Clustering Validity Evaluation Functions of Fuzzy C-means Clustering Algorithm

Guan Wang, Cheng Xing *, Jie-Sheng Wang, Hong-Yu Wang, Jia-Xu Liu

Abstract—**Fuzzy C-means (FCM) clustering algorithm is a method mainly applied to machine learning and data mining. It can cluster objects into a limited number of categories according to their similarity degree without much prior knowledge. However, FCM clustering algorithm must first give a predefined number of clusters. Therefore, it is crucial to use clustering effectiveness function to get the optimal cluster number. Therefore, partition coefficient, partition entropy, separation index, Bensaid clustering validity function, Xie and Beni clustering validity function, Dunn clustering validity function and the improved Dunn clustering validity function were selected. Clustering experiments were conducted on three typical UCI data sets in view of FCM clustering algorithm. Finally, different fuzzy indexes are used to evaluate the validity of clustering.**

Index Terms—**fuzzy c-means clustering algorithm, clustering validity function, cluster analysis**

I. INTRODUCTION

CLUSTER analysis plays an important role in data mining, which is a branch of machine learning. As a learning way without supervision, clustering technology can extract and classify information from data sets without giving prior knowledge of data sets. The core idea of clustering is to classify analogous samples into one category as far as possible and separate dissimilar samples as far as possible [1]. Researches on clustering are mainly divided into two types: hard clustering and fuzzy clustering. Hard clustering, such as K-means clustering algorithm, is based on the idea of "zero equals one", and the samples in the data set can only be divided into one category [2]. subsequently, Ruspini introduced the idea of fuzzy into hard clustering and proposed fuzzy clustering. Among them, Fuzzy C-means (FCM) clustering algorithm [3] introduces the concept of membership degree. Compared with hard clustering algorithm, it is more in line with the real data set. Just because of this, FCM clustering algorithm has been applied in medicine, biology, chemistry, statistics and other fields by virtue of its simple design, wide application range and strong ability to adapt to data sets [4-5]. Zhang et al. raised a two-stage optimal configuration model of CCHP system in view of optimized FCM clustering algorithm in 2021 [6]. Li et al. used the improved FCM algorithm to extract plateau lake water [7].

The clustering validity is a key condition to determine the effect of fuzzy clustering. When clustering, the optimal clustering number should be given in advance. Because judge optimal clustering number and improve the quality of clustering results, the study on clustering validity has been a pivotal branch of clustering problems and attracted the attention of researchers. The issue of clustering validity mainly lies in how to choose a clustering validity function to judge the best clustering number of data set. Now, many clustering validity functions have been raised and used. In 2019, Zhu et al. raised a clustering validity function in view of ratio form [8]. In 2021, Naderipour et al. raised a fuzzy clustering validity function in view of topological structure and node attributes [9]. In the same year, Wang et al. raised a new validity function in view of intra-class compactness and inter-class separation [10]. However, no clustering validity function can be applied to all data sets because of the ever-changing structure and sizes of data sets.

In the essay, the FCM clustering algorithm and some common clustering validity evaluation functions are simulated and analyzed. Firstly, the basic flow of FCM clustering algorithm is introduced and the influence of different clustering numbers given in advance is explained through experiments. After that, partition coefficient, partition entropy, separation index, Bensaid clustering validity function, Xie and Beni clustering validity function, Dunn clustering validity function and the improved Dunn clustering validity function were selected, and then experimental comparison was conducted on three UCI data sets. Finally, the stability of different clustering validity functions is explored by changing the fuzzy index. The arrangement of this essay is as follows. In Section 2, the FCM clustering algorithm is introduced. The clustering validity function is introduced in details in Section 3. The simulation and experiments analysis are described in Section 4. Finally, the conclusion is illustrated in the last part.

II. FUZZY C-MEAN CLUSTERING ALGORITHM

Fuzzy C-means (FCM) clustering algorithm is one of the most popular clustering algorithms in fuzzy clustering. This

Manuscript received August 10, 2021; revised February 28, 2022. This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. LJKZ0293 and LJKZ0307), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Guan Wang is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: 480433838@qq.com).

Cheng Xing is a Ph.D candidate in School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114044, P. R. China (Corresponding author, phone: 86-0412-2538246; fax: 86-0412-2538244; e-mail: xingcheng0811@163.com).

Jie-Sheng Wang is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: wang_jiesheng@126.com).

Hong-Yu Wang is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: wanghongyuww@126.com).

Jia-Xu Liu is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: 1849226542 @qq.com).

algorithm is proposed by Bezdek and is one of the methods widely used in pattern recognition.

FCM algorithm takes n data objects of data X (X={ x_1, x_2, ..., x_n }) is divided into c fuzzy clusters to find the least objective function as shown in Eq. (1).

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\left\|x_j-v_i\right\|^2 \quad (1)$$

where, $J_m(U,V)$ is the clustering criterion of square error, the least value of $J_m(U,V)$ is minimum square error. V={ $v_1, v_2, ..., v_n$ } is the set of partition cluster centers in dataset X. The parameter c represents the number of divided clusters. m is the fuzzy index to adjust the fuzziness of the category degree of each group of data within the range of $m\in(1,\ \infty)$. $\left\|x_j-v_i\right\|$ represents the Euclidean distance between the data object x_j and the cluster center v_i. u_{ij} ($0\le u_{ij}\le 1$) is the membership degree of x_j and v_i. $u_{ij}\in U$ is a fuzzy membership degree matrix of $c\times n$ that should meet the following restrains.

$$\sum_{i=1}^{c}u_{ij}=1,\quad j=1,2,\ldots,n \quad (2)$$

$$0<\sum_{j=1}^{c}u_{ij}<n,\quad i=1,2,\ldots,c \quad (3)$$

Clustering validity is the issue of finding the best solution under the condition of minimal J_m. In fact, if clustering analysis wants to make great contribution to engineering application, it must pay more attention to the validity of clustering. As J_m decreases monotonically with the decrease of c, an effective evaluation standard is needed. The procedure of the FCM clustering algorithm is described as follows.

Step 1: Given the clustering parameters c and fuzzy factor m (usually between 1.5 and 2.5). When m=1, FCM algorithm is equal to K-means algorithm. When m infinity approaches 1, FCM tends to harden clustering algorithm more and more. On the contrary, as m tends to infinity, all data objects x_j will coincide with cluster center v_i, and the membership degree of data object x_j to each cluster will be the same, with a value of 1/c.

Step 2: Initialize the fuzzy partition membership matrix and make it satisfy Eq. (2) and Eq. (3).

Step 3: Update cluster center $V^{(\gamma+1)}=\{v_1, v_2, ..., v_c\}$.

$$v^{(\gamma+1)}=\frac{\sum_{j=1}^{n}u^{(\gamma)}{}_{ij}{}^{m}\cdot x_i}{\sum_{j=1}^{n}u^{(\gamma)}{}_{ij}{}^{m}},\quad i=1,2,\ldots,c \quad (4)$$

Step 4: Update the fuzzy partition matrix $U^{(\gamma+1)}=(u_{ij})_{c\times n}$ according to Eq. (5).

$$u_{ij}^{(\gamma+1)}=\left[\sum_{k=1}^{c}\left(\frac{\left\|x_j-v_i\right\|^{2(\gamma)}}{\left\|x_j-v_k\right\|^{2(\gamma)}}\right)^{2/(m-1)}\right]^{-1} \quad (5)$$

where, i=1, 2, ... , c , j=1, 2, ... , n.

