Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 253 (2014) 122–137

FUZZY sets and systems

www.elsevier.com/locate/fss

A novel cluster validity index for fuzzy clustering based on bipartite modularity

Dawei Zhang [a], Min Ji [a], Jun Yang [b], Yong Zhang [a], Fuding Xie [b,c,*]

[a] School of Computer Science and Technology, Liaoning Normal University, Liaoning, Dalian 116081, PR China
[b] School of Urban and Environmental Science, Liaoning Normal University, Liaoning, Dalian 116029, PR China
[c] Academy of Mathematics and System Sciences, Chinese Academy of Science, Beijing 100080, PR China

Received 22 November 2012; received in revised form 18 December 2013; accepted 21 December 2013

Available online 28 December 2013

Abstract

A novel cluster validity index whose implementation is based on the membership degrees and improved bipartite modularity of bipartite network is proposed for the validation of partitions produced by the fuzzy c-means (FCM) algorithm. FCM algorithm is employed to group the dataset in order to obtain the membership degree of samples. Then, a weighted bipartite network is constructed by samples and centroids of each cluster. This allows the introduction of a new measurement for optimizing the numbers of clusters for fuzzy partitions. The proposed index utilizes the optimum membership as its global property and the modularity of bipartite network as its local independent property. The proposed index is compared with a number of popular validation indices on fifteen datasets. The experimental results show that the effectiveness and reliability of the proposal is superior to other indices.

Keywords: Fuzzy clustering; Cluster validity index; Bipartite modularity; FCM algorithm

1. Introduction

Clustering has become very important in areas like data mining, pattern recognition, engineering and so on. The purpose of clustering is to divide a given data set into groups (clusters), such that all data in the same group are similar to each other, while data from different clusters are dissimilar. Since clustering is an unsupervised classification process, it has no priori information of data set. A wide variety of clustering algorithms have been proposed in the past decades. Generally speaking, these algorithms can be divided into two classes: hard (crisp) cluster and soft (fuzzy) cluster. Hard clustering algorithms are based on classical set theory and require that a datum either does or does not belong to a cluster. Fuzzy clustering algorithms allow objects to belong to several clusters simultaneously, with different degrees of membership. The results of these clustering algorithms, however, depend on input parameters. For instance, c-means and FCM algorithms require a cluster number c to be predefined. In this case, the question is: What is the optimal cluster number? To answer this question, currently, cluster validity indices research has drawn

[*] Corresponding author at: School of Urban and Environmental Science, Liaoning Normal University, Liaoning, Dalian 116029, PR China.
E-mail address: Xiefd@lnnu.edu.cn (F. Xie).

http://dx.doi.org/10.1016/j.fss.2013.12.013

considerable attention in data mining. Many different cluster validity indices have been defined without any prior class knowledge. In fact, clustering validation is a technique to find a set of clusters that best fits natural partitions without any class information. Up to now, no popular index is known to be suitable for all data sets. Thus, how to define a good validity index to detect a optimal cluster number c is still a challenging problem.

Based on FCM algorithm and the modularity of weighted bipartite network, we introduce a novel cluster validity index to find a proper cluster number. To achieve this goal, the problem of clustering data is converted into detecting the community structures of bipartite network. Afterwards, the modularity of bipartite network is improved. Then we apply the improved modularity successfully to evaluate the quality of the partitions. In next section, we present different cluster validity indices after introducing the fuzzy clustering. Section 3 briefly describes the bipartite network and modularity function. The novel cluster index is proposed in Section 4. Section 5 validates the proposed index. Finally, conclusions and remarks are presented in Section 6.

2. Background

2.1. The fuzzy c-means clustering algorithm

The objective of fuzzy clustering is to partition a data set into c distinct clusters. The well-known fuzzy c-means algorithm proposed by Dunn [1], then extended by Bezdek [2] and its various variations are probably the most commonly used fuzzy clustering methods.

Let $X = \{x_1, \dots, x_n\}$ be a n points data set in a P-dimensional feature space R^P, $X \subset R^P$. The FCM clustering algorithm partitions X into $1 < C < n$ fuzzy groups by minimizing objective function J_m which is the weighted sum of squared errors within groups and is defined as follows:

$$J_m(U, V, X) = \sum_{j=1}^{n}\sum_{i=1}^{C} u_{ij}^m \|x_j - v_i\|_A^2, \quad 1 < m < \infty, \tag{1}$$

and subject to

$$u_{ij} \in [0, 1], \quad \sum_{i=1}^{C} u_{ij} = 1,$$

where $U = [u_{ij}]_{C\times n}$ is a fuzzy partition matrix composed of the membership degree of data point x_j to ith cluster; $V = (v_1, v_2, \dots, v_c)$ is a vector of unknown cluster prototype (centers), $v_i \in R^P$. Norm matrix A defines a measure of similarity between a data point and the cluster prototypes. The parameter m controls the fuzziness of membership of each datum. The cluster centers and the respective membership functions, which are solutions of the constrained optimization problem in Eq. (1), are given by the following equations:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - v_i\|_A}{\|x_j - v_k\|_A}\right)^{\frac{2}{m-1}}}, \quad 1 \leqslant i \leqslant C,\ 1 \leqslant j \leqslant n, \tag{2}$$

and

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \quad 1 \leqslant i \leqslant C. \tag{3}$$

Eqs. (2) and (3) constitute an iterative optimization procedure.

The FCM algorithm is executed in the following steps:

Step 1: Given a preselected C clustering centers set V and fuzzy factor m ($m > 1$), initialize the fuzzy partition matrix U as Eq. (2).

Step 2: Calculate the fuzzy clustering centroid matrix V by Eq. (3).

Step 3: Use Eq. (2) to update the fuzzy membership matrix U.

Step 4: If the improvement in $J_m(U, V, X)$ is less than a certain threshold (ε), then stop; otherwise go to Step 2.

The FCM algorithm detects clusters that have centroid prototypes of a roughly same size. The Gustafson–Kessel (GK) algorithm is an extension of the FCM, which can detect cluster of different orientations and shapes in a data set by employing norm-introducing matrix for each cluster [3]. In GK algorithm, they extended FCM for an inter product matrix form:

$$\|x_j - v_i\|_A^2 = (x_j - v_i)^T M_i (x_j - v_i),$$

where M_i is a positive definite matrix, and is adapted according to the actual shape of the individual clusters, described approximately by the cluster covariance matrices F_i:

$$F_i = \frac{\sum_{j=1}^n u_{ij}^m (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^n u_{ij}}, \qquad M_i = \det(F_i)^{\frac{1}{n}} F_i^{-1}.$$

These two algorithms require users to previously set the cluster number c. However, it is impossible to know the cluster number in advance. Different fuzzy partitions are obtained for different values of c. Thus, it is necessary to validate each of the fuzzy c-partitions once they are created. The problem of finding an optimal c is usually called cluster validity. In next subsection, we will review several popular validity indices.

2.2. Previous cluster validity indices

Validity indices are extremely important for automatically determining the number of clusters. In the past decades, a wide variety of cluster validity indices have been developed for evaluating quality of partitions, aiming at finding an optimal partitioning that best fits natural partitions for the given data set [4–10]. Most of the proposed indices have focused on two properties: compactness and separation. Compactness is used as a measure of the variation or scattering of the data within a particular cluster, while separation indicates the isolation of the clusters from each other. Some indices present monotonicity that flow the variation of cluster numbers on most real datasets. Thus, those indices maybe invalid for large values c. In what follows, we list some popular indices.

The first two fuzzy cluster validity functions associated with FCM algorithm are the partition coefficient and partition entropy.

(i) The partition coefficient PC and partition entropy PE [11,12]

$$PC = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{C} u_{ij}^2, \tag{4}$$

$$PE = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{C} u_{ij} \log(u_{ij}). \tag{5}$$

(ii) Both PC and PE possess monotonic evolution tendency with c [13]. Modification of the MPC index and MPE index proposed by Dave [13] can reduce the monotonic tendency and are defined respectively as

$$MPE = \frac{n \times I_{PE}}{n - C}, \tag{6}$$

$$MPC = \frac{C \times I_{PC} - 1}{C - 1}. \tag{7}$$

(iii) Davies–Bouldin index (DB) [14]

$$DB = \frac{1}{C} \sum_{i=1}^{C} \max_{i \neq j} \frac{s_i + s_j}{\|c_i - c_j\|}, \qquad s_i = \left(\frac{1}{card(v_i)} \sum_{x \in v_i} \|x - c_i\|^2 \right)^{1/2}, \tag{8}$$

where c_i denotes the centroid of cluster v_i.

