An Adaptive Cluster Validity Index for the Fuzzy C-means

CHEN Duo[1),2)], LI Xue[1),3)] and CUI Du-Wu[1)],

[1)] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China
[2)] Computer centre of TangShan College, TangShan, 063000, China
[3)] International Business School of Shaanxi Normal University, Xi'an,710062, China

Summary
Based on the basic theory of fuzzy set, this paper suggests the notion of FCM fuzzy set, which is subject to the constraint condition of fuzzy c-means clustering algorithm. The cluster fuzzy degree and the lattice degree of approaching for the FCM fuzzy set are presented, and their functions in the validation process of fuzzy clustering are deeply analyzed. A new cluster validity index is proposed, in which the two factors such as the cluster fuzzy degree and the lattice degree of approaching are taken into comprehensive account. The notable advantage of the index is that it can adaptively adjust the relative significance levels of the two factors. Also, this paper gives the algorithm to apply the cluster validity index to the cluster validation for the fuzzy c-means algorithm. The experimental results indicate the effectiveness and adaptability of the proposed cluster validity index.

Key words:
Cluster Analysis, Cluster Validity Index, Fuzzy C-means Clustering, Fuzzy Set

1. Introduction

Fuzzy clustering analysis is an important research project in knowledge discovery and data mining. In practical applications, we have the scientific data survey, database analysis, customer relations management, medicine diagnosis, weather forecast, water analysis, etc. How to determine the optimal partition and optimal number of clusters for fuzzy partitions belongs to clustering validity problems, being one of the most important issues related to fuzzy clustering analysis. The construction of cluster validity indexes is a common method for solving the problem. In the fields of fuzzy clustering analysis, the fuzzy c-means (FCM) algorithm [1] is one of the most widely used methods, and many cluster validity indexes suitable for the algorithm have been proposed. Bezdek's partition coefficient (PC) [2] and partition entropy (PE) [3] are defined based on the membership values of a fuzzy partition. With the evident mathematical significance and good mathematical character, and the advantages of simplicity and high-effect, both of them have been frequently used. However, the two indexes use only the membership values of a fuzzy partition of data and may be lack for the connection to the geometrical structure of the data [4]. To solve this problem, investigators have advanced a number of fuzzy cluster validity indexes that include both the membership values of a fuzzy partition and the information of data structure of the clusters, for example the traditional XB, Vk, FS [5,6,7] , and the PACES, FSα [4,8] appeared recently in literature, etc. Cluster properties such as compactness and separation are often considered as major characteristics by which to validate clusters [9]. Compactness is used as a measure of the closeness or scattering within clusters, and separation as a measure of the isolation of clusters from each other. A good clustering result should have the properties of being both small intra-cluster compactness and large inter-cluster separation at the same time.

The main objective of our paper is to design a cluster validity index that is suitable for FCM. Considering the constraint condition, the fuzzy partition obtained from FCM clustering algorithm has been defined as the FCM fuzzy set in this paper. Based on the definition, we further advance the cluster fuzzy degree and the lattice degree of approaching of the FCM fuzzy set and they are used as the measure of compactness and separation of fuzzy partition of data respectively. A new adaptive cluster validity index is designed, in which the fuzzy degree and the lattice degree of approaching are taken into comprehensive account. In the index, the two measures are in the symmetric position such that their important levels depend on the values themselves. Thus, the proposed cluster validity index can adaptively adjust the impact degree of the two factors. Also, the FCM validation algorithm is suggested in this paper. Experimental results on artificial and real-life data sets indicate that the index is stable and adaptive.

The remainder of this paper is organized as follows: Section 2 gives a brief introduction of FCM clustering algorithm and some cluster validity indexes suitable for the algorithm. Section 3 presents the notion of FCM fuzzy set and the definitions of cluster fuzzy degree and the lattice degree of approaching. Experimental results on both artificial and real data sets are given in Section 4. We conclude the paper and have an outlook for further research in the last section.

Manuscript received February 5, 2007
Manuscript revised February 25, 2007

2. Brief introduction of FCM algorithm and fuzzy cluster validity indexes

2.1 Fuzzy c-mean algorithm

FCM belongs to partition algorithm in the fields of fuzzy cluster analysis. The algorithm classifies a set of objects $X=\{x_1,x_2,...,x_n\}$ into c homogeneous groups represented as fuzzy sets $\widetilde{F}=\{\widetilde{F}_1,\widetilde{F}_2,...,\widetilde{F}_c\}$ [9]. The objective of FCM is to obtain a fuzzy c-partition. It can be stated as a constrained nonlinear optimization problem, which minimizes Equation (1), under the constraint of Equation (2), as follows:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^{m}\parallel x_j-v_i\parallel^2 \tag{1}$$

subject to

$$\begin{cases}\sum_{i=1}^{c}u_{i,j}=1 & 1\le j\le n\\ 0\le u_{i,j}\le 1 & 1\le j\le n,\quad 1\le i\le c\\ 0<\sum_{j=1}^{n}u_{i,j}<n & 1\le i\le c\end{cases} \tag{2}$$

where, $U=[u_{i,j}]$ is a $c\times n$ fuzzy partition matrix, n the number of the objects and c the number of clusters. The weighting exponent $m>1$, controls the fuzziness of membership. $V=\{v_1,v_2,...v_c\}$ is a set of c vectors and each v_i, $i=1,2,...c$, expresses the center of the *ith* cluster. $||x_j-v_i||$ is the Euclidean norm between x_j and v_i. FCM algorithm adopts the alternating optimization strategy, in which, $U^{(t)}$, $V^{(t)}$ are improved by turns in each iteration (where t is the iteration step). The iteration is terminated when it reaches a stable condition. The outcome of FCM can be denoted by the pair (U,V) [9].

2.2 Some cluster validity indexes suitable for FCM

We can obtain the fuzzy partition of the data set using FCM algorithm. However FCM algorithm requires the user to pre-define the number of clusters (c), and different values of c corresponds to different fuzzy partitions, so the validation of clustering results is needed. In practical application, we can define c_{min} and c_{max} (the minimal and the maximal number of clusters) in advance, and then run FCM algorithm on each value of c over the range [c_{min}, c_{max}] to get different fuzzy partitions. Thus, the optimal result can be obtained by validating each of the fuzzy partitions according to the appropriate cluster validity indexes. In recent years, investigators have put forward a number of cluster validity indexes that are suitable for FCM, some typical ones are as follows.