Step 5: Calculate e=$\left\|U^{(\gamma+1)}-U^{(\gamma)}\right\|$. When $e\le\eta$ (η is a threshold usually from 0.001 to 0.01), the algorithm stops and the final clustering result is calculated. Otherwise $\gamma=\gamma+1$ and repeat from Step 2.

Based on the FCM algorithm, a simple clustering simulation on the Motorcycle data set was performed to research the influence of the predicted number of clustering c on the clustering result. There are 133 samples in the Motorcycle data set, and the optimal number of clustering c=4. The simulation results are shown in Fig. 1. When the number of clusters is selected c=2, the clustering results are shown in Fig. 2.

The red circle represents the cluster center then the blue dot represents the sample of the dataset. From the comparison in Fig. 1 and Fig. 2, it can be obviously found that the selection of clustering number c optimizes the results of data clustering. In algorithm calculation, FCM clustering algorithm aggregates data according to the given number of clusters to synthesize the number of classes that the algorithm thinks should be reasonable, but the actual number of clusters is not the input value. Obviously, Fig. 2 incorrectly segmented the due construction of the data set, resulting in the wrong clustering effect, which is not the problem of the algorithm itself, but the problem of the initial parameters. Therefore, for the seek of solving the problem of the number of clusters and make the clustering effect of the clustering algorithm reasonable and effective, it is necessary to conduct advance processing on the data set and obtain the best number of clusters, so the study on the validity of clustering is also very important.

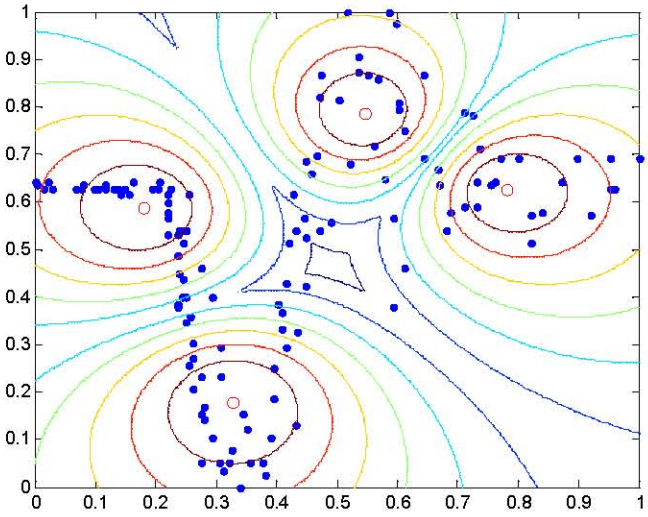


Fig. 1 FCM simulation diagram when the number of clustering centers is 4.

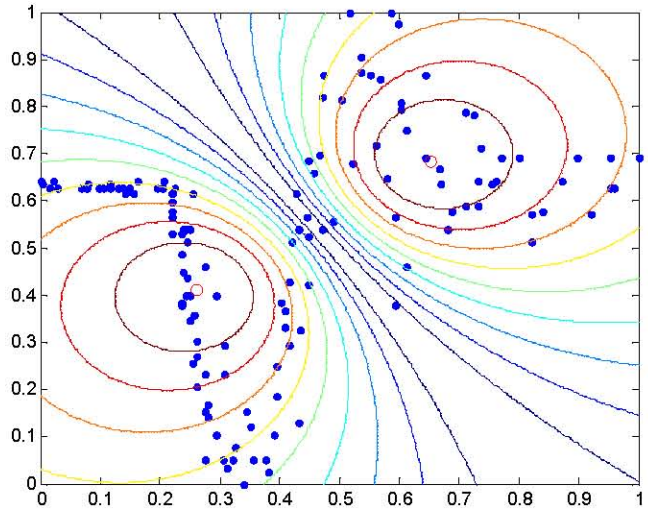


Fig. 2 FCM simulation diagram when the number of clustering centers is 2.

III. Clustering validity function

A. Clustering Validity Function Based on Fuzzy Partition

1) Partition Coefficient (V_{PC}) [11]

Bezdek defines V_{PC} to measure overlap between clusters, as expressed in Eq. (6).

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^2 \tag{6}$$

The value of V_{PC} is between 1/c and 1. If you want to find a good partition, you should find a partition with the maximal partition coefficient. The main advantage of partition coefficient is simple, but the weakness is that it is monotonously decreasing with the increase of c.

2) Partition Entropy (V_{PE}) [12]

Bezdek also uses V_{PE} to measure the fuzziness of cluster division. This index is similar to V_{PC} and is defined in Eq. (7).

$$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\left[u_{ij}\log_a(u_{ij})\right] \tag{7}$$

V_{PE} is a value between 0 and 1-V_{PC}. Bezdek proves that for all probability cluster partitions. The disadvantage also decreases monotonically as c increases, which is lack of connection to data structure. So the smaller the number, the better.

B. Clustering Validity Function Based on Geometric Structure

1) Xie-Beni fuzzy clustering validity function (V_{XB})[13]

This function is a cluster validity index proposed by Xie and Beni, which measures the proportion of total change and cluster separation within a cluster with high utilization rate and excellent effect. The V_{XB} index is expressed in Eq. (8).

$$V_{XB}(\mathrm{U},\mathrm{V},\mathrm{c}) = \frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^m\left\|v_i - x_j\right\|^2}{\min_{i\neq j}\left\|v_i - v_j\right\|^2} \tag{8}$$

V_{XB} is the first clustering validity function that takes the structure of data set into account, which is the ratio of intra-class compactness to inter-class separation. Obviously, the more discrete between the classes, the more compact within the classes, indicating the better clustering results.

Therefore, a smaller V_{XB} value corresponds to a better number of partition categories. This index has the following shortcomings.

(1) When the number of categories is close to the number of samples, namely $c \to n$, the center of the class is the sample point itself, namely $v_i \to x_j$. At this point, $V_{XB}(\mathrm{U},\mathrm{V},\mathrm{c}) \to 0$, V_{XB} will gradually lose its judgment ability.

(2) When the fuzzy index $m \to \infty$, all FCM class centers coincide, that is, the fuzzy is too high, and the membership degree of samples belonging to all classes is equal. Obviously, when $\lim_{m\to\infty}\left\|v_i - v_j\right\|^2$, $V_{XB}(\mathrm{U},\mathrm{V},\mathrm{c}) \to \infty$.