(iv) Fukuyama–Sugeno index (FS) [15]

$$FS = J_m(C, X, U, V) - K_m(C, U, V), \qquad K_m(C, U, V) = \sum_{j=1}^{n} \sum_{i=1}^{C} u_{ij}^m \|c_i - \bar{c}\|^2, \tag{9}$$

where $\bar{c} = \frac{1}{n} \sum_{j=1}^n x_j$.

(v) Fuzzy Hyper Volume (*FHV*) [16]

$$FHV=\sum_{i=1}^{C}V_i,\quad V_i=|\Sigma_i|^{1/2},\ \Sigma_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}(x_j-c_i)(x_j-c_i)^T}{\sum_{j=1}^{N}u_{ij}^{m}}. \tag{10}$$

(vi) Xie–Beni index (*XB*) [17]

$$XB=\frac{J_m}{n\min_{i\neq j}\|c_i-c_j\|^2}, \tag{11}$$

where J_m is defined as Eq. (1) with Euclidean distance.

(vii) Partition Coefficient And Exponential Separation (*PCAES*) [18]

$$PCAES=\sum_{i=1}^{C}\left(\sum_{j=1}^{n}u_{ij}^2/u_M-\exp\left(-\min_{k\neq i}\left\{\|c_i-c_k\|^2\right\}/\beta_T\right)\right) \tag{12}$$

where $u_M=\min_{1\leqslant i\leqslant C}\{\sum_{j=1}^{n}u_{ij}^2\}$, and $\beta_T=\frac{\sum_{l=1}^{C}\|c_l-\bar{c}\|^2}{C}$.

(viii) PBM-index for Fuzzy c-means (*PBMF*) [19]

$$PBMF=\left(\frac{1}{C}\times\frac{E_1}{J_m'}\times D_c\right)^2, \tag{13}$$

where C is the number of clusters, $E_1=\sum_{j=1}^{n}u_{ij}\|x_j-\bar{c}\|$, $\bar{c}$ being the centroid of data set, and $D_c=\max_{i,j=1}^{C}\|c_i-c_j\|$, with J_m' defined by $J_m'=\sum_{j=1}^{n}\sum_{i=1}^{C}(u_{ij})^m\|x_j-c_i\|$, $m=1.5$ is considered in [19].

(ix) Rezaee Compactness and Separation (*RSC*) [20]

$$RSC=Sep^N(C,U)+Comp^N(C,X,U,V), \tag{14}$$

where $Sep^N(C,U)=\frac{Sep(C,U)}{\max_{2\leqslant C\leqslant C_{\max}}Sep(C,U)}$ and $Comp^N(C,X,U,V)=\frac{Comp(C,X,U,V)}{\max_{2\leqslant C\leqslant C_{\max}}Comp(C,X,U,V)}$. Furthermore $Sep(C,U)=\frac{2}{C\times(C-1)}\sum_{p=1}^{C-1}\sum_{q=p+1}^{C}\sum_{j=1}^{n}(\min(u_{pj},u_{qj})\times(-\sum_{i=1}^{C}u_{ij}\log u_{ij}))$ and $Comp(C,X,U,V)=\sum_{i=1}^{C}\sum_{j=1}^{n}u_{ij}^2\|x_j-c_i\|^2$.

In (i)–(ix), an optimal c can be found by solving the maximum or the minimum of the corresponding objective functions.

3. The bipartite network and bipartite modularity

Recently, complex networks have attracted considerable attention in physics, computer science and other fields. It can be considered as the foundation of the mathematical representation for a variety of complex systems, such as biological and social systems, the Internet, the worldwide web, and many others. Among them, the bipartite network is an important kind of complex network. In fact, many real-world networks are naturally bipartite, such as the actors–films network [21], the papers–scientists network [22], and so on. In bipartite network, there are two non-overlapping sets of nodes called top nodes and bottom nodes. The edges only connect pairs vertices which belong to different sets.

A common feature of many bipartite networks is "community structure", the tendency for vertices to be divided into groups, with dense connections within groups and only sparser connections between them. Thus, of great current interest is the identification of the community structure of the bipartite network. Detecting communities allows quantitative investigation of relevant subnetworks, which have different properties from the aggregate properties of the network as a whole. Informally, a community in bipartite network is a subgraph in which vertices are more likely to be connected to each other than to those outside the subgraph.

The bipartite modularity is a quantitative measurement for the quality of a particular division of a bipartite network. Guimera [23] proposed a projection-based method. He transforms the bipartite network into one-mode network and uses a method for one-mode network to discover communities. However, the projection process is usually considered to cause information loss which leads to a bad result. Barber [24] extended Newman's modularity to bipartite network and proposed a method to find communities by minimizing the modularity, but his bipartition-based method is based on an assumption that the number of communities is specified in advance. These two bipartite modularities are, however, not sufficient for evaluating the degree of correspondence between communities of different vertex types, which is often important to understand the characteristics of the communities. Very recently, T. Murata proposed a new bipartite modularity for evaluating community extracted from bipartite networks. Experimental results show that this bipartite modularity is appropriate for discovering communities of many bipartite networks.

The bipartite modularity defined by Murata [25] is described as follows.

Suppose G is a bipartite network whose vertices V are divided into two disjoint sets V^+ and V^-, such that every edge in the network connects a pair of nodes from V^+ and V^- respectively. We denote $G = (V^+, V^-, E)$ to a bipartite network and call V^+ and V^- top nodes and bottom nodes respectively. Conveniently, X-vertices $\{x_0, x_1, \dots, x_n\}$ stands for the top nodes and $\{y_0, y_1, \dots, y_p\}$ for Y-vertices the bottom nodes.

Suppose that M is the number of edges in a bipartite network. Consider a particular division of the bipartite network into X-vertex communities and Y-vertex communities. The numbers of the communities are L^+ and L^- respectively. V_i^+ and V_j^- are the individual community that belong to the sets $(V^+ = \{V_1^+, V_2^+, \dots, V_{L^+}^+\},\ V^- = \{V_1^-, V_2^-, \dots, V_{L^-}^-\})$. $A(i, j)$ is an element of its adjacency matrix. $A(i, j)$ is equal to 1 if vertices i and j are connected and 0 otherwise.

Assume that the vertices of V_l and V_m are different types (which means $(V_l \in V^+ \wedge V_s \in V^-) \vee (V_l \in V^- \wedge V_s \in V^+)$), e_{lm} (the fraction of all edges that connect vertices in V_l to those in V_s) and a_i (its row sums) can be defined as follows:

$$e_{lm} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V_s} A(i, j), \tag{15}$$

$$a_i = \sum_j e_{ij} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V} A(i, j). \tag{16}$$

The bipartite modularity Q_B is defined as:

$$Q_B = \sum_i (e_{ij} - a_i a_j), \quad j = \max_k (e_{ik}). \tag{17}$$

High Q_B value indicates strong community structure in a bipartite network.

4. The proposed validity index

In this section, we define a novel validity index to validate the partitions produced by the fuzzy c-means algorithm.