Bezdek advanced two cluster validity indexes [2,3], the partition coefficient (*PC*) and partition entropy (*PE*), defined as

$$PC=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{i,j}^2 \tag{3}$$

$$PE=-\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{i,j}\log_a(u_{i,j}) \tag{4}$$

PC and *PE* is used to measure the fuzziness of the fuzzy partition matrix, the lower the fuzziness of a partition is, the larger the *PC* value (or the smaller the *PE* value). The two indexes use only the membership values of the fuzzy partition; therefore, they are devoid of connection to the structure of the data set.

Xie and Beni proposed a validity index (*XB*) [5] as follows:

$$XB=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^2\parallel x_j-v_i\parallel^2}{n(\min_{h\ne k}\parallel v_h-v_k\parallel^2)} \tag{5}$$

In Eq. (5), the numerator denotes the compactness by the sum of square distances within clusters, while the denominator denotes separation by the minimal distance between clusters. The index is of high reliability and accuracy and has been widely uses for fuzzy clustering validation.

Kwon extended the index XB and proposed a cluster validity index V_k [6]:

$$V_k=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^2\parallel x_j-v_i\parallel^2+\frac{1}{c}\sum_{i=1}^{c}\parallel v_i-\bar{v}\parallel^2}{(\min_{i\ne k}\parallel v_i-v_k\parallel^2)} \tag{6}$$

The validity index, *FS*, proposed by Fukuyama and Sugeno [7] was defined as

$$FS=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^m\parallel x_j-v_i\parallel^2-\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^m\parallel v_i-\bar{v}\parallel^2 \tag{7}$$

The first item is the FCM objective function which measures the compactness and the second measure the separation in Eq. (7).

In the Eq. (6) and Eq. (7), $\bar{v}=\sum_{i=1}^{c}v_i/c$.

Wu and Yang [4] proposed $PCAES(c)$ index as follows:

$$PCAES(c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^2/u_M-\sum_{i=1}^{c}\exp(-\min_{k\ne i}\{\parallel v_i-v_k\parallel^2\}/\beta_T) \tag{8}$$

where, $u_M=\max_{1\le i\le c}\sum_{j=1}^{n}u_{i,j}^2$, and $\beta_T=\frac{1}{c}\sum_{i=1}^{c}\parallel v_i-\bar{v}\parallel^2$.

PACES index consists of two terms. The first term is the normalized partition coefficient to measure the compactness. The second term is an exponential-type separation measure, which takes advantage of exponential function that measures the sum distances between the closest pairs of cluster centers [4].

Campello and Hruschka [8] proposed a cluster validity index, named Fuzzy silhouette (*FSα,* called in this paper), is defined as follows:

$$FS_\alpha = \frac{\sum_{j=1}^{n}(u_{p,j}-u_{q,j})^\alpha s_j}{\sum_{j=1}^{n}(u_{p,j}-u_{q,j})^\alpha} \quad (9)$$

where $u_{p,j}$ and $u_{q,j}$ are the first and second largest elements of the *j*th column of the fuzzy partition matrix respectively, $\alpha \geq 0$ is a weighting coefficient, and s_j is the silhouette of *j*th object defined as:

$$s_j = \frac{b_{pj}-a_{pj}}{\max\{a_{pj}, b_{pj}\}} \quad (10)$$

where, a_{pj} is the average distance of *j*th object to all other objects belonging to *p*th cluster, while d_{pj} the minimum d_{qj}, which is the average distance of *j*th object to all object belonging to another cluster *q*, $q \neq p$. Exponent α is an optional parameter (unit by default) [8]. It can be known that the a_{pj} and b_{pj} represent the compactness and the separation respectively. Campello and Hruschka pointed out that the time complexity of computing s_j for all objects is $O(n^2)$. In order to reduce the computational burden, the simplified method based on the distances among the objects and the cluster centers of the corresponding clusters[10] are used in the literature [8]. This modification has shown not to degrade accuracy while being able to significantly reduce the time complexity to $O(n)$ [8].

To sum up the above descriptions, the major cluster properties such as compactness and separation are used in most of the cluster validity indexes proposed in recent decades, and obtaining the minimum compactness of intra-cluster under the premise of the as large as possible separation of inter-clusters in the fuzzy partition is a fundamental characteristic of fuzzy cluster validity indexes.

3. The proposed cluster validity index

3.1 The measure of compactness

In the fuzzy cluster validity indexes, $\sum_{j=1}^{n} u_{i,j}^m \| x_j - v_i \|^2 \ (m>1)$ is a usual type of the compactness measure, in which both square distances and membership values of fuzzy partition are considered at the same time. However, the measure suffers from a tendency to monotonically decrease when the number of clusters to the number of objects, because $\lim_{c \to n} \| x_j - v_i \| = 0$ [6]. Also, it is difficult to distinguish between the two clusters by the measure in some special cases [9]. Thus, there are some limitations in using of the measure of compactness. If using only the distances within clusters to measure the compactness of fuzzy partition, the process of defuzzification is required in advance. This method will lead to an extra computational error, obviously. In literature [13] and [9], the inclusion degree and overlap degree between clusters are respectively used as the measures of compactness. However, these measures are indirect, and some relevant parameters are needed to provide in advance for the overlap degree measure. To tackle these problems, we define the cluster fuzzy degree to measure the compactness of a fuzzy partition obtained from FCM algorithm.

Definition 1. Let $X = \{X_1, X_2, ..., X_n\}$ be a *n*-data set and $c \times n$ fuzzy partition matrix $U = [u_{i,j}]$ satisfy the constraint conditions (2), and then FCM fuzzy set $\widetilde{F}_i \in U, i \in \{1,...,c\}$ is defined as follows:

$$\widetilde{F}_i = \sum_{j=1}^{n} \mu_{\widetilde{F}_i}(x_j)/x_j = \sum_{j=1}^{n} u_{i,j}/x_j \qquad i = 1,...,c \quad (11)$$

We adopt L. Zadeh's [11] convenient expressions in Eq. (11). It can be known from Definition 1 that the FCM fuzzy set is derived from *c*-fuzzy sets that satisfy the constraint condition (2) of FCM.

Definition 2. The cluster fuzzy degree of FCM fuzzy set is defined as follows:

$$D(U;c) = \sum_{i=1}^{c} D_i(U;c) \quad (12)$$

where

$$D_i(U;c) = \frac{1}{n}\sum_{j=1}^{n} \left| u_{ij} - (u_{i,j})_T \right| \quad (13)$$

$$(u_{ij})_T = \begin{cases} 1 & u_{ij} \geq \dfrac{1}{c} \\ 0 & \text{其它} \end{cases} \quad (14)$$

The cluster fuzzy degree is defined on the FCM cluster fuzzy set, so it is suitable for the FCM clustering. Following from Definition 2 immediately, we have

Property 1. $0 \leq D_i(U;c) \leq (c-1)/c$, $0 \leq D(U;c) \leq c-1$

Property 2. $D(U;c) = 0$ iff U is crisp.

