optimal number of clusters.

The calculation process of the fuzzy clustering validity index (CSBM) is as follows:

(1) Input: the dataset S and threshold ε used in the FCM algorithm

(2) Output: the clustering result.

(3) Running the FCM algorithm on the given dataset

(4) Constructing the weighted bipartite network according to the membership degree u_{ci} and the cluster centre v_c obtained from formulas (3) and (4) and the adjacency matrix A (i, j) from formula (10).

(5) Calculating the bipartite modularity Q_B according to formulas (5)–(10)

(6) Calculating CSBM according to formulas (11)–(15)

4. Experiments

4.1. Datasets. In order to verify the effectiveness and validity of the index CSBM, we select six datasets from UCI database. The detailed information of these six datasets is shown in Table 2.

4.2. Evaluation Criteria for Clustering Result. F-measure (FM) and entropy (EN) are used to evaluate the clustering result of the FCM algorithm. F-measure is often used to

Table 2: The detailed information of datasets.

Datasets	Samples	Attributes	Clusters	Description
Iris	150	4	3	The dataset is the most famous database in the pattern recognition literatures, including 150 samples, 4 attributes, and 3 classes containing 50 instances each, and each class refers to a type of iris.
Wine	178	13	3	The dataset is the result of chemical analysis on three different kinds of wines from the same region of Italy. It contains 178 results. The analysis determines the number of 13 components found in each wine. The dataset is divided into three categories.
Wpbc	194	33	2	The dataset is a Wisconsin breast cancer dataset with 194 samples and 33 attributes, and each sample represents a subsequent data of breast cancer case. The dataset is divided into two categories.
Hayesroth	132	5	3	The dataset describes the evidence of behaviour classification, confidence degree, and recognition behaviour that are closely related to human behaviour. It contains 132 samples and 5 attributes using numerical values instead of actual values. The dataset is divided into three classes.
Zoo	101	16	7	The dataset describes the classification of animals, including 101 samples and 17 attributes. The dataset is divided into 7 classes.
Glass	214	9	6	The study of glass classification is motivated by criminological investigations. If the identification is correct, the remaining glass can act as evidence. The dataset consists of 214 samples and 9 attributes, which is divided into 6 classes.

evaluate the partition result of clustering algorithms and can be defined as

$$F = \frac{2\mathrm{PR}}{P + R}, \tag{16}$$

where P represents the accuracy rate, which means the proportion of the related files retrieved by the system to the total number of all the related files in the system and R represents the recall rate, which indicates the proportion of the related files retrieved by the system to the total number of all the files retrieved by the system. In general, the accuracy rate and recall rate interact on each other. Considering these two factors, the index F-measure can be calculated and reveals the overall performance. The value range of F-measure is [0, 1], and the larger its value, the better the clustering result.

Entropy, also known as Shannon entropy [28], is represented by a set of discrete probabilities p_i, which, in the case of sending a message, are the ones that a particular message is actually sent. The entropy of message system is used to measure the average amount of information in a message and can be defined as follows:

$$E = \sum_{i=1}^{N} p_i \log_2 p_i. \tag{17}$$

The value range of entropy is [0, 1], and the smaller its value, the better the clustering result.

4.3. Experimental Results. In order to evaluate the accuracy of CSBM in terms of predicting the optimal number of clusters, we compare it with six prevalent fuzzy clustering validity indices, including PC, PE, MPC, MPE, CO, and WGLI. Tables 3–8 show the values of the index CSBM and other comparative indices obtained from running the FCM algorithm on six datasets, in which the values of F-measure and entropy are also included, so the practicability and effectiveness of the FCM algorithm can be confirmed. The bold values in the tables denote the values of indices corresponding to the optimal number of clusters.

The experimental results can be clearly interpretive from the following three aspects:

(1) On the iris, wine, and zoo datasets (Tables 3, 4 and 7, respectively), the index CSBM outperforms other indices, which means its matching classification number indicates the best partition result. On the iris and wine datasets shown in Tables 3 and 4, respectively, other indices could produce their own best index values when the cluster number equals 2. And actually, these two datasets have both 3 classes. Table 7 lists the results on the zoo dataset which has 7 classes. However, WGLI, PC, MPC, and CO all generate their own best index values when there are 4 clusters, and PE and MPE think the best number of clusters is 3.