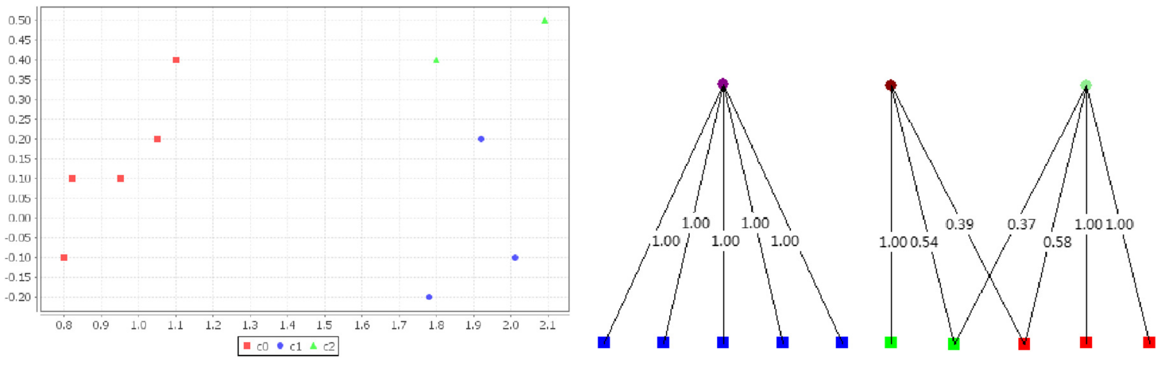
For a given data set, we first employ the FCM algorithm with a proper distance norm to partition it into c $(c \geqslant 2)$ clusters. Once we obtain the matrix of membership degree, then a weighted bipartite network is constructed whose top nodes consist of all clustering centroids and bottom nodes are made up of all samples. The membership degrees can be considered as its weighted edges. Obviously, it is improper to directly create a bipartite network in this way because that results in a fully-connected bipartite network and expensive computation.

To deal with this problem and extend the bipartite modularity Q_B to the weighted bipartite network, we would like to take the following strategy:

$$A(i, j) = \delta(u_{ij}) = \begin{cases} 1.0 & u_{ij} > \alpha, \\ u_{ij} & (1 - \alpha) \leqslant u_{ij} \leqslant \alpha, \\ 0.0 & u_{ij} < (1 - \alpha), \end{cases} \tag{18}$$

where one takes $\alpha = 0.7$ throughout this paper. Generally, it should be noted that $\alpha > 0.5$ is assumed.

Example. As shown in Fig. 1(a), if we partition the data set by FCM algorithm for $c = 3$, a matrix of membership degree is obtain. In terms of the idea mentioned above and Eq. (18), we can easily construct a weighted bipartite network. The top nodes consist of three clustering centers and bottom nodes are made up of ten samples. The number in Fig. 1(b) stands for the weight on each edge in network.



(a)

(b)

Fig. 1. The data set and a bipartite network.

If we define

$$L^{+} = L^{-} = C, \qquad V_i^{+} = \{v_i\}, \qquad V_j^{-} = \Big\{x_i \;\Big|\; u_{ji} = \max_k(u_{ki})\Big\}, \qquad M = \sum_{i=1}^{C}\sum_{j=1}^{n} A(i, j), \tag{19}$$

the bipartite modularity Q_B can be successfully extended to the weighted bipartite network.

To avoid the disadvantage of achieving the local maximum or minimum of the proposal, it is necessary to introduce a global quantity in the proposed index. Sine the mean maximum membership degree (MMD) is related to global information, we will integrate it into the novel index.

MMD is defined as follows

$$MMD = \frac{1}{n}\sum_{j=1}^{n} \max_{1\leqslant i\leqslant C} u_{ij}. \tag{20}$$

An improved bipartite modularity Q'_B can be achieved by Combining Eqs. (15)–(19), one can get. Based on this bipartite modularity Q'_B and *MMD*, we now propose a novel weighted global–local based validity index (*WGLI*):

$$WGLI = \big(2MMD + Q'_B\big)/3. \tag{21}$$

The proposed index consider the local information and global information, thus it allows us to find an optimal cluster number. The bigger the value of *WGLI* is, the better the cluster result is. The maximum value of the index *WGLI* over different fuzzy partitions indicates the appropriate cluster. The value of *WGLI* is varied in the range [0, 1].

In what follows, we would like to describe an algorithm to partitioning a data set into proper groups.

Input: Data set S; threshold ε (used in FCM algorithm).

Output: The clustering result.

Step 1: Execute FCM algorithm for the given data set.

Step 2: Construct a weighted bipartite network by matrix of membership degree U and clustering centroids v_i according to Eq. (18) and Eq. (19).

Step 3: Calculate the bipartite modularity Q'_B and *MMD* in terms of Eqs. (15)–(20).

Step 4: Compute index *WGLI* in Eq. (21).

From $c = 2$ to $c_{\max}$, we repeat the above process. Among these values, it is easy to find the maximum value of *WGLI*. Generally, $C_{\max}$ is unknown. Usually, $C_{\max} = \ln(n)$ or $C_{\max} = \sqrt{n}$ is adopted although they have not been proved up to now. In this paper, we will adopt Bezdek's suggestion: $C_{\min} = 2$ and $C_{\max} = \sqrt{n}$.

5. Experiments

To validate our proposal, we compare the performance of the proposed index with some other indices on six artificial data sets and nine well-known real data sets. Experimental environment is Eclipse RCP on Windows7 with

Intel(R) Core(TM)2 Duo CPU T9550@2.66 GHz 2.67 GHz and 4.00 GB memory. The programming language is Java. The fuzzifier m in the proposed algorithm is set to 2 in all experiments.

5.1. Evaluation of clustering result

To assess the quality of partition, the F-measure and P-measure are used to quantify the cluster results. The F-measure is a harmonic combination of the precision and recall values used in information retrieval [35].

If n_i is the number of the members of class i, and n_{ij} is the number of the members of class i in cluster j, then the precision P_{ij} and recall R_{ij} can be defined as:

$$P_{ij} = \frac{n_{ij}}{n_j}, \qquad R_{ij} = \frac{n_{ij}}{n_i}. \tag{22}$$

The F_{ij} is denoted by:

$$F_{ij} = \frac{2 \times P_{ij} \times R_{ij}}{P_{ij} + R_{ij}}. \tag{23}$$

The corresponding F-measure (FM) of the whole clustering result is defined as:

$$FM = \sum_i \frac{n_i}{N} \max_j F_{ij}, \tag{24}$$

where N is the total number of the members in the data set.

In general, high values of F-measure indicate better cluster results.

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of data assigned to that cluster, thus the purity of cluster j is defined as:

$$P_j = \frac{1}{n_j} \max_i n_{ij}. \tag{25}$$

The purity of the whole clustering result is defined as:

$$PM = \sum_j \frac{n_j}{N} P_j. \tag{26}$$

In general, the larger the purity value is, the better the clustering result is.

5.2. Examination on data sets

In this subsection, we would like to select six artificial data sets and nine well-known real data sets for testing the introduced cluster validity index *WGLI*. Meanwhile, we also compare it with a number of popular validation indices on these data sets. The six artificial data sets are called data_3_2, spherical_4_3, spherical _5_2, spherical_6_2, elliptical_10_2 and st900_9_2. The numbers in the title of each dataset imply the numbers of clusters and dimensions respectively. For example, there are three clusters and the dimension of the data is 2 in the data_3_2 data set. The data set st900_9_2 denotes that there are 900 samples in this data set. As it indicates, the number of clusters ranges from two to ten. The nine well-know real data sets are iris, wine, WBCD (Wisconsin Breast Cancer Database), BUPA, Haberman, ionosphere, Pima (pima–indians–diabetes), WPBC and glass.

It is well known that the pre-selection of c clustering centroids will generally impact on the clustering result obtained by FCM algorithm. To tackle this problem, we randomly select 10 groups of initial values to repeatedly execute the FCM algorithm for each pre-selected cluster number c on the fifteen data sets, which means that FCM algorithm will run 10 times for the same cluster number c with the different initial values. The run times (RTs) from 10 to 19 correspond to the same cluster number 3. The RTs listed in the first column in Tables 2–16 imply that at least one of the indices have achieved its maximum or minimum value. To evaluate the quality of partition, we here employ the F-measure and P-measure to quantify the cluster results on the tested fifteen data sets.

In Table 1, we compare the original cluster numbers (OC) with the cluster numbers c found by *WGLI* and other eight indices on fifteen data sets. The bold type numbers in Table 1 indicate that the cluster numbers found by the

Table 1
The comparison of the original cluster number OC with those found by the proposal and other eight indices on fifteen data sets.