Property 3. $D(U;c) = c-1$ iff $U = [1/c]$

The more distinct the fuzzy partition is, the smaller the value of cluster fuzzy degree is, and reaches its minimum (zero) under the condition of crisp partition (the most distinct). On the contrary, the fuzzier the partition is, the larger the value of cluster fuzzy degree is. If every object belongs to all clusters uniformly, i.e. $u_{i,j} = 1/c \ (1 \leq j \leq n, \ 1 \leq i \leq c)$, the cluster fuzzy degree takes its maximum, *c*-1, which is the fuzziest state of a fuzzy partition. In the FCM cluster algorithm, $u_{i,j}$, the value of membership of an object x_j to the *j*th cluster, is dependent on $\|x_j - v_i\|$, the distance between object x_j and cluster center v_i. The larger the different of distances of an object to all cluster centers, the more clearly the object

belongs to the cluster being the closest to the object, and the further the distribution of fuzzy partition deviate from the fuzziest state. As a result the cluster fuzzy degree will be smaller, which corresponds to a better property of the fuzzy partition. On the basis of the above analysis, the cluster fuzzy degree can be used as the measure of compactness of fuzzy partitions.

3.2 The separation measure

In most of fuzzy cluster validity indexes, such as the above-mentioned indexes, *XB*, *Vk*, *FS*, and *PACES*, the separation measures are calculated based on the distances among cluster centers, i.e. $\| v_i - v_j \|$, $i \neq j$. However, the measures have a limited capacity to differentiate the geometric structures of clusters because the calculation is based only on centroids information and does not consider the overall cluster shape [9]. To overcome the shortcomings, the lattice degree of approaching is used as a measure of separation in this paper. The lattice degree of approaching is calculated from the fuzzy partition matrix, in which the membership values of all objects belonging to each cluster centers are included. We introduce the definition of lattice degree of approaching usual used in fuzzy set to the FCM fuzzy set.

Definition 3. Let the FCM fuzzy set $\widetilde{F}_i, \widetilde{F}_k \in U$ $i,k \in \{1,2,...c\}$, the lattice degree of approaching is defined as follows:

$$N(\widetilde{F}_i, \widetilde{F}_k) = \min[(\widetilde{F}_i \circ \widetilde{F}_k), (\widetilde{F}_i \hat{\circ} \widetilde{F}_k)^c] \quad (15)$$

where, $(\widetilde{F}_i \circ \widetilde{F}_k) = \bigvee_{j=1}^{n} (u_{i,j} \wedge u_{k,j})$ is a inner product, and $(\widetilde{F}_i \hat{\circ} \widetilde{F}_k) = \bigwedge_{j=1}^{n} (u_{i,j} \vee u_{k,j})$ a outer product. $\wedge, \vee$ and $(\cdot)^c$ denote operators of maximization, minimization and complementation respectively. Let a FCM fuzzy set $\widetilde{F}_i$ be fixed, if FCM fuzzy set $\widetilde{F}_k$ $(k \neq i)$ approaches $\widetilde{F}_i$ gradually, the inner product will increase while the out product decrease. As a result, the lattice degree of approaching will get larger. Thus, the lattice degree of approaching can measure the degree of similarity between FCM fuzzy sets.

According to the basic fuzzy set theory, the lattice degree of approaching satisfies the following properties.

Property 1. $0 \leq N(\widetilde{F}_i, \widetilde{F}_k) \leq 1$

Property 2. $N(\widetilde{F}_i, \widetilde{F}_k) = N(\widetilde{F}_k, \widetilde{F}_i)$

Property 3 $\widetilde{F}_i \subseteq \widetilde{F}_j \subseteq \widetilde{F}_k \Rightarrow$

$$N(\widetilde{F}_i, \widetilde{F}_k) \leq N(\widetilde{F}_i, \widetilde{F}_j) \wedge N(\widetilde{F}_j, \widetilde{F}_k)$$

In this paper, the lattice degree of approaching is used as a measure of the separation of a fuzzy partition. The smaller the measure is, the larger the degree of separation, and the better the result of clustering.

3.3 The proposed cluster validity index

Based on above-mentioned measures of compactness and separation, an adaptive cluster validity index is advanced in this paragraph.

Definition 4. Let *U* be a $c \times n$ fuzzy partition matrix, the cluster validity index *DN(U;c)* is defined as follows:

$$DN(U;c) = \begin{cases} 0 & DC = 0 \ \ and \ NM = 0 \\ \dfrac{2 * DC * NM}{DC + NM} & otherwise \end{cases} \quad (16)$$

where,

$$DC = \frac{1}{c-1} D(U;c) \quad (17)$$

and

$$NM = \max_{k \neq h} [N(\widetilde{F}_h, \widetilde{F}_k)] \quad (18)$$

According to the Property 1 of lattice degree of approaching the value of *NM* in (18) will belong to interval [0,1]. We exploit the maximum value of lattice degree of approaching to measure the separation of a fuzzy partition in the cluster validity index. It is worth noticing that the maximum value is corresponding to the closest pair of cluster centers, which is the most unfavorable case for the fuzzy partition. From this viewpoint, the measure *NM* in (18) is similar to the separation measure $\min_{i \neq k} \| v_i - v_k \|$ in (5) defined by the *XB* index. The main purpose of introducing the term $1/(c-1)$ into *DC* in (17) lies in counteracting the interference of the changing in number of clusters (*c*), and having the value of *DC* limited to the interval [0,1] so as to match with *NM* in value.

The cluster validity index *DN* can comprehensively reflect the two factors of *DC* ad *NM*. It is easy to proof that *DN* is the monotonic increasing function with respect to *DC* and *NM*. Because the values of both *DC* and *NM* belong to the interval [0,1], we have $DN \in [0,1]$ and, *DN*=0 iff *DC*=0 or *NM*=0 or *DC*=*NM*=0; *DN*=1,iff $DC=N_M=1$. It can be known from Definition 4 that *DC* and *NM* in *DN* are two symmetrical terms. When *DC* and *NM* values are closer, they will have the similar effect upon *DN*; when the value of one of the terms is much smaller than that of the other, the term having the small value will have the greater effect upon *DN*. Hence, *DC* and *NM* can act on the cluster validity index *DN* adaptively. In some practical application, as the cluster shapes in data sets are possibly quite different, the sensitivity of *DC* and *NM* in the cluster validation is also quite different under such situation. It just because *DN* can automatically adjust the

relative significance levels of the two factors, *DC* and *NM*, the cluster validity index has a better discriminating ability in various kinds of cluster shapes. In Eq. (16), the term *DC* indicates the compactness of a fuzzy partition, while the term *NM* indicates the degree of the separation between clusters. A good fuzzy partition produces a small value of the *DC*, and that well-separated cluster centers will bring a small value of *NM*. Thus, the smaller the *DN* is, the better the performance of fuzzy clustering, and the most desirable fuzzy clustering result is obtained by minimizing *DN* for c from c_{min} to c_{max}.