2) Bensaid validity indicators (V_{SC}) [14]

Bensaid et al. proposed V_{SC} in 1996. V_{SC} is the ratio of the sum of the compactness of a cluster to the sum of the degree of separation, normalizing the sum of individual cluster validity measures by the fuzzy cardinality of each cluster. Its definition is shown in Eq. (9).

$$V_{SC}(U,V,c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{n} u_{ij}\left\|v_i - x_j\right\|^2}{n_i\sum_{i=1}^{c}\left\|v_i - v_j\right\|^2} \tag{9}$$

where, $n_i = \sum_{j=1}^{n} u_{ij}$. In this index, the measure function of the separation degree $\min_{i\neq j}\left\|v_i - v_j\right\|^2$ in V_{XB} is replaced by the function $\sum_{i=1}^{c}\left\|v_i - v_j\right\|^2$, which can be used to compare the same number of clusters but different dispersion.

The measure of intra-class compactness is changed from the average of the whole and the upper. This is replaced by the average sum of compactness within a class. V_{SC} is useful when comparing different partitions with the same number of clusters, and the lower the V_{SC} value, the better the partition.

3) Separation Index (V_S)

Different from V_{SC} index, V_S uses minimal clustering to partition data, which is defined as Eq. (10).

$$V_S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(u_{ij}\right)^2\left\|x_j - v_i\right\|^2}{N\min_{i,j}\left\|v_k - v_i\right\|} \tag{10}$$

It should be noted that the only difference between V_{SC}, V_S, and V_{XB} is the clustering of the separated methods.

C. Clustering Validity Function Based on Hard Partition

1) Dunn clustering validity function (V_{DI})[15]

V_{DI} is an internal validity function, which refers to the ratio of the shortest distance between the data objects of any two clusters to the largest distance in any cluster. The larger the V_{DI} is, the larger the distance between clusters is; the smaller the V_{DI} is, the smaller the distance between clusters is. Therefore, the efficiency is better when V_{DI} is at its maximum. Its definition is shown in Eq. (11).

$$V_{DI} = min_{i\in c}\left\{min_{j\in c, i\neq j}\left\{\frac{min_{x\in C_i, y\in C_i} d(x,y)}{max_{k\in c}\{max_{x,y\in c} d(x,y)\}}\right\}\right\} \tag{11}$$

This metric was raised to distinguish compact and well-separated clusters. The disadvantage of V_{DI} is that it becomes computationally extensive as c and m increases. Therefore, V_{DI} has a good clustering effect on discrete points, but a poor clustering effect on circular distributed data.

2) Alternative Dunn Index based on V_{DI} (V_{ADI})

V_{ADI} defined in Eq. (12) is an improvement on V_{DI}, which makes calculation easier.

$$V_{ADI} = min_{i\in c}\left\{min_{j\in c, i\neq j}\left\{\frac{min_{x\in C_i, y\in C_i}\left|d(y,v_i) - d(x_i,v_i)\right|}{max_{k\in c}\{max_{x,y\in c} d(x,y)\}}\right\}\right\} \tag{12}$$

In the case of overlapping clusters, both V_{ADI} and V_{DI} values are less reliable because of the method of re-partitioning the results using physical partitions. Smaller V_{ADI} value corresponds to better clustering partitioning.

IV. Simulation Experiment and Result Analysis

A. Selection of Experimental Data

In this paper, seven cluster validity indexes (V_{PC}, V_{PE}, V_{XB}, V_{SC}, V_S, V_{DI}, V_{ADI}) are selected to verify the clustering validity of data, and three kinds of data that are commonly used in fuzzy clustering validity problem are selected, which are Iris flowers, Wine and Wheat Seeds data sets from UCI database. Here are the brief properties of these three data sets.

(1) Iris data set. Iris data set was collected by Fisher in 1936. This data set measured the calyx length, calyx width, petal length and petal width of Iris. The data set includes 150 data objects in total, each of which contains 4 attributes. The data objects can be divided into 3 types of Iris, and each type of Iris occupies 50 data objects. The size of these four attributes can be used to judge which category of Iris flowers in three categories. Partial samples of the Iris datasets are listed in Table 1.

(2) Wine data set. This data set is based on a study of three different varieties of wine from the same area. The data set includes 178 data objects, which are roughly divided into three categories. The first category contains 59 data objects, the second category contains 71 data objects, and the third category contains 48 data objects. Each data object has 13 attributes of ethanol, malic acid, ash content, ash alkalinity, magnesium content, phenolic compounds, flavonoids, yellow alkanes of phenol, anthocyanins, color strength, hue, dilute the wine of OD280 and proline. By detecting the 13 attributes content in wine can judge wine belongs to the kind of wine. A partial sample of the Wine dataset is listed in Table 2.

(3) Seeds data set. This dataset is used to predict wheat seeds of different varieties, and the given quantitative data are seeds. There are 210 data objects in total, and each data includes 7 attributes, namely region A, periphery P, compactness, kernel length, kernel width, symmetry coefficient and core slot length. These seven attributes are used to determine which wheat seeds belong to which type. All data objects can be divided into three categories of wheat seeds, and each category contains 70 data objects. A partial sample of the Seeds dataset is listed in Table 3.

B. Cluster Validity Function Simulation and Result Analysis

1) Iris data set

In the simulation experiment of clustering effectiveness on Iris data set, the selected c value changes from 2 to 10, and the fuzzy index m=2. Simulation results of clustering validity function experiment are shown in Fig. 3(a)-(g). It can be seen from Fig. 3 (a) and Fig. 3 (b), V_{PC} and V_{PE} indices change monotonically. V_{PC} takes the maximum value and V_{PE} takes the minimum value, so they are c=2.

Table 1. Iris data set

Feature	Calyx length	Calyx width	Petal length	Petal width
Sample of Iris cerasus 1	5.1	3.5	1.4	0.2
Sample of Iris cerasus 2	4.9	3.4	1.4	0.2
...	...	...	...	...
Sample of iris variegata 1	5.9	3.0	4.2	1.5
Sample of iris variegata 2	6.0	2.2	4.0	1.0
...	...	...	...	...
Sample of Iris Virginia 1	6.9	3.1	5.6	2.4
Sample of Iris Virginia 2	6.4	3.1	5.5	1.8
...	...	...	...	...

Table 2. Wine dataset

Feature	Sample 1	Sample 2	Sample 3	Sample 4	...	Sample 178
Ethanol	14.23	13.86	12.72	12.42	...	13.52
Malic acid	1.71	1.35	1.81	1.61	...	3.17
Ash	2.43	2.27	2.22	2.19	...	2.72
The ash alkalinity	15.6	16	18.8	22.5	...	23.5
Magnesium content	127	98	86	108	...	97
Phenolic compound	2.8	2.98	2.2	2.0	...	1.55
Flavonoids	3.06	3.15	2.53	2.09	...	0.52
Yellow alkanes of phenol	0.28	0.22	0.26	0.34	...	0.5
Anthocyanins	2.29	1.85	1.77	1.16	...	0.55
Color intensity	5.64	7.22	3.9	2.06	...	4.35
Tonal	1.04	1.01	1.16	1.06	...	0.89
OD280	3.92	3.55	3.14	2.96	...	2.06
Proline	1065	1045	714	345	...	520