Data set	*OC*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*
data_3_2.data	3	**3**	2	2	2	**3**	**3**	4	8	2
spherical_4_3.data	4	**4**	2	2	2	**4**	**4**	6	**4**	**3**
spherical_5_2.data	5	**5**	2	2	2	4	**5**	**5**	**5**	2
spherical_6_2.data	6	**6**	**6**	**6**	**6**	**6**	4	8	**6**	2
elliptical_10_2.data	10	**10**	2	2	2	**10**	**10**	13	22	2
st900_9_2.data	9	**4/9**	2	2	2	**9**	**9**	8	**9**	7
iris.data	3	**3**	2	2	2	2	2	7	12	2
wine.data	3	**3**	2	2	2	2	2	4	13	2
WBDC.data	2	**2**	**2**	**2**	**2**	**2**	**2**	3	**2**	**2**
BUPA.data	2	**2**	**2**	**2**	**2**	**2**	**2**	**2**	18	**2**
Haberman.data	2	**2**	**2**	**2**	**2**	**2**	4	4	**2**	**2**
ionosphere.data	2	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Pima.data	2	3	**2**	**2**	**2**	**2**	3	4	24	**2**
WPBC.data	2	4	**2**	**2**	**2**	**2**	4	5	**2**	**2**
glass.data	6	2	2	2	2	3	3	3	2	2

indices are the same as the original. It is easy to see that we can successfully detect the optimal cluster numbers c by the introduced index *WGLI* in most cases except the last three data sets, which indicates that the proposed index can not determine successfully the correct cluster number for all data sets. We can also see that the optimal cluster number c for data set glass is not found by all indices listed here, partially due to its complex data structure or maybe the application of improper similarity function. In Table 1, the correct cluster number c by *WGLI* and *MPC* can be found easily on most data sets. However, the indices *PE*, *PC*, *MPE*, *MPC* and *XB* have good performances on most real data sets.

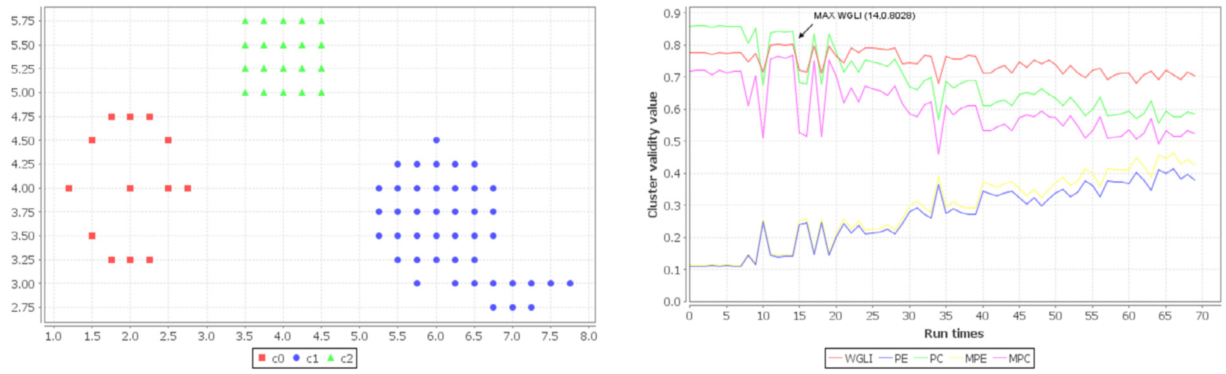
5.2.1. Six artificial data sets

In this subsection, six artificial datasets [26,27] are used as test sets for comparisons of validity indices. The readers can also browse the web page http://www.isical.ac.in/~sanghami/data.html to find the information in detail.

The dentate curves of five indices in Figs. 2–7 show the relationship between the results produced by FCM algorithm and the initial values. It shows that the different initial values will lead to different index values. Fortunately, optimal cluster numbers c are unchanged with the variation of the initial values. The fifth row in Table 2 demonstrates that the proposed index *WGLI* achieves its maximum value when we take the fifth group of initial values for $c = 3$. Meanwhile, the other indices *MPC* and *DB* also gain their maximum or minimum values, respectively. Therefore, the optimal cluster number 3 is obtained by these three indices. Table 3 and Fig. 3 describe the case when we apply the proposed algorithm and index *WGLI* on data set spherical_4_3. The values of *FM* and *PM* in Table 2 and Table 3 account for the fact that we can entirely group them and discover the optimal numbers c. Shown from the left panel (a) in Fig. 4, it is not easy to partition data set spherical_5_2 into five clusters because there is no clear border among groups. This fact has also been proved by panel (b) in Fig. 4 and Table 4 since there is no distinct difference between values of indices *WGLI*, *MPC*, *FS* and *DB* while $c = 4$ and $c = 5$. In spite of this, the optimal cluster number $c = 5$ is still obtained by four indices including the proposal. The spherical_6_2 and elliptical_10_2 are demonstrated in Fig. 5(a) and Fig. 6(a). These two data sets are widely used to test the validity of the new cluster index because there are some groups to be near each other. The cluster number $c = 6$ for spherical_6_2 and $c = 10$ for elliptical_10_2 are successfully detected by six indices and three indices respectively. The values (1 and 0.996) of *FM* in Table 5 and Table 6 show that the clustering results are good enough. The data set st.900_9_2 includes 900 samples and has strong overlap among clusters shown in Fig. 7. 4 clusters on this data set are found by index *WGLI* and this number is obviously not well. Although the correct number 9 can be found by the second big value (0.6750) described in Table 7, this is not always right for other data sets because we generally do not know the optimal cluster number in advance. Maybe, this result gives us a hint that the better partition can be found by the second big or small index value.

Table 2
The cluster numbers c found by nine indices and its corresponding index value for data_3_2 data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7767	**0.1091**	0.8605	**0.1120**	0.7210	0.7521	−135.8481	0.8795	0.0002	0.8611	0.8289
3	2	0.7720	0.1132	0.8540	0.1163	0.7080	0.6423	−172.9437	0.8594	**0.0001**	0.8611	0.8289
4	2	0.7772	0.1092	**0.8609**	0.1122	0.7218	0.6739	−159.7721	0.8536	0.0002	0.8611	0.8289
12	3	0.8019	0.1400	0.8436	0.1458	0.7654	**0.4615**	−254.8700	0.6947	0.0003	1.0000	1.0000
14	3	**0.8028**	0.1408	0.8444	0.1466	**0.7666**	0.5018	−234.9998	0.7101	0.0004	1.0000	1.0000
20	4	0.7647	0.2013	0.7762	0.2125	0.7016	0.7185	**−305.1515**	0.7789	0.0024	0.9487	1.0000
60	8	0.7136	0.3681	0.5937	0.4114	0.5357	0.9906	−174.3500	**0.6141**	0.0173	0.5685	1.0000



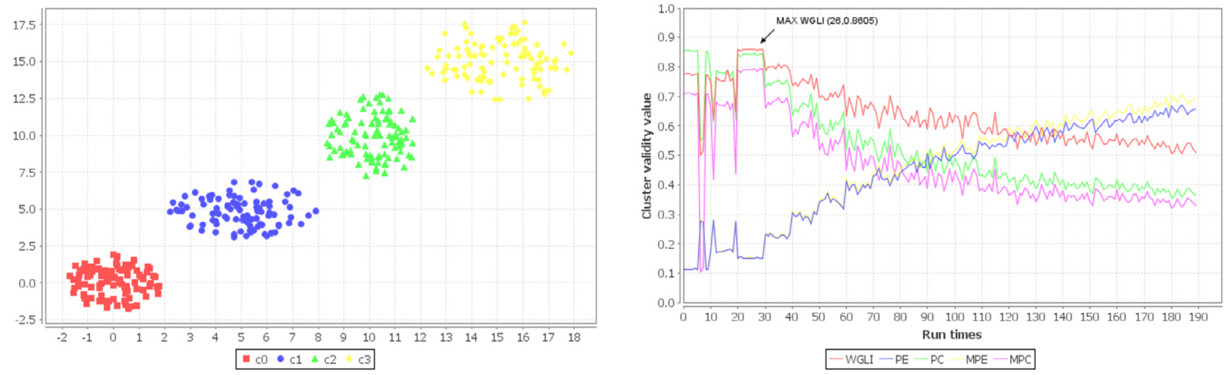
(a) Data_3_2.