3.4 The FCM validation algorithm

The FCM validation algorithm is described as:

1. Initialization: Chose a cluster validity that is suitable for FCM, set the initial values : c_{min}, c_{max}, *TIME,* and m (weighting exponent used in FCM), and let $c=c_{min}$
2. Run FCM algorithm *TIME* times and select the optimal (U,V) (corresponding to the minimum of $J_m(U,V)$)
3. Compute and store the value of the cluster validity index
4. if $c<c_{max}$, then $c \leftarrow c+1$ and go to 2, else go to 5
5. Output the optimal (U^*, V^*) and the optimal c^* according to the cluster validity index.

The performance of a fuzzy cluster validity index depends on the outcome of a fuzzy clustering algorithm, and a validity index is not able to provide desirable evaluation when the used clustering algorithm is not appropriate to the partitioning of a given data set[9]. Considering that FCM algorithm is sensitive to the initial-choices (cluster centers or partition matrix), we run the algorithm *TIME*=100 times starting form random initialization of cluster centers for each validating, and the optimal (U,V) that is correspondent to the minimum of $J_m(U,V)$ is selected as the outcome of FCM in each iteration of the FCM validation algorithm.

4. Experimental results

Five artificial and two real-life data sets are considered for testing the performance of cluster validity indexes. In our experiments, we execute The FCM validation algorithm defined in 3.4, and the proposed index $DN(U;c)$ is compared with seven fuzzy cluster validity indexes mentioned in Section 2.2: *PC* [2], *PE* [3], *XB* [5], V_k [6], *FS* [7], *PACES* [4] and *FSα* [8].

4.1 Data sets

The artificial data sets, Data_A, Data_B, Data_C, Data_D and Data_E, are all uniformly distributed inside the bi-dimensional region appointed, being illustrated in Figs 1-5, respectively. Data_A is comprised of 60 objects that are classified into 3 clusters with some "bridge points" in it as shown in Fig. 1. The distribution of Data_B are demonstrated in Fig. 2, we can see that there are 4 clusters in the data set and two of them are adjacent, while the others are well separated from each other. As shown in Fig. 3 Data_C has 300 objects and 5 clusters with 60 objects per cluster, which are well separable from each other. Data_D has 490 objects, in which 480 objects are classified into 6 clusters with 80 objects per cluster, and the other 10 objects are the points of noise as shown in Fig. 4. Data_E is comprised of 400 objects. As displayed in Fig. 5, there is some overlapping between some clusters. Intuitively, the data set should be classified into 8 clusters.

The real-life data sets, Iris data set and Cancer data set, are both obtained form the UCI Machine Learning Repository [14]. Iris data set expresses different categories of iris flowers, having 150 objects with 4 numeric attributes, namely sepal length, sepal width, petal length, and petal width. It has three classes, i.e. Setosa, Versicolor and Virginica, each containing 50 objects. It is known that two classes Versicolor and Virginica have some overlap while the class Setosa is well separated from the other two Thus, we can accept that there are 2 or 3 clusters in the Iris data set. The Iris data is wildly used for examining the performance of clustering algorithms and the cluster validity indexes. Cancer data set is the Wisconsin Breast Cancer data set, in which contains 699 objects with 9 numeric attributes, they are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. The data set has two categories: malignant and benign. In addition, there are 16 objects that contain a single missing attribute value, which are not considered in our experiments.

In our experiments, the algorithm proposed in Section 3.4 is executed for the 5 artificial data sets and 2 real-life data sets respectively. We appoint c_{min}=2. As far as the c_{max} is concerned, its value can often be obtained from the domain knowledge; however, as this is not always possible, a rule of thumb that many investigators use is $c_{\max} \le \sqrt{n}$ [5,9,12]. We adopt the upper limit value of the range. The parameter *TIME*=100 is given. Pal and Bezdek showed that the FCM algorithm provided the best results for $m \in [1.5,2.5]$ [12], then we take the medium value m=2 in the experiments.

4.2 Results

The variation of the *DN* index and its two terms, *DC* and *NM*, with the number of clusters for the above-mentioned experimental data sets are shown in Figs. 6-10.

Fig.6 shows that the minimal values of *NM* at c=3 and *DC* at c=8. Although *DC* does not detect the correct cluster number, *NM* has a steeper valley at c=3 such that

the *DN* index prefers the optimal cluster number. In Data_B, there are 4 clusters and two of them are adjacent, while the others are well separated from each other, thus the *NM* is acutely changed at about the optimal cluster number *c*=4 as shown in Fig. 7. The term *DC* reaches the valley point at cluster numbers 4, as it play a great role in *NM* for this data set, the index *DN* can successfully finish the validation and the optimal cluster number *c*=4 is obtained. As far as Data_C is concerned, the clusters are well separable from each other, as expected, Fig. 8 shows that the *DN* index and its two terms *DC* and *NM* arrive at the valley points at *c*=5 simultaneously, and we can see that *NM* has a better resolution performance than *DC*. Also, the clusters in Data_D are well separable form each other, but it is different form Data_C that there are some points of noise in Data_D data set. In spite of this, the *NM* still has a steeper valley as shown in Fig. 9. In contrast, curve of *DC* is smoother. Under the action of *NM* and *DC* together the index *DN* reaches the optimal point at *c*=6. In Data_E there is some overlapping. For this data set, the term *DC* plays a major role than *NM* in *DN* index as shown in Fig. 10. Although the term *NM* has a larger value than *DC* such that *NM* has less effect to the index *DN*, the term *NM* has a steeper valley. So it enhances the examination performance of the index *DN* to some degree. In this Data set, the index *DN* can also work well.

The values of each cluster validity index for the experimental data sets are shown in Tables 1-7 respectively, where the optimum values of indexes are presented in boldface.