Table 3. Seeds dataset

Feature	Seed samples 1	Seed samples 2	Seed samples 3	Seed samples 4	...	Seed samples 210
Area A	15.26	14.88	15.38	16.16	...	12.30
Surrounding P	14.84	14.57	14.90	15.33	...	13.73
degree	0.8710	0.8811	0.8706	0.8830	...	0.8081
Kernel length	5.763	5.554	5.884	5.776	...	5.394
Kernel width	3.312	3.333	3.268	3.408	...	2.745
Symmetrical coefficient	2.221	1.018	4.462	4.972	...	4.825
Nuclear slot length	5.220	4.956	5.795	5.847	...	5.220

In Fig. 3(c), it can be found that V_S index reaches the maximum value when c=5, so c=5. In Fig. 3(d), it can be found V_{SC} index reaches local minimum value at c=3. In Fig. 3(e), it is obvious that V_{XB} index reaches the local minimum value at c=8, so c=8 is chosen. In Fig. 3(f), it is obvious that V_{DI} index reaches its maximum value at c=3, so c=3; It can be seen from Fig. 3(g) that V_{ADI} index does not change significantly after c=5, and the best number of clusters cannot be determined. In line with the introduction of Iris data set, it can be known that Iris data set is divided by the number of clusters c=3 as the best cluster. Therefore, it can be found that V_{DI} and V_{SC} are the most effective for this data cluster class. Fig. 3 is the simulation result of clustering validity function obtained by fixed value of fuzzy index m=2, and its influence on clustering validity index is observed when different m values are taken. Since V_{DI} and V_{ADI} belong to the hardening fraction, the fuzzy index does not affect the final calculation result, so they are not calculated. The experimental results are listed in Table 4. It can be found in Table 4 that at different m values, V_{PC} and V_{PE} are all c=2 except that the optimal clustering number of V_{PC} becomes c=3 when m=1.5.

Therefore, it is considered that the effect of m on V_{PC} and V_{PE} is very weak, and the lack of connection between V_{PC} and V_{PE} and the geometry of the dataset can also be found. V_{XB} only finds the optimal number of clusters when m=1.2, and fails when m increases. V_{SC} is invalid only when m=5. V_S is only valid if m=5. In summary, it can be seen that V_{SC} is the most effective method to estimate the number of clusters for Iris data sets. Because V_{SC} changes with m of the data set, the change is very weak, the probability of finding the best number of clusters is highest.

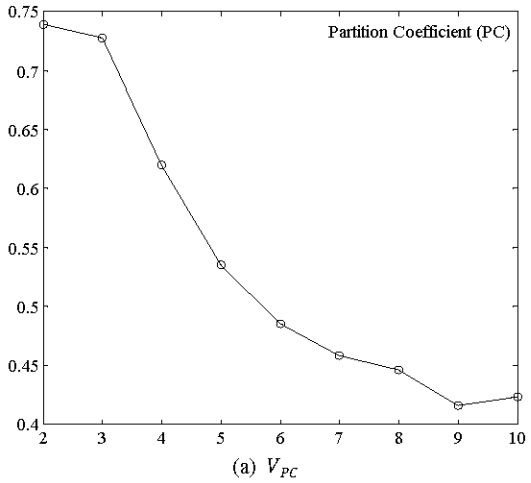
2) Wine data set

For the clustering effectiveness simulation experiment of Wine data set, the simulation results of clustering validity function experiment are shown in Fig. 4(a)-(g). The abscissa of the simulation diagram is the number of clustering c, and the ordinate is the value of clustering validity function. V_{PC} and V_{PE} in Fig. 4(a) and Fig. 4(b) lack direct connection with data, and the clustering division corresponding to a larger V_{PC} value is c=2, while the clustering division corresponding to a smaller V_{PE} value is c=3, but the accuracy of the partitioning results is low.

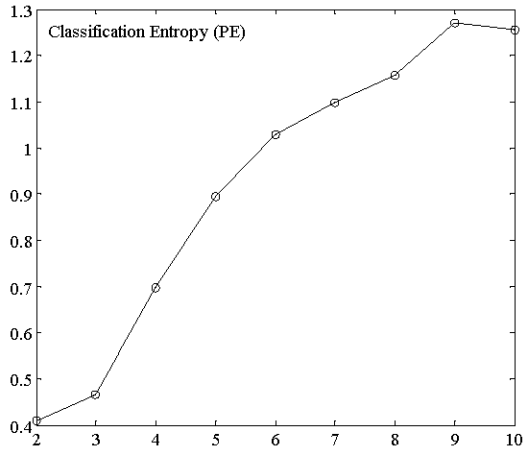
TABLE 4. CLUSTERING VALIDITY INDEX VALUES UNDER DIFFERENT M VALUES

c	2	3	4	5	6	7	8	9	10
m=1.2									
V_{PC}	**0.9269**	0.9194	0.8807	0.8485	0.8278	0.8170	0.8088	0.7923	0.7951
V_{PE}	**0.0732**	0.0816	0.1190	0.1315	0.1622	0.1930	0.1952	0.2077	0.2049
V_S	0.0058	0.0084	0.0120	0.0140	0.0124	0.0127	**0.0132**	0.0117	0.0125
V_{SC}	0.8696	**0.8581**	1.2626	1.6548	1.1757	1.2551	1.3488	1.1012	1.1622
V_{XB}	3.1078	**3.0984**	3.1204	3.1783	4.4718	4.1797	3.3119	4.5107	3.8302
m=1.5									
V_{PC}	0.8415	**0.8506**	0.7547	0.6932	0.6554	0.6354	0.6230	0.5957	0.6013
V_{PE}	**0.1585**	0.1594	0.2452	0.3566	0.4444	0.5646	0.5770	0.5043	0.5987
V_S	0.0051	0.0066	0.0090	**0.0102**	0.0088	0.0086	0.0087	0.0076	0.0079
V_{SC}	0.7589	**0.6835**	0.9614	1.2137	0.8361	0.8673	0.8992	0.7236	0.7434
V_{XB}	2.6841	2.6036	2.4671	2.3803	3.1661	2.9278	**2.1950**	2.9844	2.4544
m=2									
V_{PC}	**0.7389**	0.7279	0.6199	0.5350	0.4856	0.4581	0.4460	0.4157	0.4230
V_{PE}	**0.2611**	0.2721	0.3801	0.4650	0.5144	0.5419	0.5540	0.5843	0.5770
V_S	0.0042	0.0049	0.0062	**0.0066**	0.0054	0.0051	0.0049	0.0042	0.0042
V_{SC}	0.6302	**0.5137**	0.6603	0.7787	0.5182	0.5174	0.5097	0.4021	0.4020
V_{XB}	2.2251	2.1629	1.8913	1.6963	2.0805	1.9021	**1.3301**	1.8163	1.4260
m=5									
V_{PC}	**0.4662**	0.4553	0.3230	0.2332	0.1809	0.1523	0.1479	0.1359	0.1422
V_{PE}	**0.5338**	0.5447	0.6770	0.7668	0.8191	0.8477	0.8521	0.8641	0.8578
V_S	0.0023	**0.0024**	0.0021	0.0016	0.0011	0.0010	0.0009	0.0006	0.0007
V_{SC}	0.3452	0.2356	0.1938	0.1608	0.1027	0.0912	0.0802	**0.0594**	0.0595
V_{XB}	2.2251	2.1629	1.8913	1.6963	2.0805	1.9021	**1.3301**	1.8163	1.4260