(2) On the wpbc and Hayesroth datasets shown in Tables 5 and 6, respectively, some indices including CSBM all gain their own optimal values, while others are not as satisfactory. On the wpbc dataset, there are 5 indices that could find the realistic number of

Table 3: Values of each index on the iris dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
2	0.7582	**0.7966**	**0.8724**	**0.0833**	**0.7668**	**0.0844**	**0.8840**	0.8399	0.2507
3	**0.7744**	0.7427	0.7660	0.1682	0.6599	0.1716	0.8129	0.9815	0.1338
4	0.7625	0.6926	0.6699	0.2468	0.5672	0.2536	0.7130	0.8652	0.1304
5	0.7593	0.6824	0.6508	0.2871	0.5690	0.2970	0.7072	0.8458	0.0888
6	0.7646	0.6472	0.5931	0.3474	0.5161	0.3618	0.6035	0.7968	0.0825
7	0.7647	0.6248	0.5447	0.3856	0.4725	0.4045	0.5832	0.7232	0.0446

Table 4: Values of each index on the wine dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
2	0.7436	**0.7760**	**0.8518**	**0.0912**	**0.7313**	**0.0922**	**0.8515**	0.8533	0.3846
3	**0.7564**	0.7427	0.7690	0.1606	0.6674	0.1634	0.8294	0.8383	0.3304
4	0.7488	0.7314	0.7613	0.1767	0.6910	0.1807	0.8322	0.7379	0.3382
5	0.7483	0.7146	0.7235	0.2152	0.6613	0.2214	0.7933	0.7587	0.2875
6	0.7475	0.7192	0.7258	0.2204	0.6765	0.2280	0.8005	0.6555	0.282
7	0.7411	0.7208	0.7368	0.2134	0.6976	0.2221	0.7973	0.6425	0.2743

Table 5: Values of each index on the wpbc dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
2	**0.7187**	**0.7460**	**0.7640**	**0.1180**	0.5997	**0.1192**	0.7407	0.7833	0.2788
3	0.7109	0.7098	0.7132	0.1686	**0.6057**	0.1713	**0.7654**	0.6914	0.2723
4	0.7142	0.7003	0.6786	0.2079	0.5954	0.2122	0.7526	0.5835	0.2778
5	0.7158	0.6726	0.6341	0.2527	0.5605	0.2594	0.6769	0.5362	0.2781
6	0.7106	0.6635	0.6013	0.2889	0.5359	0.2981	0.6703	0.4124	0.2758
7	0.7175	0.6547	0.5922	0.3045	0.5362	0.3159	0.6621	0.4014	0.2729

Table 6: Values of each index on the Hayesroth dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
2	0.7429	**0.7697**	**0.8141**	**0.1022**	**0.6749**	**0.1038**	0.7985	0.4485	0.4618
3	**0.7439**	0.7419	0.7616	0.1529	0.6657	0.1564	**0.8072**	0.4328	0.4544
4	0.7404	0.7252	0.7318	0.1856	0.6579	0.1914	0.773	0.3603	0.4574
5	0.7368	0.7125	0.711	0.2099	0.6504	0.2182	0.7778	0.3369	0.4562
6	0.7355	0.7041	0.6945	0.2301	0.6428	0.241	0.7625	0.3269	0.4437
7	0.7343	0.6958	0.6786	0.2493	0.6329	0.2633	0.7405	0.2997	0.4427

Table 7: Values of each index on the zoo dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
3	0.7853	0.6493	0.6147	**0.2940**	0.4248	**0.3030**	0.5777	0.6000	0.3421
4	0.7891	**0.6711**	**0.6196**	0.3219	**0.4946**	0.3352	**0.6767**	0.6825	0.2475
5	0.7885	0.622	0.5326	0.4117	0.4171	0.4332	0.5685	0.6652	0.2183
6	0.7778	0.5996	0.4917	0.4697	0.3911	0.4993	0.5662	0.7103	0.1625
7	**0.8149**	0.5542	0.4496	0.5117	0.3588	0.5498	0.4215	0.5737	0.1800
8	0.7969	0.556	0.4476	0.5342	0.3695	0.5802	0.4691	0.6003	0.1248

Table 8: Values of each index on the glass dataset.