Data_A is comprised of 60 objects and classified into 3 clusters clearly, although there are some "bridge points" in it, every index can correctly find the optimal cluster number as indicated in Table 1. In Data_B data set, the 4 clusters are separated clearly, in which two clusters are closer while the others are more distant. From Table 2 it is known that the index *FS* and *DN* can correctly recognize the optimal cluster number, and *FSα* only obtain the approximate result of *c*=3, while the other indexes do not correctly work. The distribution of Data_C is symmetric, and the 5 clusters in it are correctly discerned by the indexes: *PC*, *XB*, *Vk*, *FS*, *PACES* and *DN* as shown in Table 3. There are some points of noise in Data_D, the index *XB*, *Vk*, *FS*, *PACES*, *FSα* and *DN* can overcome the influence of noise such that the optimal result *c*=6 can be examined. In contrast, *PC* and *PE* provide cluster numbers *c*=2 as shown in Table 4. There are 8 clusters in Data_E that is a few more cluster number, the index *XB*, *Vk*, *FS, FSα* and *DN* can correctly validate while the others fail to do so.

As for the real-life Iris data set, it is acceptable that both *c*=2 and *c*=3 can be used as the optimal cluster number. The Table 6 lists the validation results of each index for the data set. The optimal *c*=2 is identified by *PC*, *PE*, *XB*, *Vk*, *PACES*, *FSα* and *DN*, while only the *FS* index yield cluster number *c*=5. In the other real-life data set, the cancer data set, we have the optimal cluster number c=2. From Table 7, we can know that the index *PC*, *PE*, *XB*, *Vk*, *PACES* and *DN* provide the optimal value at *c*=2, but the index *FS* and *FSα* are wrong for the data set.

Table 8 summarizes the results obtained when the above validity indexes are applied to the artificial and real-life data sets. In the Table, the column $c_{optimal}$ expresses the optimal number of clusters for each data set and the other columns show the optimal cluster numbers obtained from each validity index. The last row indicates the accuracy, i.e. the ratio of the number of data sets that are correctly validated by the cluster validity index in corresponding column and the total of data sets. We can observe that the proposed cluster validity index *DN* has the highest accuracy of the indexes considered.

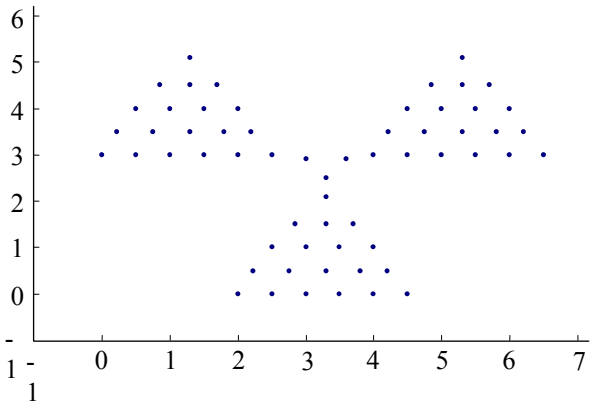

Fig. 1 Data_A data set

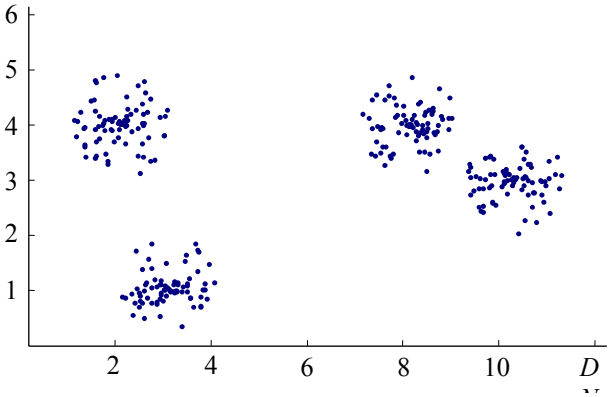

Fig. 2 Data_B data set

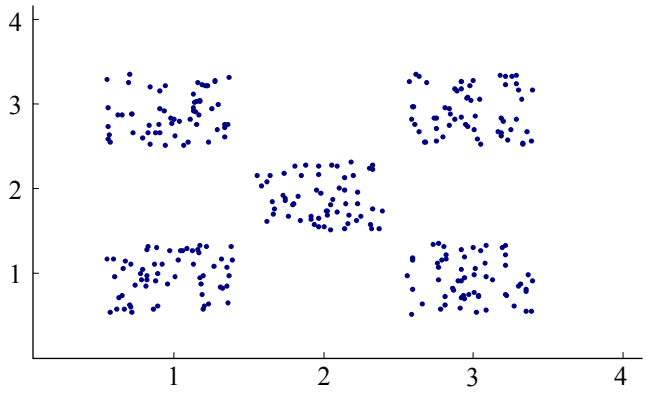

Fig. 3 Data_C data Set

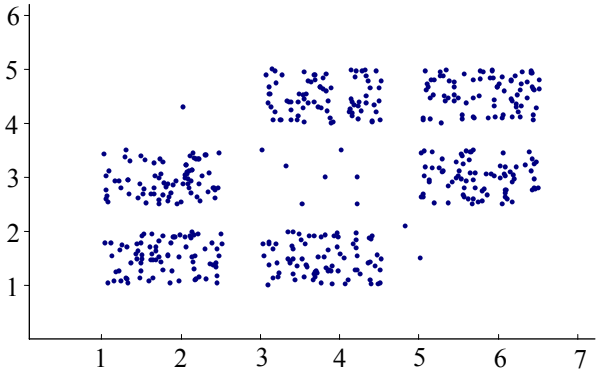

Fig. 4 Data_D data set.

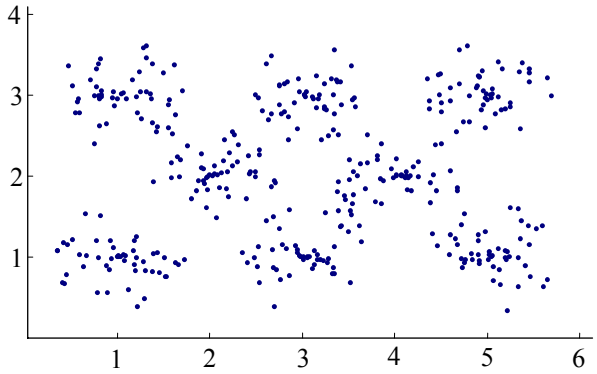

Fig. 5 Data_E data set.

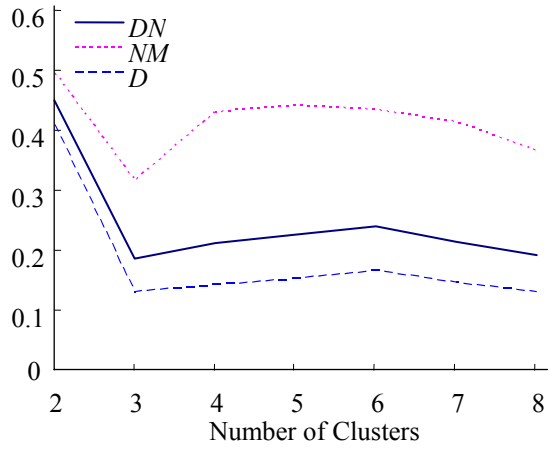

Fig. 6 Variation of the DN, DC and NM with the number of clusters for Data_A.