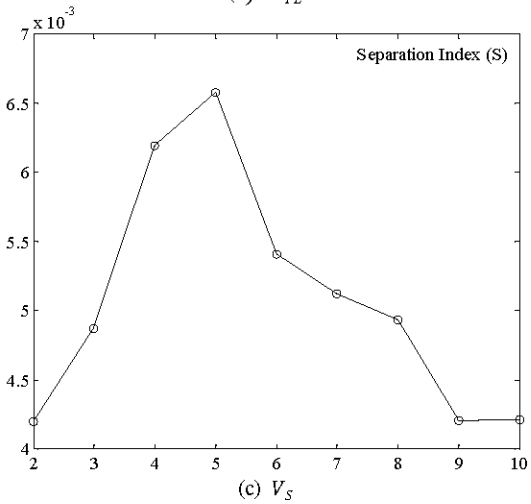
As shown in Fig. 4(c), V_{XB} minimizes at c=10, so c=10 is selected. As shown in Fig. 4(d), V_{SC} is monotonically decreasing on the whole and the minimum number is not clear, so the indicator is considered invalid. In Fig. 4(e), it can be found that V_S is the best cluster at c=2. In Fig. 4(f) and Fig. 4(g), it can be found that V_{DI} reaches its maximum value at c=2, so the number of clusters is c=2. V_{ADI} reaches local minimum value at c=3, so c=3 is selected as the clustering division. According to the introduction of Wine data set, the optimal clustering number is 3, so it is considered that the division of Wine data set V_{ADI} is the most reliable. Then, the influence of different fuzzy indexes on various validity indexes was observed. The experimental results are listed in Table 5.



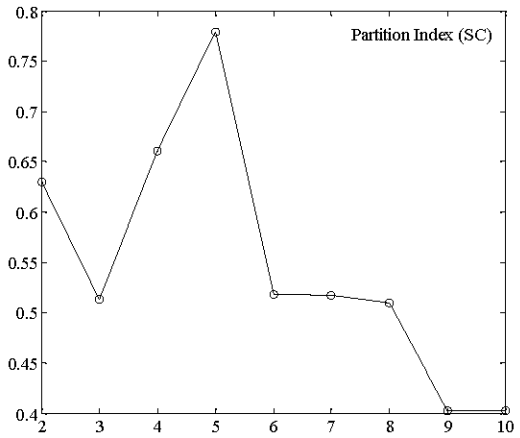
(a) V_{PC}



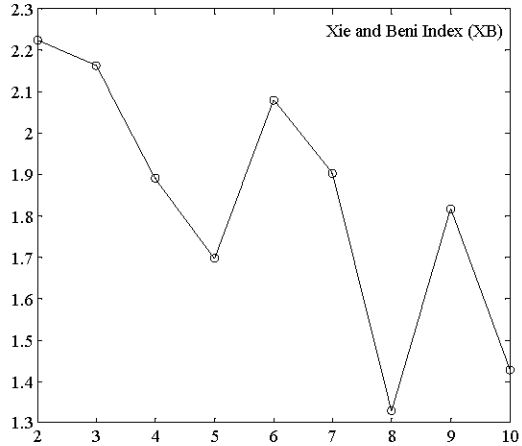
(b) V_{PE}



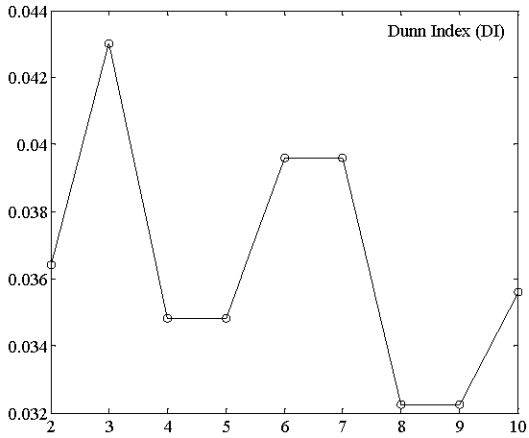
(c) V_S



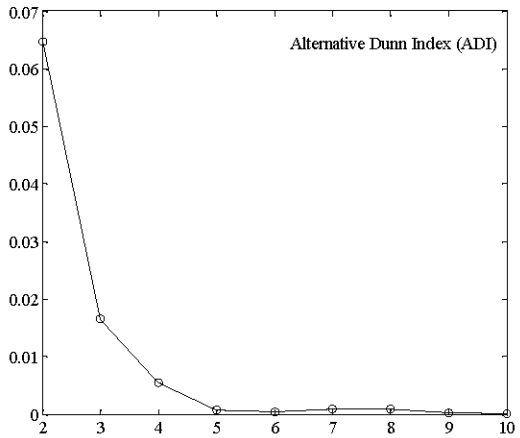
(d) V_{SC}



(e) V_{XB}

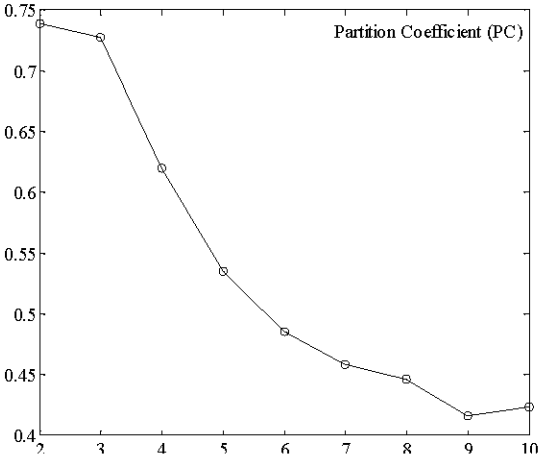


(f) V_{DI}

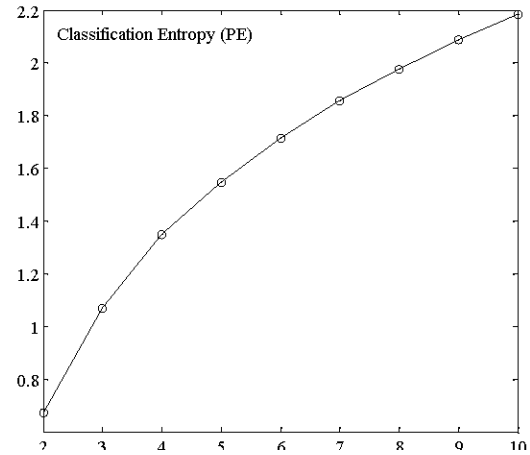


(g) V_{ADI}

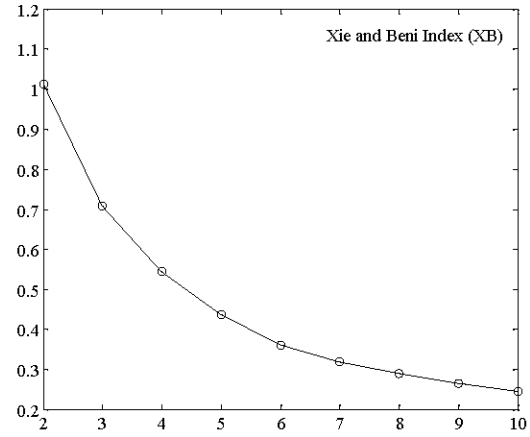
Fig. 3 Simulation results of clustering validity of Iris data set.