Clusters	CSBM	WGLI	PC	PE	MPC	MPE	CO	FM	EN
2	0.7252	**0.7296**	**0.7702**	**0.1314**	0.5864	**0.1326**	**0.7498**	0.6782	0.6697
3	0.7036	0.6958	0.7215	0.1909	**0.6054**	0.1936	0.7392	0.7240	0.6119
4	0.7350	0.6421	0.6051	0.2859	0.4889	0.2913	0.6576	0.7038	0.5901
5	**0.7852**	0.5568	0.4763	0.3925	0.3569	0.4019	0.3512	0.6558	0.5753
6	0.7746	0.5541	0.4703	0.4127	0.3736	0.4246	0.5503	0.6776	0.5364
7	0.7845	0.5307	0.4193	0.4799	0.3302	0.4961	0.4578	0.6967	0.5378

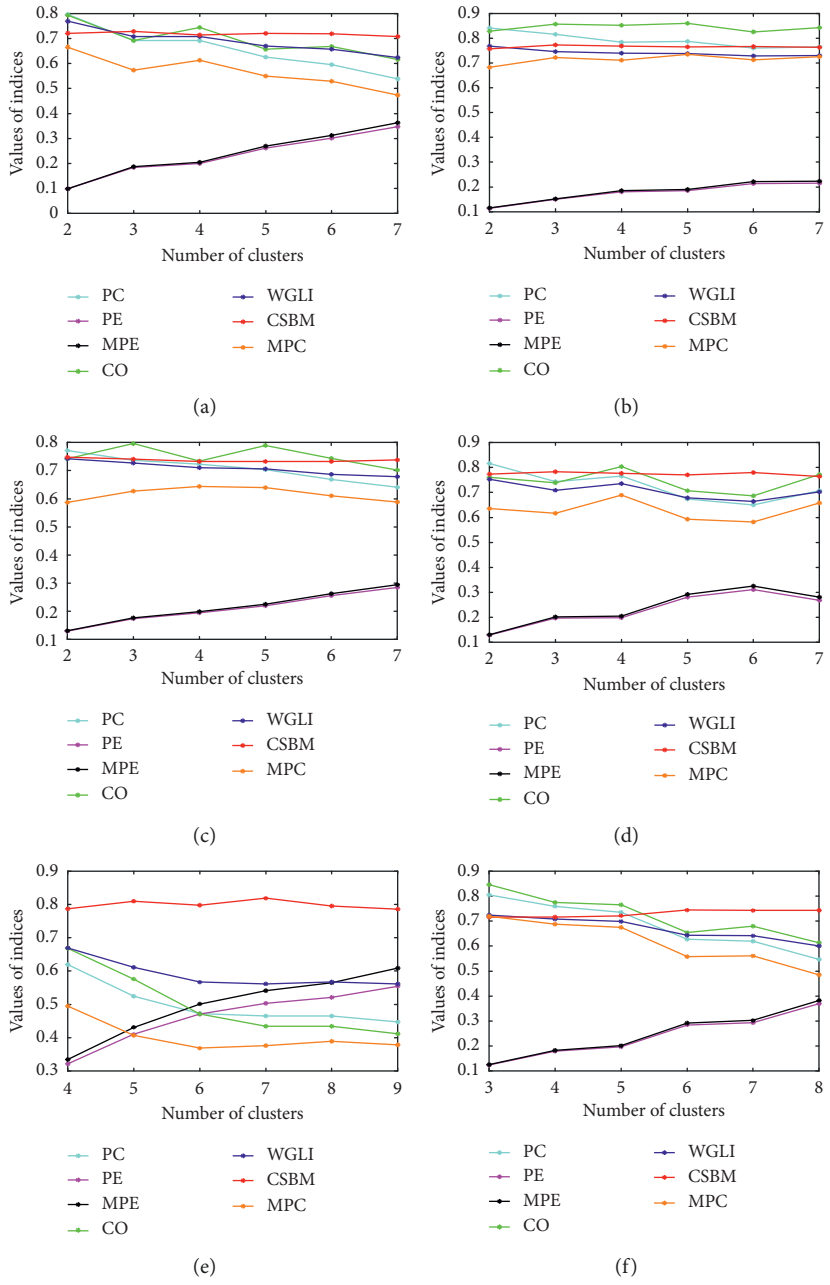

FIGURE 2: The variation trend of each index with the number of cluster (C) after adding noise data. (a) Iris. (b) Wine. (c) Wpbc. (d) Hayesroth. (e) Zoo. (f) Glass.

classes accurately, including CSBM, WGLI, PC, PE, and MPE. Other two indices, MPC and CO, find more clusters. On the Hayesroth dataset, only the CSBM and CO generate accurate number of clusters and the other 5 indices generate less clusters.