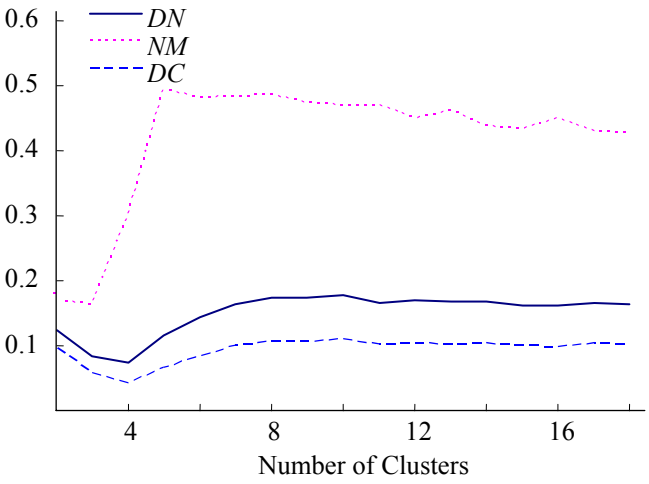

Fig. 7 Variation of the DN, DC and NM with the number of clusters for Data_B.

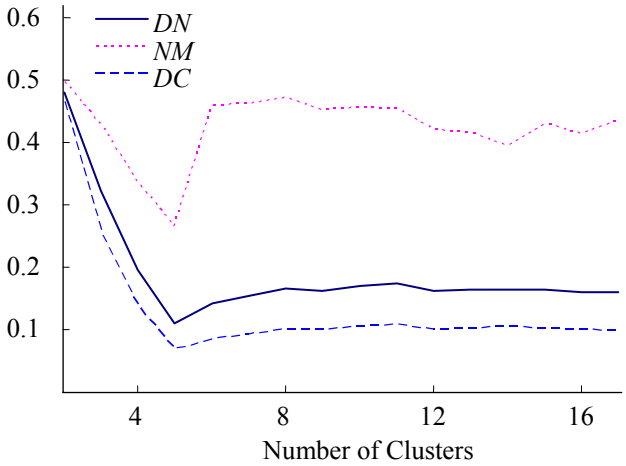

Fig. 8 Variation of the DN, DC and NM with the number of clusters for Data_C.

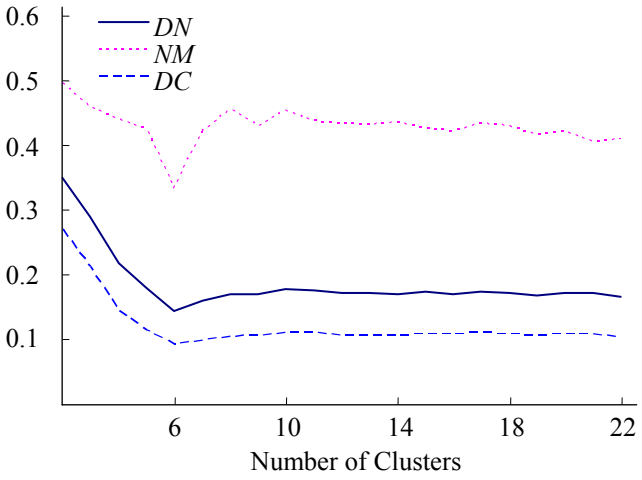

Fig. 9 Variation of the DN, DC and NM with the number of clusters for Data_D.

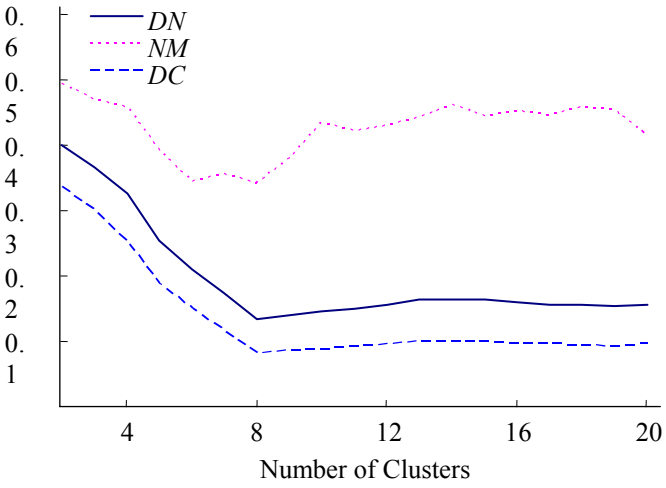

Fig. 10 Variation of the DN, DC and NM with the number of clusters for Data_E.

Table 1. Cluster validity value for Data_A data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	0.717907	0.43657	0.253279	15.4467	38.0565	1.89149	0.632466	0.449523
3	**0.78833**	**0.41852**	**0.05924**	**3.91529**	**−169.97**	**2.80754**	**0.76035**	**0.18523**
4	0.694391	0.609075	0.114483	7.7477	−162.943	2.26164	0.697261	0.212696
5	0.633037	0.745828	0.334781	23.5366	−163.689	0.957353	0.669325	0.225353
6	0.581302	0.855665	0.266549	19.9055	−153.968	0.142718	0.619255	0.239482
7	0.578629	0.894818	0.193881	15.5662	−163.59	1.52618	0.644751	0.214932
8	0.570312	0.943899	0.154688	13.0464	−166.57	1.1012	0.655424	0.191606

Table 2. Cluster validity value for Data_B data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	**0.90999**	**0.18391**	**0.04754**	**15.4628**	−2575.05	**1.92965**	0.78948	0.123082
3	0.895088	0.229788	0.096144	31.965	−3583.82	1.18534	**0.82993**	0.084807
4	0.891205	0.238918	0.054949	19.9247	**−3653.8**	1.62214	0.821602	**0.07318**
5	0.821232	0.362328	0.41685	154.824	−3595.36	0.069141	0.780002	0.116415
6	0.758776	0.476893	0.331143	129.545	−3094.67	−1.58631	0.737225	0.144094
7	0.710105	0.564557	0.274648	111.599	−2964.06	−2.69017	0.673697	0.164517
8	0.6743	0.633289	0.212948	90.7198	−2836.52	−1.22183	0.63449	0.173598
9	0.652746	0.688439	0.360252	158.595	−2796.37	−2.40878	0.63764	0.173658
10	0.627307	0.751592	0.390907	176.072	−2674.23	−2.7965	0.632401	0.177553
11	0.620138	0.778612	0.358626	168.148	−2679.66	−3.04609	0.640839	0.166914
12	0.599727	0.834893	0.404284	202.879	−2506.11	−5.64023	0.642695	0.17024
13	0.593443	0.865882	0.329548	160.331	−2577.81	−4.89203	0.653765	0.167156
14	0.580713	0.909997	0.311471	154.04	−2496.52	−5.55757	0.660653	0.167326
15	0.584568	0.907383	0.273262	144.635	−2497.63	−5.4117	0.668671	0.161349