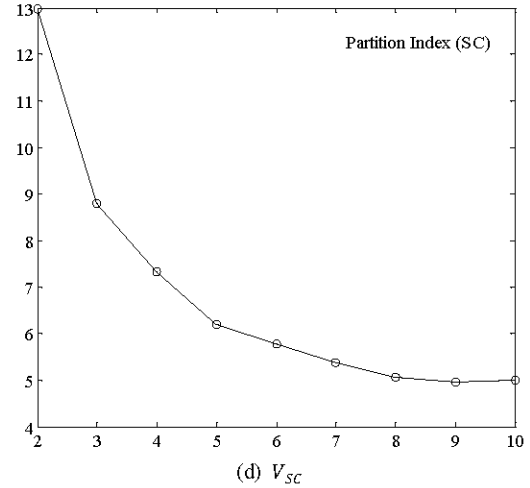
(a) V_{PC}



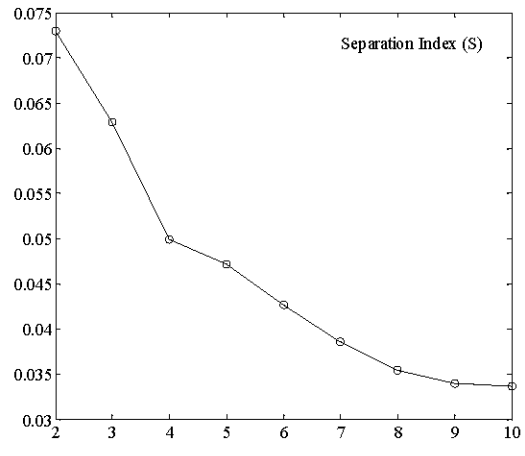
(b) V_{PE}



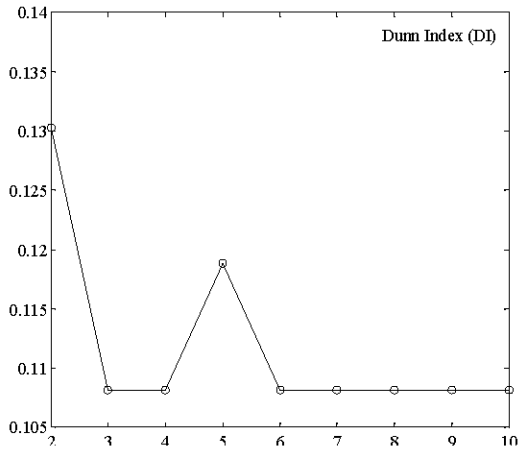
(c) V_{XB}



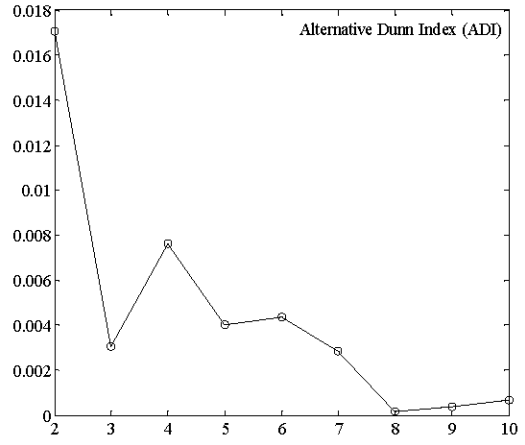
(d) V_{SC}



(e) V_S



(f) V_{DI}



(g) V_{ADI}

Fig. 4 Simulation results of clustering validity of Wine data set.

It can be found in Table 5 that, under different m values, only V_S index found the best division when m =1.2, while the best cluster division was not found for the rest indexes. Therefore, it can be found that selecting V_S as the clustering index for Wine data sets is far more reliable than other clustering validity functions.

3) Seeds data set

Simulation experiment on clustering effectiveness of Seeds data set. Simulation results of clustering validity function experiment are shown in Fig. 5(a)-(g). The abscissa of Fig. 5 is the number of clustering, and the ordinate is the value of clustering validity function. As shown in Fig. 5(a) and Fig. 5(b), V_{PC} and V_{PE} change

monotonically, so the number of clustering of both is c=2. It is found in Fig. 5(c) that the overall change of V_{SC} is monotonically decreasing, and the best number of clusters is c=9. It can be found in Fig. 5(d) that V_S achieves the maximum value when the number of clusters is c=3, thus realizing the optimal number of clusters. In Fig. 5(e), it can be found that the optimal number of V_{XB} clusters is c=9. It can be clearly seen from Fig. 5(f) that V_{DI} index achieves the maximum value at c=3, so the optimal number of clusters is 3. It can be found from Fig. 5(g) that V_{ADI} reaches the local minimum value at c=3, but the overall minimum value is c=8, so the clustering effectiveness is only for reference.

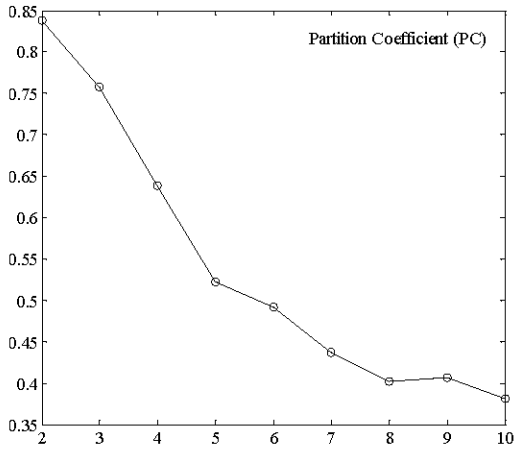
In summary, the most effective clustering validity functions for Seeds data sets should be V_S and V_{DI}. In order to judge whether V_S is always valid for Seeds data set and find the optimal clustering validity function by comparison, the experiment then observed the influence of various validity indexes on different fuzzy index values. The experimental results are listed in Table 6. It can be obviously found in Table 6 that V_S index has a better effect on Seeds data clustering than other five indexes under different values. Although the optimal clustering division was not found at m=1.2 and m=1.5, the optimal clustering division was found at m=2 and 5. V_{PC} and V_{PE} indexes still lack direct connection with data structure, so it can be seen that the change of fuzzy index has very little influence on V_{PC} and V_{PE}, so the clustering results obtained can only be used as reference. The V_{XB} index becomes invalid as the value increases. V_{SC} index does not obtain the optimal number of clusters =3. Therefore, V_S is best index to divide Seeds dataset.