(b) The change of the values of five indices with run times for data_3_2.

Fig. 2. Data_3_2 and the comparison of *WGLI* with the other four indices.

Table 3
The optimal cluster c found by nine indices and its corresponding index value for Data set spherical_4_3.

RTC	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
2	2	0.7766	**0.1128**	**0.8552**	**0.1134**	0.7104	0.5170	−29 530.1782	11.3003	0.0000	0.6667	0.5000
26	4	**0.8605**	0.1496	0.8459	0.1511	**0.7945**	0.4629	−32 251.8495	6.2289	**0.0000**	1.0000	1.0000
28	4	0.8599	0.1500	0.8449	0.1515	0.7932	**0.4610**	−32 358.6408	**6.2245**	0.0000	1.0000	1.0000
47	6	0.7757	0.2680	0.7091	0.2721	0.6509	1.1428	**−36 100.3970**	8.7235	0.0002	0.8824	1.0000



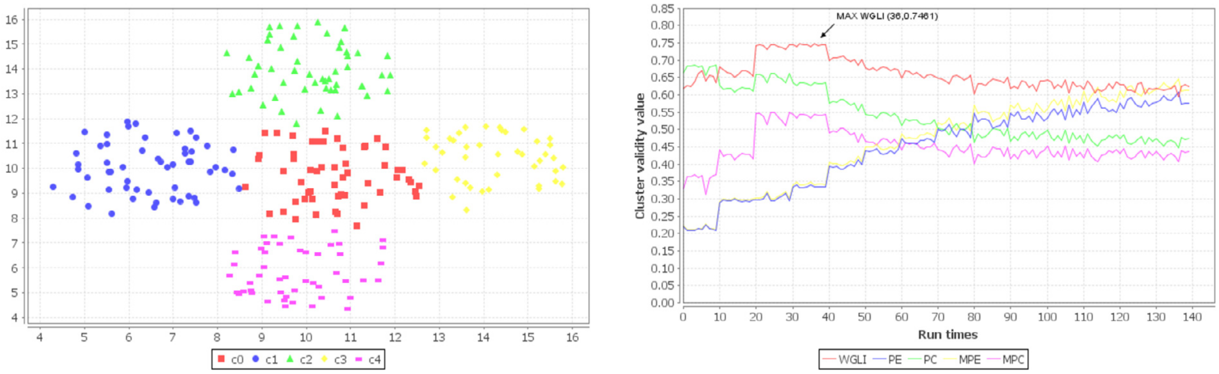
(a) Spherical_4_3.

(b) The change of the values of five indices with run times for spherical_4_3.

Fig. 3. Data set spherical_4_3 and the comparison of *WGLI* with the other four indices.

Table 4
The cluster numbers *c* found by nine indices and its corresponding index value for spherical_5_2 data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.6257	0.2096	0.6825	0.2113	0.3650	1.3719	−834.2000	8.9870	**0.0000**	0.4863	0.4000
3	2	0.6360	**0.2095**	**0.6840**	**0.2112**	0.3680	1.3439	−849.9728	8.8609	0.0000	0.4993	0.4000
24	4	0.7441	0.2943	0.6623	0.2991	**0.5497**	0.7095	−2125.6824	5.4290	0.0001	0.7604	0.8000
32	5	0.7460	0.3325	0.6352	0.3393	0.5440	0.7061	**−2137.9079**	5.1881	0.0001	0.9444	0.9440
33	5	0.7435	0.3325	0.6350	0.3393	0.5438	0.7080	−2128.1266	**5.1842**	0.0001	0.9445	0.9440
36	5	**0.7461**	0.3331	0.6346	0.3399	0.5432	0.7079	−2110.8827	5.1950	0.0001	0.9403	0.9400
38	5	0.7433	0.3347	0.6328	0.3415	0.5410	**0.7035**	−2118.0582	5.2241	0.0001	0.9680	0.9680



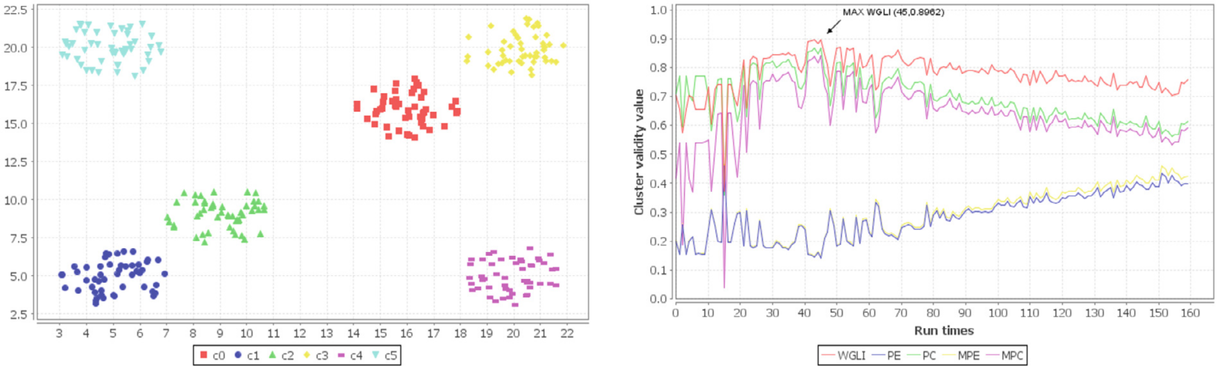
(a) Spherical_5_2.

(b) The change of the values of five indices with run times for spherical_5_2.

Fig. 4. Data set spherical_5_2 and the comparison of *WGLI* with the other four indices.

Table 5
The cluster numbers *c* found by nine indices and its corresponding index value for spherical_6_2 data set.

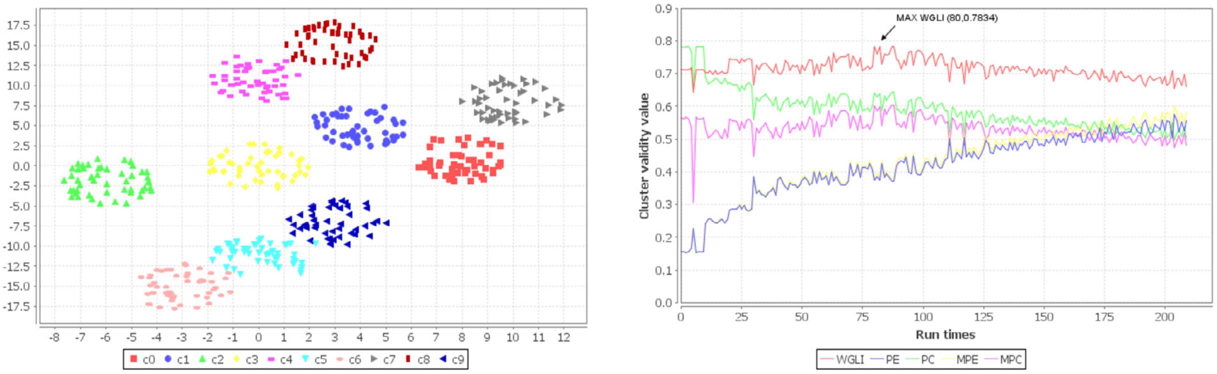
RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7047	0.1991	0.7094	0.2004	0.4188	1.2906	−5968.8520	44.7270	**0.0000**	0.5000	0.3333
28	4	0.8320	0.1754	0.8150	0.1778	0.7533	**0.3753**	−20 298.6105	9.6881	0.0000	0.7778	0.6667
45	6	**0.8962**	**0.1414**	**0.8677**	**0.1443**	**0.8412**	0.3831	−22 062.4865	**5.5364**	0.0000	1.0000	1.0000
69	8	0.8606	0.2061	0.7968	0.2117	0.7678	0.6459	**−22 442.6663**	6.1443	0.0003	0.9216	1.0000



(a) Spherical_6_2.

(b) The change of the values of five indices with run times for spherical_6_2.

Fig. 5. Data set spherical_6_2 and the comparison of *WGLI* with the other four indices.



(a) Elliptical_10_2.

(b) The change of the values of five indices with run times for elliptical_10_2.