Table 3. Cluster validity value for Data_C data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	0.662822	**0.51401**	0.347776	104.583	118.309	1.95388	0.511135	0.480961
3	0.653924	0.618972	0.125014	37.8528	−100.064	2.48942	0.680572	0.321761
4	0.743706	0.52889	0.054877	16.994	−318.574	3.33579	**0.80026**	0.196307
5	**0.7688**	0.521067	**0.05317**	**16.8096**	**−342.35**	**3.33582**	0.77377	**0.10933**
6	0.715155	0.633081	0.409304	130.74	−337.77	1.62945	0.747637	0.142393
7	0.670341	0.728931	0.346188	112.263	−307.15	−0.06343	0.721368	0.153258
8	0.638778	0.796174	0.295515	97.3355	−306.864	−1.66013	0.691204	0.165522
9	0.617947	0.863065	0.285866	94.9036	−300.704	−2.51452	0.698925	0.162777
10	0.595571	0.918822	0.280587	93.6427	−296.571	−3.24794	0.677133	0.169631
11	0.569326	0.970828	0.248969	85.5558	−276.462	−1.12418	0.628183	0.174458
12	0.573157	0.986365	0.189081	65.4294	−275.868	−1.27442	0.654324	0.161508
13	0.566736	1.01427	0.16902	59.6568	−271.899	−2.12502	0.66248	0.164237
14	0.555051	1.04932	0.187455	67.6596	−271.052	−2.56168	0.655751	0.164727
15	0.554185	1.06479	0.255923	94.4244	−284.276	−4.5834	0.673213	0.163918

Table 4. Cluster validity value for Data_D data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	**0.79551**	**0.33986**	0.115832	57.0075	−468.149	1.94358	0.691938	0.349093
3	0.694361	0.556088	0.191033	94.3278	−751.929	1.83221	0.67249	0.290824
4	0.68218	0.632649	0.111323	55.3446	−970.371	2.19397	0.689987	0.217821
5	0.68277	0.669871	0.13078	65.5763	−1173.53	1.62344	0.711313	0.179818
6	0.694526	0.673733	**0.10603**	**54.0547**	**−1309**	**2.83438**	**0.72675**	**0.14499**
7	0.656996	0.764082	0.189353	97.2691	−1308.05	1.38258	0.719184	0.159115
8	0.621141	0.849361	0.237335	122.717	−1237.77	0.133122	0.702011	0.170465
9	0.597376	0.915847	0.245337	128.261	−1198.91	−0.81367	0.693468	0.16993
10	0.578514	0.964591	0.25149	132.48	−1177.73	−1.42709	0.676501	0.17809
11	0.570492	0.99286	0.218359	116.054	−1169.23	−2.60876	0.665802	0.176957
12	0.558055	1.04528	0.276692	148.004	−1155.49	−3.46056	0.678192	0.171706
13	0.548339	1.06907	0.239029	129.712	−1132.3	−1.32462	0.659903	0.171441
14	0.535508	1.11233	0.221208	120.918	−1114.36	−2.45799	0.662066	0.17047
15	0.525838	1.15331	0.232902	128.373	−1096.1	−3.06898	0.661184	0.173171

Table 5. Cluster validity value for Data_E data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	**0.75552**	**0.39101**	0.15924	63.9458	−82.0912	1.94124	0.659599	0.400286
3	0.594601	0.705188	0.224432	90.3998	−196.361	2.25491	0.552711	0.366068
4	0.577277	0.799811	0.196535	79.779	−431.872	1.81867	0.633739	0.325125
5	0.605335	0.806627	0.090208	37.0885	−661.777	2.73353	0.723333	0.254295
6	0.646228	0.774014	0.078865	32.8561	−757.442	**3.62034**	0.769347	0.209718
7	0.655341	0.78293	0.081482	34.3136	−777.748	2.94469	0.76699	0.174396
8	0.670109	0.769567	**0.06264**	**26.7605**	**−789.89**	3.21537	**0.77084**	**0.13338**
9	0.646533	0.839803	0.194469	83.3159	−766.287	2.09655	0.764381	0.140938
10	0.625762	0.892571	0.21686	94.3074	−753.197	0.575882	0.749632	0.14626
11	0.61004	0.936029	0.22222	97.8051	−755.809	−0.96664	0.737718	0.150237
12	0.593829	0.987371	0.25876	114.249	−748.259	−2.20504	0.730073	0.156873
13	0.564428	1.05635	0.301355	134.831	−704.223	−3.88959	0.699977	0.163652
14	0.555949	1.08662	0.28858	130.323	−662.488	−3.88495	0.688237	0.164556
15	0.546322	1.12379	0.274431	124.713	−662.74	−5.34366	0.687569	0.163013

Table 6. Cluster validity value for the Iris data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	**0.89202**	**0.19606**	**0.05425**	**8.38702**	−401.073	**1.57435**	**0.8052**	**0.20931**
3	0.783196	0.395927	0.137108	21.9837	−449.735	1.43672	0.73829	0.228204
4	0.706524	0.56173	0.195753	32.0404	−475.062	0.246231	0.712571	0.21931
5	0.665464	0.675854	0.228209	38.3145	**−543.86**	−0.83608	0.703556	0.209765
6	0.597752	0.796956	0.301621	54.2332	−392.182	−0.07311	0.599421	0.228372
7	0.55697	0.907886	0.373145	68.3604	−399.107	−1.30135	0.597351	0.235752
8	0.534346	0.98472	0.25315	46.8747	−393.286	−2.27352	0.599586	0.235567
9	0.496311	1.07288	0.373904	72.6405	−332.005	−1.53992	0.580652	0.241001
10	0.472859	1.14613	0.328013	65.6562	−332.004	−2.87783	0.575366	0.243481
11	0.462675	1.19794	0.30272	61.6189	−330.858	−3.76337	0.58116	0.239209
12	0.450905	1.2475	0.395271	81.322	−329.749	−4.93582	0.584671	0.237494