V. Conclusions

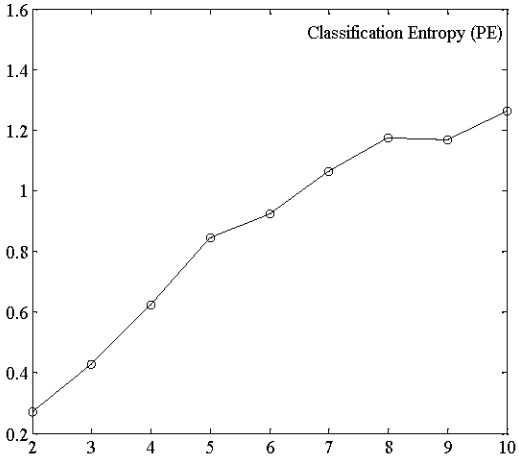
In this paper, several typical clustering validity functions were used for simulation analysis of Iris data set, Wine data set and Seeds data set in UCI database based on FCM clustering algorithm. It was found that V_{SC} and V_{DI} were the most accurate methods to judge the number of clusters in Iris data set, V_{ADI} was the most reliable method to classify Wine data set, and V_S and V_{DI} were the best methods to classify Seeds data set. Finally, by changing the fuzzy index m, it is found that V_{SC} and V_S are more stable than other validity functions, and the effect of dividing three data sets is the best.

Table 5. Clustering validity index values under different m values

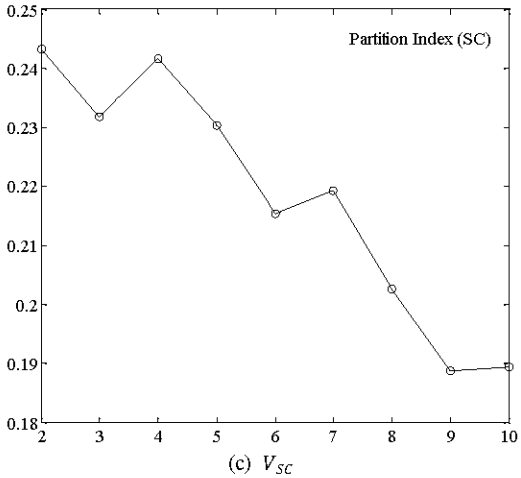
c	2	3	4	5	6	7	8	9	10
m=1.2									
V_{PC}	**0.8745**	0.8087	0.7652	0.7363	0.7122	0.6928	0.6764	0.6623	0.6498
V_{PE}	**0.1254**	0.1910	0.2347	0.2635	0.2879	0.3071	0.3234	0.3379	0.3503
V_S	0.0965	**0.0980**	0.0874	0.0909	0.0887	0.0856	0.0832	0.0838	0.0866
V_{SC}	17.177	13.691	12.773	11.793	11.842	11.687	**11.564**	11.858	12.470
V_{XB}	1.7332	1.6531	1.5820	1.4720	1.3916	1.3675	1.3576	1.3488	**1.3346**
m=1.5									
V_{PC}	**0.7172**	0.5907	0.5154	0.4696	0.4334	0.4054	0.3832	0.3644	0.3487
V_{PE}	**0.2829**	0.4092	0.4845	0.5303	0.5665	0.5945	0.6167	0.6355	0.6512
V_S	**0.0868**	0.0829	0.0707	0.0709	0.0671	0.0632	0.0601	0.0594	0.0604
V_{SC}	15.451	11.582	10.357	9.2414	9.0320	8.7141	**8.4564**	8.5231	8.8257
V_{XB}	1.4135	1.1972	1.0529	0.9222	0.8278	0.7785	0.7439	0.7145	**0.6860**
m=2									
V_{PC}	**0.5188**	0.3544	0.2712	0.2284	0.1966	0.1746	0.1584	0.1457	0.1359
V_{PE}	**0.4811**	0.6455	0.7287	0.7715	0.8033	0.8255	0.8476	0.8541	0.8640
V_S	**0.0729**	0.0628	0.0499	0.0472	0.0426	0.0386	0.0354	0.0340	0.0336
V_{SC}	12.977	8.7896	7.3230	6.1844	5.7812	5.3737	5.0498	**4.9461**	4.9916
V_{XB}	1.0136	0.7085	0.5439	0.4360	0.3613	0.3184	0.2886	0.2652	**0.2451**
m=5									
V_{PC}	**0.0874**	0.0233	0.0103	0.0094	0.0094	0.0108	0.0130	0.0151	0.0170
V_{PE}	**0.9126**	0.9767	0.9897	0.9906	0.9906	0.9892	0.9870	0.9849	0.9830
V_S	**0.0269**	0.0130	0.0069	0.0050	0.0036	0.0027	0.0021	0.0017	0.0015
V_{SC}	4.7827	1.8053	0.9785	0.6203	0.4485	0.3361	0.2619	0.2164	**0.1871**
V_{XB}	0.1629	0.0441	0.0190	0.0152	**0.0125**	0.0130	0.0151	0.0172	0.0190