Fig. 6. Data set elliptical_10_2 and the comparison of *WGLI* with the other four indices.

Table 6
The cluster numbers c found by nine indices and its corresponding index value for elliptical_10_2 data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7104	0.1550	0.7817	0.1556	0.5634	0.7770	−27 595.6543	39.7042	**0.0000**	0.3211	0.2000
6	2	0.7105	**0.1534**	**0.7836**	**0.1540**	0.5672	0.7720	−28 002.7688	39.5687	0.0000	0.3181	0.2000
80	10	**0.7834**	0.3730	0.6401	0.3806	0.6001	0.6385	−34 933.2356	17.1382	0.0000	0.9960	0.9960
88	10	0.7825	0.3712	0.6429	0.3788	**0.6032**	**0.6168**	−35 493.3529	17.0918	0.0000	0.9739	0.9740
110	13	0.7281	0.4430	0.5880	0.4548	0.5537	0.8240	**−38 833.2169**	19.8859	0.0002	0.8137	0.8680
203	22	0.7083	0.5033	0.5508	0.5265	0.5294	0.9050	−30 971.4077	**15.4450**	0.0004	0.6898	0.9880

Table 7
The cluster numbers c found by nine indices and its corresponding index value for st900_9_2 data set.

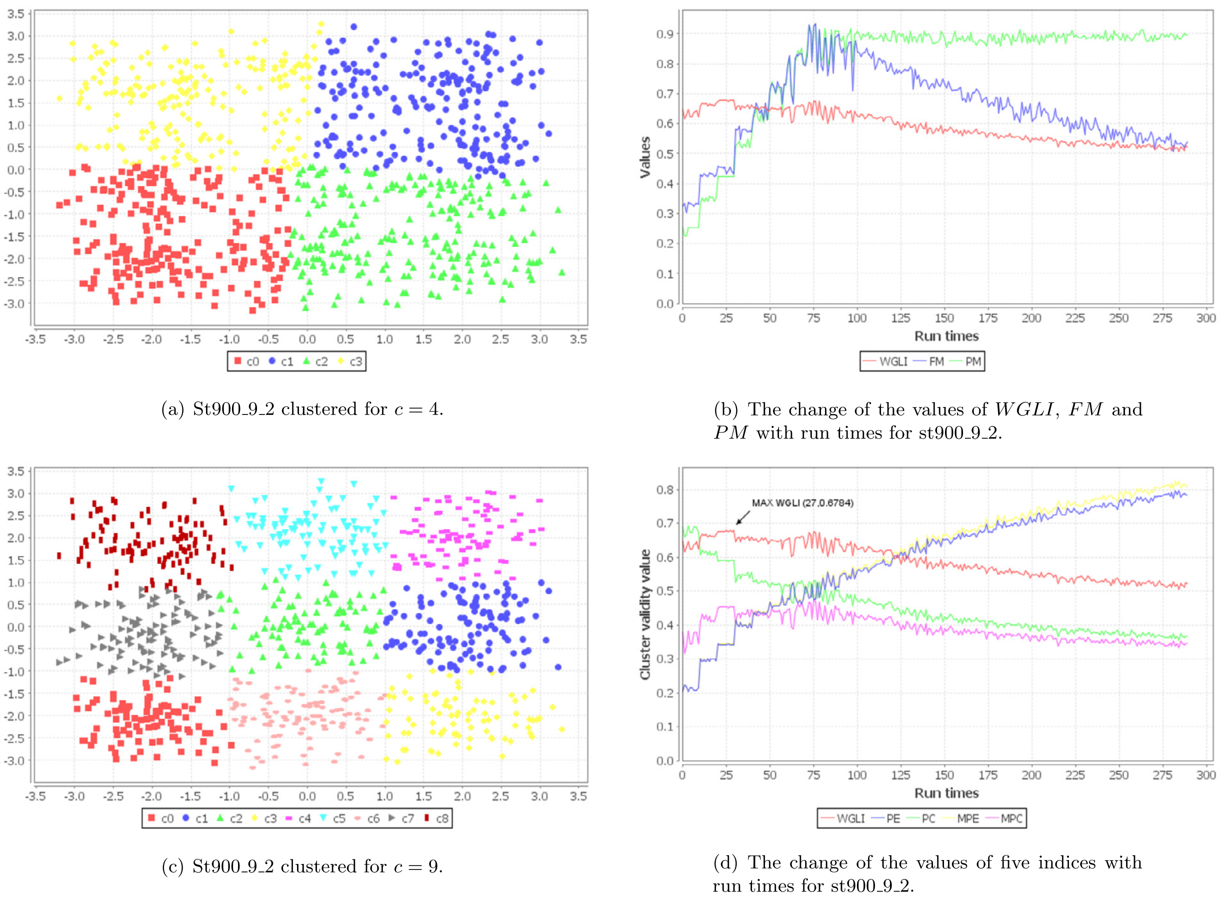
RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
3	2	0.6456	**0.2051**	**0.6915**	**0.2056**	0.3830	1.3140	−1297.8018	3.4819	0.0228	0.3366	0.2522
27	4	**0.6784**	0.3428	0.5900	0.3443	0.4533	0.8163	−2516.4990	3.0741	0.0047	0.4548	0.4244
53	7	0.6495	0.4576	0.5206	0.4612	0.4407	0.9242	−2493.3794	2.8738	**0.0002**	0.7020	0.7000
68	8	0.6628	0.4626	0.5306	0.4667	0.4635	0.8559	**−2715.8880**	2.5512	0.0054	0.8313	0.8500
75	9	0.6750	0.4762	0.5309	0.4810	**0.4723**	**0.7347**	−2691.6125	**2.3974**	0.0159	0.9266	0.9267
289	30	0.5217	0.7815	0.3682	0.8084	0.3464	1.2062	−1988.9565	2.8726	0.0028	0.5397	0.8933

5.2.2. Nine real data sets

To demonstrate the effectiveness of the proposed index, we also compare *WGLI* with other indices on nine real data sets. For each of c $(2, 3, \ldots, \sqrt{n})$, we execute the FCM algorithm ten times with different initial values on every data sets. For simplicity, we only list the optimal cluster numbers c found by nine indices and its corresponding index value in following tables.

The Iris data set [28] has 150 samples in four-dimensional space that represent three physical clusters. On this data set, two of the three clusters are hardly discernable while the third one is well separated from the other two. Table 8 shows that the optimal cluster number $c = 3$ is only found by the proposed index. Most of indices recognize $c = 2$ for this data set. The high values of *FM* (0.8918) indicates that the obtained partition is good enough. Meanwhile most of techniques reported in the literature usually provide two clusters for this data set [29]. Without any class information, the researchers usually argue that there are two, not three clusters.

The Wine data set with three clusters includes 178 samples in 13-dimensional space. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determines the quantities of 13 constituents found in each of the three types of wines [30]. The other eight indices



(a) St900_9_2 clustered for $c = 4$.

(b) The change of the values of $WGLI$, FM and PM with run times for st900_9_2.

(c) St900_9_2 clustered for $c = 9$.

(d) The change of the values of five indices with run times for st900_9_2.

Fig. 7. Two cases of clustering data set st900_9_2 (a), (c), the changes of the values of *FM* and *PM* with *RT* (b) and the comparison of *WGLI* with the other four indices (d).

Table 8
The cluster numbers c found by nine indices and its corresponding index value for iris data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7680	0.0851	0.8920	0.0862	0.7840	**0.4690**	−530.0622	0.0216	**0.0000**	0.7635	0.6667
2	2	0.7692	**0.0840**	**0.8942**	**0.0851**	**0.7884**	0.4736	−519.8439	0.0215	0.0000	0.7635	0.6667
16	3	**0.7797**	0.1722	0.7835	0.1757	0.6752	0.7678	−518.4038	0.0207	0.0004	0.8918	0.8933
59	7	0.6593	0.4087	0.5636	0.4287	0.4909	1.3262	**−595.0651**	0.0249	0.0116	0.6944	0.9600
102	12	0.5973	0.5423	0.4501	0.5895	0.4001	1.4156	−334.0411	**0.0160**	0.0182	0.4584	0.9733

fail to detect the optimal cluster number $c = 3$ except the index *WGLI* shown in Table 9. For these two data sets, the correct cluster numbers is successfully detected by the proposed index. However the value of *FM* tells us that near half of samples are put into wrong clusters.