Table 7. Cluster validity value for the Cancer data set

c	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
2	**0.8409**	**0.26668**	**0.11032**	**75.599**	−13508.1	**1.30428**	0.761968	**0.29563**
3	0.715577	0.50641	1.21502	833.701	−25728.2	−0.31611	0.772319	0.321573
4	0.640335	0.698262	7.99857	5490.88	**−31025**	−1.4738	**0.78493**	0.303156
5	0.489004	0.959756	802.516	552898	−16542	−2.40121	0.533705	0.37474
6	0.463938	1.06667	1162.68	801812	−20133.9	−3.40966	0.560177	0.368005
7	0.364448	1.29084	693.82	479709	−12118.5	−4.25905	0.479503	0.379245
8	0.351203	1.366	1307.98	905573	−14042.3	−5.2789	0.47525	0.382447
9	0.342588	1.43248	452.299	313439	−15818.6	−6.28647	0.497428	0.369175
10	0.301087	1.57226	1853.19	1287430	−12002.4	−7.05034	0.507574	0.3475
11	0.283688	1.64406	36308.1	25255200	−12494.3	−8.1859	0.515976	0.35236
12	0.274435	1.7006	2427.45	1689990	−13258.3	−9.21378	0.52687	0.353241
13	0.256133	1.78838	7506.32	5243300	−10872.6	−8.65539	0.526126	0.35927
14	0.251199	1.83203	3872.6	2708090	−11650.5	−9.67722	0.544633	0.359416
15	0.244489	1.89451	715.938	501612	−10822.7	−10.4992	0.571269	0.359949

Table 8. Values of preferred by validity indexes for the experimental data sets

Data set	$c_{optimal}$	*PC*	*PE*	*XB*	*Vk*	*FS*	*PACES*	*FSα*	*DN*
Data_A	3	3	3	3	3	3	3	3	3
Data_B	4	2	2	2	2	4	2	3	4
Data_C	5	5	2	5	5	5	5	4	5
Data_D	6	2	2	6	6	6	6	6	6
Data_E	8	2	2	8	8	8	6	8	8
iris	2 or 3	2	2	2	2	5	2	2	2
breast	2	2	2	2	2	4	2	3	2
accuracy	--	4/7	3/7	6/7	6/7	5/7	5/7	4/7	7/7

5. Conclusions

This paper applies fuzzy set theory to the clustering analysis, and further proposes the notion of FCM fuzzy set, which subject to the constraints of FCM. The cluster fuzzy degree and the lattice degree of approaching for the FCM fuzzy set is defined; the former is used as the measure of compactness and the latter used as the measure of separation of a fuzzy partition matrix obtained using FCM algorithm, and their functions in the process of validation are deeply analyzed. The two measures express two aspects of clustering process respectively. Compactness is used as a measure of the closeness or scattering of clusters, and separation as a measure of the isolation of clusters from one another. A good clustering result will have small intra-cluster compactness and large inter-cluster separation. The new adaptive cluster validity index, called *DN* is designed, which consists of two terms: *DC* and *NM*, where *DC* is a type of normalized cluster fuzzy degree and *NM* is

the maximum of lattice degree of approaching corresponding to the nearest pair of cluster centers. In order to take the effect of the two terms upon the cluster validation into comprehensive consideration, *DC* and *NM* are in the symmetric position, whereby rendering the cluster validity index *DN* to be able to adjust the action level of *DC* and *NM* adaptively. When *DC* and *NM* values are closer, they will have the similar effect upon *DN*; when the value of one of the terms is much smaller than that of the other, the term having the small value will have the greater effect upon the index. Also, this paper proposes the FCM validation algorithm, in which the FCM algorithm is repetitiously executed to obtain the optimal clustering result. Experimental results indicate that the *DN* cluster validity index is stable and adaptive. The future research work will include the further improvement of the validation efficiency as well as the practical applications of the cluster validity index.

References

[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[2] J. C. Bezdek, Numerical Taxonomy with Fuzzy Sets, J. Math.Biol. 1974, 1: 57–71.
[3] J. C. Bezdek, Cluster Validity with Fuzzy Sets, J. Cybernet. 1974, 3: 58–72.
[4] K. -L. Wu, M. -S. Yang, A Cluster Validity Index for Fuzzy Clustering, Pattern Recognition Letters 2005, 26: 1275–1291
[5] X. L. Xie, G. Beni, A Validity Measure for Fuzzy Clustering, IEEE Trans. Pattern Anal. Mach. Intell. 1991,13(8): 841–847.
[6] S. H. Kwon, Cluster Validity Index for Fuzzy Clustering, Electron, Lett. 1998, 34(22): 2176–2177
[7] Y. Fukuyama, M. Sugeno, A New Method of Choosing the Number of Clusters for the Fuzzy C-means Method, in: Proceedings of the Fifth Fuzzy Systems Symposium,1989: 247–250
[8] R. J. G. B. Campello, E. R. Hruschka, A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis, Fuzzy Sets and Systems, 2006, 157(21): 2858-2875
[9] D. -W. Kim, K. H. Lee, D. Lee, On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters, Pattern Recognition, 2004, 37: 2009–2025
[10] E. R. Hruschka, L. N. de Castro, R. J. G. B. Campello, Evolutionary Algorithms for Clustering Gene-expression Data, in: Proc. IEEE Internat. Conf. on Data Mining, Brighton, England, 2004: 403–406.
[11] L. A. Zadeh, A fuzzy set theoretic interpretation of linguistic hedges, J. Cybernet. 1972, 2(3): 4–34.
[12] N. R. Pal, J. C. Bezdek, On Cluster Validity for the Fuzzy C-means Model, IEEE Trans. Fuzzy Syst. 1995, 3(3): 370–379.
[13] J. -L. FAN , C. -M. WU, Subsethood Measures Applied to Clustering Validity Judgement, Fuzzy Systems and Mathematics, 2002, 16(1): 80–86 (in Chinese)
[14] http://www.ics.uci.edu/~mlearn/MLRepository.html

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported in part by the National Natural Science Fund of China under Grant 60501006

CHEN Duo, received his B.S. degree in 1984 from Hebei University of Technology, Tianjin, China , received the M.S. degree in 1995 from Yanshan University, Qinhuangdao, China. He is an associate professor in the Computer Center of Tangshan College, Tangshan, China. He is currently working toward the Ph.D degree in computer science and engineering in Xi'an University of Technology, Xi'an, China. His research interests include data mining, intelligent computing and power electronics and electric drive, etc.

LI Xue, received her B.S. degree in 1998 and the M.S. degree in 2004 respectively from Xi'an University of Technology, Xi'an, China. She is an assistant in International Business School of Shaanxi Normal University, Xi'an, China. She is currently working toward the Ph.D degree in computer science and engineering in Xi'an University of Technology, Xi'an, China. Her research interests include intelligent computing, data mining, and electronic commerce, etc.