(3) On the glass dataset as in Table 8, all the indices cannot achieve the desired results. But compared with the other indices, the predicted value, 5, of the index CSBM is closest to the standard value, 6. The second best is the index MPC, whose predicted number of clusters is 3. The rest indices, including WGLI, PC, PE, MPE, and CO, all generate their own best values when the number of clusters is 2.

According to the above experimental results, the index CSBM performs better than these comparative indices in terms of predicting the optimal number of clusters.

In order to verify the robustness of the index CSBM, we add some noisy data to each dataset at a rate of 10% and then run the FCM algorithm again. The change trend of values of all the indices with the number of clusters is shown as Figure 2. At the same time, the OC (original clusters) is the original classification number of each dataset, and the optimal numbers of clusters of each dataset determined by all the indices before and after adding noise data are shown in Table 9.

Table 9: The original classification number of each dataset and the optimal number of clusters determined by each index before and after adding noise data.

Datasets	OC	CSBM	WGLI	PC	PE	MPC	MPE	CO
Iris	3	**3(3)**	2(2)	2(2)	2(2)	2(2)	2(2)	2(2)
Wine	3	**3(3)**	2(2)	2(2)	2(2)	2(5)	2(2)	2(**3**)
Wpbc	2	**2(2)**	**2(2)**	**2(2)**	**2(2)**	3(4)	**2(2)**	3(3)
Hayesroth	3	**3(3)**	2(2)	2(2)	2(2)	2(4)	2(2)	**3**(4)
Zoo	7	**7(7)**	4(4)	4(4)	3(3)	4(4)	3(3)	4(4)
Glass	6	5(**6**)	2(2)	2(3)	2(3)	3(3)	2(3)	2(3)

The numbers outside and within parentheses, respectively, represent the optimal number of clusters determined by the corresponding indices before and after the addition of noise data, and the coarser values indicate that the number of clusters matches the standard values

It can be seen from Figure 2 and Table 9 that the index CSBM is less vulnerable to the added noise data than other comparative indices and can still determine the optimal number of clusters accurately. Meanwhile, due to the noise data, the results of indices WGLI, PC, PE, and MPE on zoo and glass datasets, index MPC on wine, wpbc and Hayesroth datasets and index CO on wine, and Hayesroth and glass datasets are all changed in different degrees. According to Figure 2, after adding noise data, the values of some indices remain unchanged, such as the results of WGLI on iris dataset, and the values of some indices are slightly closer to standard values, such as the results of CO on wine dataset. However, some indices differ from standard values to a greater extent, such as the results of MPC on wine dataset. The overall change shows that there is a big difference between the value of each index and the standard values in theory after adding noise data, but the actual results are relatively random and lack of the corresponding rules, which furthermore verifies the excellent performance of the index CSBM in terms of robustness.

5. Conclusions

In this paper, a new fuzzy clustering validity index named CSBM is proposed. This index modifies the intraclass compactness and interclass separateness on the basis of conventional indices, which means that it enhances the robustness and weakens the impact of noise data. At the same time, the optimal cluster number of fuzzy clustering can be predicted more accurately by integrating with bipartite modularity. Six datasets from the UCI database are selected for the sake of validating the feasibility and validity of the index CSBM. The results show that the index CSBM performs better than other comparative indices in terms of clustering accuracy and robustness and predicts the optimal number of clusters more precisely.

Data Availability

All the datasets used in this paper are derived from the UCI (University of California Irvine) Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank members of the IR&DM Research Group from Henan Polytechnic University for their invaluable advice that makes this paper successfully completed. The authors would also like to thank the support of the Foundation for Scientific and Technological Project of Henan Province under Grant 172102210279.

References

[1] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and M. A. Awadallah, "A krill herd algorithm for efficient text documents clustering," in *Proceedings of the IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 67–72, IEEE Computer Society, Washington, DC, USA, 2016.

[2] N. Akhtar, M. N. Qureshi, and M. V. Ahamad, "An improved clustering method for text documents using neutrosophic logic," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 36, no. 4, pp. 197–203, 2017.

[3] T. Nie, Y. Ding, C. Zhao et al., "Clustering search engine suggests by integrating a topic model and word embeddings," in *Proceedings of the 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, IEEE Computer Society, Washington, DC, USA, June 2017.