The Wisconsin Breast Cancer Database (WBCD) data set was collected by Dr. William H. Wolberg at the University of Wisconsin Hospitals. There are 699 records in this data set. Each record has nine attributes, which are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. In this database, 241 records are malignant and 458 records are benign [31]. Table 10 tells us that the optimal cluster number $c = 2$ can be detected by eight indices. The value of *FM* (0.9540) shows that not only do we find the correct cluster number but also obtain the ideal cluster result.

The BUPA data set with two clusters consists of 345 samples in 7-dimensional space. The first 5 variables are all blood tests which are considered to be sensitive to liver disorders that might arise from excessive alcohol consumption.

Table 9
The cluster numbers *c* found by nine indices and its corresponding index value for wine data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7003	0.1497	0.7902	0.1514	0.5804	0.8878	−19 036.0304	0.0049	**0.0000**	0.5285	0.5112
2	2	0.6975	**0.1476**	**0.7938**	**0.1493**	**0.5876**	**0.8672**	−20 170.3140	0.0048	0.0000	0.5229	0.5056
11	3	**0.7202**	0.2392	0.6930	0.2433	0.5395	0.9665	−18 819.7797	0.0034	0.0000	0.5065	0.5449
20	4	0.7073	0.2896	0.6481	0.2963	0.5308	0.8976	**−29 958.2791**	0.0021	0.0000	0.4991	0.5506
111	13	0.4744	0.6931	0.3322	0.7477	0.2766	1.6520	−16 385.1983	0.0007	0.0006	0.3864	0.7360
112	13	0.4930	0.6922	0.3396	0.7467	0.2846	1.4981	−14 056.8847	**0.0006**	0.0003	0.4725	0.7809

Table 10
The cluster numbers *c* found by nine indices and its corresponding index value for WBCD data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7235	0.1247	0.8275	0.1251	0.6550	0.8747	−20 035.6833	926.0295	**0.0000**	0.9540	0.9542
2	2	0.7186	**0.1192**	**0.8351**	**0.1195**	**0.6702**	0.8238	−24 166.1988	916.3943	0.0000	0.9347	0.9356
3	2	0.7188	0.1193	0.8351	0.1196	0.6702	**0.8235**	−24 107.3236	916.2646	0.0000	0.9362	0.9371
4	2	0.7198	0.1197	0.8345	0.1200	0.6690	0.8280	−23 657.6724	**915.8268**	0.0000	0.9363	0.9371
5	2	**0.7236**	0.1245	0.8278	0.1249	0.6556	0.8726	−20 131.4721	925.3969	0.0000	0.9540	0.9542
18	3	0.6517	0.2255	0.7106	0.2265	0.5659	2.1641	**−28 385.4531**	1443.7660	0.0000	0.8599	0.9585
245	26	0.2823	0.9665	0.2417	1.0038	0.2114	63.9859	−12 399.0459	7872.1270	1.3275	0.3092	0.9671

Table 11
The cluster numbers *c* found by nine indices and its corresponding index value for BUPA data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.6625	0.1406	0.8046	0.1414	0.6092	1.2217	−388 030.5798	3 807 103.4231	**0.0000**	0.6243	0.5797
4	2	0.6625	**0.1083**	**0.8569**	**0.1089**	**0.7138**	**0.8929**	**−826 094.7966**	4 562 527.1153	0.0000	0.6451	0.5797
8	2	**0.6651**	0.1339	0.8156	0.1347	0.6312	1.1405	−449 215.1833	3 918 502.4305	0.0000	0.6266	0.5797
169	18	0.2672	0.9419	0.1827	0.9937	0.1346	3.2973	−195 455.9879	**1 721 138.9298**	0.0007	0.2466	0.6638

Table 12
The cluster numbers *c* found by nine indices and its corresponding index value for Haberman data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.6824	0.1799	0.7397	0.1811	0.4794	1.1371	−17 624.4305	267.9019	**0.0000**	0.5537	0.7353
6	2	0.6819	**0.1792**	**0.7407**	**0.1804**	**0.4814**	1.1201	−18 016.0197	**267.1881**	0.0000	0.5559	0.7353
8	2	**0.6827**	0.1800	0.7391	0.1812	0.4782	1.1289	−17 796.4454	267.2267	0.0000	0.5448	0.7353
20	4	0.6728	0.3434	0.5871	0.3479	0.4495	**0.9801**	**−23 345.0880**	337.4214	0.0000	0.4564	0.7582
153	17	0.4835	0.7523	0.3316	0.7966	0.2898	1.1596	−22 748.6417	294.6365	0.0001	0.2088	0.7680

Each line in the BUPA data file constitutes the record of a single male individual. In Table 11, we can easily see that all indices except the index *FHV* may find the optimal cluster number $c = 2$ for the first group of initial values.

Haberman's Survival Data [32]: The data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. This data set includes 306 samples in 4-dimensional space and has two clusters. From Table 12, it can be seen that the optimal number $c = 2$ can be found by seven indices. However, the value of *FM* proves that the partition result is not well.

Johns Hopkins University Ionosphere database [33]: The radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kW. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. This data set consists of 351 samples in 34-dimensional space and has two clusters. For this data set, all indices in Table 13 can determine the optimal cluster number.

Table 13
The cluster numbers *c* found by nine indices and its corresponding index value for ionosphere data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.5943	0.2299	0.6448	0.2312	0.2896	2.2166	−316.0723	**0.0000**	0.0001	0.7122	0.7066
5	2	0.4646	**0.2072**	**0.6844**	**0.2084**	**0.3688**	**0.8301**	**−1148.2816**	0.0000	**0.0000**	0.6902	0.6439
9	2	**0.6029**	0.2240	0.6551	0.2253	0.3102	2.0265	−379.4385	0.0000	0.0001	0.7095	0.7037
162	18	0.2411	1.0163	0.1797	1.0712	0.1314	373.3521	−338.4100	0.0000	2294.6255	0.3973	0.8718

Table 14
The cluster numbers *c* found by nine indices and its corresponding index value for Pima data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7047	0.1316	0.8198	0.1319	0.6396	0.9526	−5 929 555.2572	47 003 538.5097	**0.0000**	0.6405	0.6519
8	2	0.6949	**0.1260**	**0.8289**	**0.1263**	**0.6578**	0.8684	−7 581 701.8924	47 586 534.6489	0.0000	0.6370	0.6558
11	3	**0.7170**	0.1954	0.7506	0.1962	0.6259	**0.8244**	−11 727 937.1964	38 756 613.2310	0.0000	0.5634	0.6545
28	4	0.7018	0.2731	0.6617	0.2745	0.5489	0.9432	**−12 633 369.2549**	30 047 313.0653	0.0000	0.5129	0.6610
220	24	0.3458	0.9073	0.2270	0.9366	0.1934	1.9927	−4 222 498.1937	**18 637 742.3099**	0.0000	0.2279	0.7458
256	27	0.2401	1.0264	0.1639	1.0638	0.1317	11.0516	−3 263 582.5446	74 222 179.0577	0.0011	0.2197	0.7458

Table 15
The cluster numbers *c* found by nine indices and its corresponding index value for WPBC data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.7313	**0.1271**	0.8230	**0.1284**	0.6460	0.7882	−49 320 314.3104	**0.0000**	**0.0000**	0.6465	0.7626
4	2	0.7275	0.1272	**0.8233**	0.1285	**0.6466**	0.7662	−53 374 615.9442	0.0000	0.0000	0.6586	0.7626
27	4	**0.7540**	0.2346	0.7193	0.2394	0.6257	0.7685	−69 153 204.6872	0.0000	0.0000	0.4867	0.7626
29	4	0.7507	0.2253	0.7292	0.2299	0.6389	**0.7238**	−90 296 432.6546	0.0000	0.0000	0.4966	0.7626
38	5	0.7467	0.2622	0.6963	0.2690	0.6204	0.7718	**−105 252 045.0466**	0.0000	0.0000	0.4786	0.7727
121	14	0.6072	0.5514	0.4540	0.5934	0.4120	1.3382	−47 165 127.5498	0.0000	0.0000	0.2271	0.7727

Table 16
The cluster numbers *c* found by nine indices and its corresponding index value for glass data set.