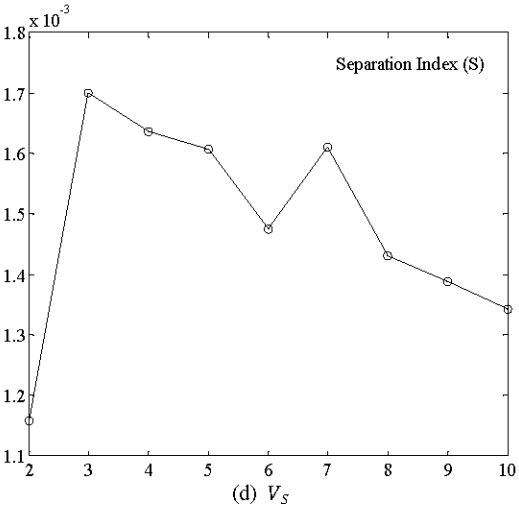
(a) V_{PC}



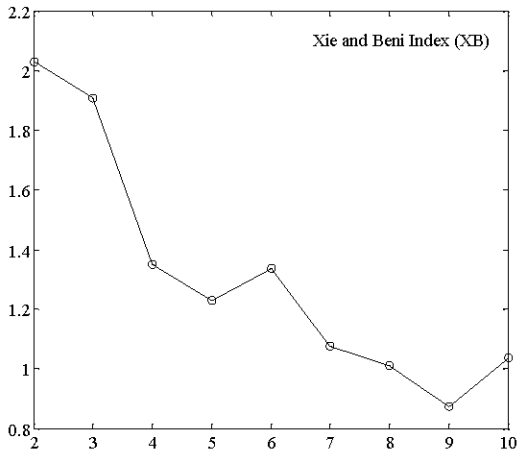
(b) V_{PE}



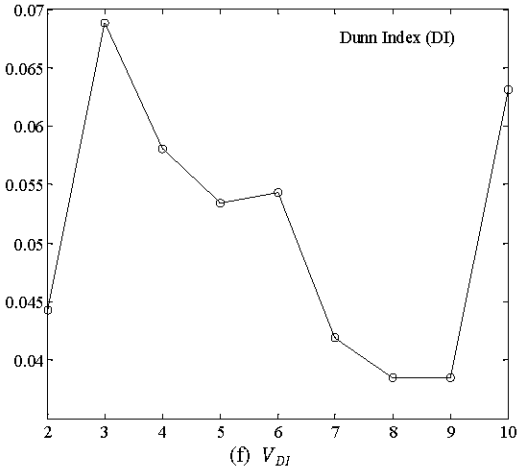
(c) V_{SC}



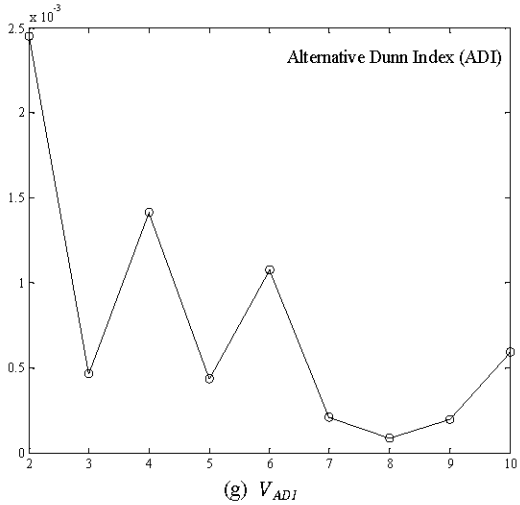
(d) V_S



(e) V_{XB}



(f) V_{DI}



(g) V_{ADI}

Fig. 5 Simulation results of clustering validity of Seeds data set.

TABLE 6. CLUSTERING VALIDITY INDEX VALUES UNDER DIFFERENT M VALUES

c	2	3	4	5	6	7	8	9	10
m=1.2									
V_{PC}	**0.9530**	0.9268	0.8920	0.8538	0.8417	0.8188	0.8021	0.8035	0.7897
V_{PE}	**0.0471**	0.0731	0.1079	0.1463	0.1584	0.1815	0.1980	0.1966	0.2106
V_S	**0.0016**	0.0030	0.0033	0.0035	0.0035	0.0041	0.0038	0.0040	0.0041
V_{SC}	**0.3463**	0.4002	0.4720	0.4927	0.5047	0.5490	0.5373	0.5345	0.5601
V_{XB}	2.5177	2.6800	**2.1788**	2.4357	2.8704	2.4828	2.6118	2.2067	2.9627
m=1.5									
V_{PC}	0.9000	**0.9074**	0.7733	0.6953	0.6725	0.6304	0.6010	0.6044	0.5807
V_{PE}	**0.0999**	0.1525	0.2266	0.3043	0.3273	0.3697	0.3991	0.3954	0.4191
V_S	0.0014	0.0023	0.0024	0.0025	0.0024	**0.0027**	0.0025	0.0025	0.0025
V_{SC}	**0.2938**	0.3125	0.3492	0.3525	0.3463	0.3676	0.3513	0.3369	0.3483
V_{XB}	2.2696	2.2761	1.7486	1.8047	2.0557	1.7253	1.7324	**1.4705**	1.8786
m=2									
V_{PC}	**0.8376**	0.7568	0.6387	0.5221	0.4914	0.4369	0.4020	0.4059	0.3806
V_{PE}	**0.2689**	0.4253	0.6226	0.8453	0.9229	1.0647	1.1742	1.1680	1.2725
V_S	0.0012	**0.0017**	0.0016	0.0016	0.0015	0.0016	0.0014	0.0014	0.0013
V_{SC}	0.2433	0.2318	0.2416	0.2304	0.2153	0.2193	0.2026	**0.1887**	0.1894
V_{XB}	2.0292	1.9084	1.3488	1.2276	1.3373	1.0755	1.0102	**0.8720**	1.0371
m=5									
V_{PC}	**0.6608**	0.5075	0.3145	0.1758	0.1453	0.1087	0.0917	0.0943	0.0859
V_{PE}	**0.3392**	0.4925	0.6856	0.8240	0.8547	0.8914	0.9073	0.9056	0.9141
V_S	0.7797	**0.9043**	0.6721	0.4766	0.3870	0.3513	0.2669	0.2643	0.2214
V_{SC}	0.1637	0.1148	0.0879	0.0606	0.0511	0.0424	0.0344	0.0329	**0.0282**
V_{XB}	1.4986	1.1572	0.5910	0.3265	0.2899	0.2022	0.1617	0.1496	**0.1487**

REFERENCES

[1] H. Huang, F. Z. Meng, S. H. Zhou, F. Jiang, and G. Manogaran, "Brain Image Segmentation Based on FCM Clustering Algorithm and Rough Set," *IEEE Access*, vol. 7, pp. 12386-12396, 2019.

[2] J. A. Hartigan, and M. A. Wong, "A K-means Clustering Algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100-108, 1979.

[3] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The Fuzzy C-means Clustering Algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[4] M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K. H. Phoe, "Fuzzy Clustering Method to Compare the Spread Rate of Covid-19 in the High Risks Countries," *Chaos, Solitons & Fractals*, vol. 140, pp. 110230, 2020.

[5] S. J. Cao, J. W. Ding and C. Ren, "Sensor Deployment Strategy Using Cluster Analysis of Fuzzy C-Means Algorithm: Towards Online Control of Indoor Environment's Safety and Health," *Sustainable Cities and Society*, vol. 59, pp. 102190, 2020.

[6] K. S. Zhang, P. J. Feng, G. Zhang, T. Jie, J. W. Hou, and X. Hou, "The Bi-Level Optimal Configuration Model of the CCHP System Based on the Improved FCM Clustering Algorithm," *Processes,* vol. 9, pp. 907, 2021.

[7] Y. X. Li, S. H. Li, S. Y. Peng, S. L. Zhao, W. X. Yan, and L. D. Qiu, "Extraction of Plateau Lake Water Bodies Based on an Improved FCM Algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 30, pp. 1727-1740, 2021.

[8] L. F. Zhu, J. S. Wang and H. Y. Wang, "A Novel Clustering Validity Function of FCM Clustering Algorithm," *IEEE Access*, vol. 7, pp. 152289-152315, 2019.

[9] M. Naderipour, M. H. F. Zarandi, and S. Bastani, "A Fuzzy Cluster-validity Index Based on the Topology Structure and Node Attribute in Complex Networks," *Expert Systems with Applications*, vol. 6, pp. 115913, 2021.

[10] H. Y. Wang, J. S. Wang and L. F. Zhu, "A New Validity Function of FCM Clustering Algorithm Based on the Intra-class Compactness and Inter-class Separation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 6, pp. 12411-12432, 2021.

[11] J. C. Bezdek, "Numerical Taxonomy with Fuzzy Sets," *Journal of Mathematical Biology*, vol. 1, no. 1, pp. 57-71, 1974.

[12] J. C. Bezdek, "Cluster Validity with Fuzzy Sets," *Journal of Cybernetics,* vol. 3, no. 3, pp. 58-73, 1973.

[13] X. L. Xie, and G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

[14] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity-guided Clustering with Applications to Image Segmentation," *IEEE Transactions on fuzzy systems*, vol. 4, no. 2, pp. 112-123, 1996.

[15] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters," *Journal Cybern.*, vol. 3, no. 3, pp. 32-57, 1973.