[4] T. D. Rajkumar, S. P. Raja, and A. Suruliandi, "Users' click and bookmark based personalization using modified agglomerative clustering for web search engine," *International Journal on Artificial Intelligence Tools*, vol. 26, no. 6, article 1730002, 2017.

[5] M. Łuczak, "Hierarchical clustering of time series data with parametric derivative dynamic time warping," *Expert Systems with Applications*, vol. 62, pp. 116–130, 2016.

[6] Y. Liu, J. Chen, S. Wu, Z. Liu, and H. Chao, "Incremental fuzzy *C* medoids clustering of time series data using dynamic time warping distance," *PLoS One*, vol. 13, no. 5, Article ID e0197499, 2018.

[7] S. R. B. Prabhu, R. Mahalakshmi, S. Nithya et al., *A Review of Energy Efficient Clustering Algorithm for Connecting Wireless Sensor, Network Fields*, Social Science Electronic Publishing, New York, NY, USA, 2017.

[8] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth*

Observations & Remote Sensing, vol. 8, no. 5, pp. 2015–2030, 2017.

[9] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, "An incremental CFS algorithm for clustering large data in industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 13, pp. 1193–1201, 2017.

[10] Y. Zhang, P. Wang, P. Cheng, and S. Lei, "Wind speed prediction with wavelet time series based on lorenz disturbance," *Advances in Electrical and Computer Engineering*, vol. 17, no. 3, pp. 107–114, 2017.

[11] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *Advanced Applications in Pattern Recognition*, vol. 22, no. 1171, pp. 203–239, 1987.

[12] Y. Hu, C. Zuo, Y. Yang et al., "A robost cluster validity index for fuzzy *c*-means clustering," in *Proceedings of the International Conference on System Science, Engineering Design and Manufacturing Informatization*, pp. 263–266, IEEE, Guiyang, China, October 2011.

[13] J. Chen and D. Pi, "A cluster validity index for fuzzy clustering based on non-distance," in *Proceedings of the Fifth International Conference on Computational and Information Sciences*, pp. 880–883, IEEE, Niterói, Brazil, November 2013.

[14] D. Zhang, M. Ji, J. Yang, Y. Zhang, and F. Xie, "A novel cluster validity index for fuzzy clustering based on bipartite modularity," *Fuzzy Sets and Systems*, vol. 253, no. 9, pp. 122–137, 2014.

[15] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of Mathematical Biology*, vol. 1, no. 1, pp. 57–71, 1974.

[16] J. C. Bezdek†, "Cluster validity with fuzzy sets," *Journal of Cybernetics*, vol. 3, no. 3, pp. 58–73, 1974.

[17] R. N. Dave, "Validating fuzzy partitions obtained through *c*-shells clustering," *Pattern Recognition Letters*, vol. 17, no. 6, pp. 613–623, 1996.

[18] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[19] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy *c*-means method," in *Proceedings of the Fifth Fuzzy System Symposium*, Kobe, Japan, 1989.

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

[21] K. R. Žalik, "Cluster validity index for estimation of fuzzy clusters of different sizes and densities," *Pattern Recognition*, vol. 43, no. 10, pp. 3374–3390, 2010.

[22] T. Murata, "Detecting communities from bipartite networks based on bipartite modularities," in *Proceedings of the International Conference on Computational Science and Engineering*, pp. 50–57, IEEE Computer Society, Washington, DC, USA, 2009.

[23] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 69, no. 2, article 026113, 2004.

[24] J. Hu, J. W. Locasale, J. H. Bielas et al., "Heterogeneity of tumor-induced gene expression changes in the human metabolic network," *Nature Biotechnology*, vol. 31, no. 6, pp. 522–529, 2013.

[25] M. Guye, G. Bettus, F. Bartolomei et al., "Graph theoretical analysis of structural and functional connectivity MRI in normal and pathological brain networks," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 23, no. 5-6, pp. 409–421, 2010.

[26] M. J. Barber, "Modularity and community detection in bipartite networks," *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 76, no. 6, article 066102, 2007.

[27] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Module identification in bipartite and directed networks," *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 76, no. 2, article 036102, 2007.

[28] R. Balian, "Entropy, a protean concept," *Poincaré Seminar, Progress in Mathematical Physics*, vol. 38, pp. 119–144, 2004.