RT	*C*	*WGLI*	*PE*	*PC*	*MPE*	*MPC*	*DB*	*FS*	*FHV*	*XB*	*FM*	*PM*
1	2	0.6834	**0.1368**	**0.8091**	**0.1381**	0.6182	1.3397	−501.4581	**0.0000**	**0.0001**	0.5217	0.4486
9	2	**0.6853**	0.1503	0.7859	0.1517	0.5718	1.4818	−369.5523	0.0000	0.0001	0.5181	0.4486
11	3	0.6766	0.1970	0.7601	0.1998	**0.6402**	**1.2283**	**−961.5440**	0.0000	0.0001	0.5475	0.5000

Pima Indians Diabetes Database [33]: The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. The population lives near Phoenix, Arizona, USA. Several constraints were placed on the selection of these instances from a larger database. The data set consists of 768 samples in 8-dimensional space and has 2 clusters. In Table 14, there are five indices by which one can find the optimal cluster number $c = 2$. Our proposed index fails to detect the correct cluster number and the corresponding values of *FM* is only 0.5634. Obviously, this result is not good.

Wisconsin Prognostic Breast Cancer (WPBC) [34]: Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The first 30 features are computed from a digitized image of a fine needle aspirate of a breast mass. This data set consists of 198 samples in 34-dimensional space and has 2 clusters. From Table 15, it can be seen that the cluster number $c = 2$ is obtained by six indices. The proposed index *WGLI*, *DB* and *FS* fail to find the optimal cluster number.

The Glass data set with six clusters [33] has 214 points in 9-dimensional space. Since the clusters of this data set are heavily overlapped, it is difficult to find an optimal cluster number by the validity indices. The optimal cluster number is not detected by all nine indices shown in Table 16. This data set is only clustered two or three groups by these nine indices and the values of *FM* is not less than or equal to 0.5. This fact indicates that we can not successfully partition this data set into proper groups. Maybe, it is a good way to deal with the problem that one could simply choose the ideal cluster centers and compute the membership degrees w.r.t. these cluster centers and then check what the validity index says about the result.

It is a difficult problem to correctly cluster high-dimensional data in data mining. We can sometimes obtain the desired clustering results for some particular real datasets (in high-dimensional space) with well-separated clusters. However, as pointed in Ref. [36], FCM algorithm fails to group some artificial data sets in high-dimensional space. This case will lead to the failure of our index because our index depends on the cluster result obtained by FCM.

6. Conclusions

In this paper, we proposed a novel cluster validity index *WGLI* to detect the optimal cluster number. The proposed index utilizes global optimum membership as its global property and modularity of bipartite network as its local independent property. To prove the effectiveness of our proposal, we select six artificial data sets and nine real data sets widely used in data mining to evaluate the performance of the proposed index *WGLI*. As a result, we can successfully find most optimal cluster numbers on these 15 data sets (12 correct cluster numbers among 15). The comparisons of the introduced index *WGLI* with other eight popular indices on above mentioned data sets show that we can obtain the good fuzzy c-partition by the indices *WGLI* and *MPC*. The results of this study suggest that the new validation index can achieve the optimal result for most datasets. However, none of the above mentioned indices correctly recognizes optimal cluster numbers c for all the mentioned datasets. Furthermore, none of the nine indices detects the cluster number 6 for the glass data set. Therefore, it is worthwhile to further introduce novel indices to deal with this problem.

Acknowledgements

The author Y. Zhang's work is supported by NSFC under the grants No. 61373127. The authors would like to express gratitude to anonymous referee for providing several exceptionally helpful comments which led to a significant improvement of the article.

References

[1] J.C. Dunn, A fuzzy relative of the ISODATA process its use in detecting compact well-separated clusters, J. Cybern. 3 (1974) 32–57.
[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[3] D. Gustafson, W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: Proceedings of the IEEE CDC, San Diego, CA, USA, 1979, pp. 761–766.
[4] S.H. Kwon, Cluster validity index for fuzzy clustering, Electron. Lett. 34 (22) (1998) 2176–2177.
[5] D.W. Kim, K.H. Lee, D. Lee, On cluster validity index for estimation of the optimal number of fuzzy clusters, Pattern Recognit. 37 (2004) 2009–2025.
[6] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets Syst. 158 (2007) 2095–2117.
[7] J.C. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man Cybern. 28 (1998) 301–315.
[8] K.Y. Huang, Applications of an enhanced cluster validity index method based on the Fuzzy C-means and rough set theories to partition and classification, Expert Syst. Appl. 37 (2010) 8757–8769.
[9] K.R. Zalik, Cluster validity index for estimation of fuzzy clusters of different sizes and densities, Pattern Recognit. 43 (2010) 3374–3390.
[10] J. Liang, X. Zhao, D. Li, et al., Determining the number of clusters using information entropy for mixed data, Pattern Recognit. 45 (2012) 2251–2265.
[11] J.C. Bezdek, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 57–71.
[12] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybern. 3 (1974) 58–72.
[13] R.N. Dave, Validating fuzzy partition obtained through c-shells clustering, Pattern Recognit. Lett. 17 (1996) 613–623.
[14] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224–227.
[15] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, in: Proceedings of Fifth Fuzzy Systems Symposium, 1989, pp. 247–250.
[16] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 773–781.
[17] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 841–847.
[18] K.L. Wu, M.S. Yang, A cluster validity index for fuzzy clustering, Pattern Recognit. Lett. 26 (2005) 1275–1291.
[19] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognit. 37 (2004) 487–501.
[20] B. Rezaee, A cluster validity index for fuzzy clustering, Fuzzy Sets Syst. 161 (2010) 3014–3025.
[21] D. Watts, S. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[22] M.E.J. Newman, Scientific collaboration networks. I. Network construction and fundamental results, Phys. Rev. E 64 (2001) 016131/1–016131/8.

[23] R. Guimera, M. Sales-Pardo, A. Lan, Module identification in bipartite and directed networks, Phys. Rev. E 76 (2007).

[24] M.J. Barber, Modularity and community detection in bipartite network, Phys. Rev. E 76 (2007).

[25] T. Murata, Modularity for bipartite networks, in: N. Memon, et al. (Eds.), Annals of Information Systems, in: Data Mining for Social Network Data, vol. 12, Springer Science+Business Media, LLC, 2010.

[26] S. Bandyopadhyay, U. Maulik, Non-parametric genetic clustering: comparison of validity indices, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 31 (1) (2001) 120–125.

[27] S. Bandyopadhyay, U. Maulik, Genetic clustering for automatic evolution of clusters and application to image classification, Pattern Recognit. 35 (2002) 1197–1208.

[28] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annu. Eugen. 7 (1936) 179–188.

[29] R. Kothari, D. Pitts, On finding the number of clusters, Pattern Recognit. Lett. 20 (1999) 405–416.

[30] S. Aeberhard, D. Coomans, O. de Vel, Comparison of Classifiers in High Dimensional Settings, Tech. Rep. No. 92-02, James Cook University of North Queensland, 1992.

[31] O.L. Mangasarian, W.H. Wolberg, Cancer diagnosis via linear programming, SIAM Soc. Newsl. 23 (1990) 1–18.

[32] S.J. Haberman, Generalized residuals for log-linear models, in: Proceedings of the 9th International Biometrics Conference, Boston, 1976, pp. 104–122.

[33] UCI data sets, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[34] W.H. Wolberg, W.N. Street, D.M. Heisey, O.L. Mangasarian, Computerized breast cancer diagnosis and prognosis from fine needle aspirates, Arch. Surg. 130 (1995) 511–516.

[35] C. Luo, Y. Li, S.M. Chung, Text document clustering based on neighbors, Data Knowl. Eng. 68 (2009) 1271–1288.

[36] R. Winkler, F. Klawonn, R. Kruse, Fuzzy C-means in high dimensional spaces, Fuzzy Syst. Appl. 1 (2011) 1–17.