



Exponent and Logarithm Component-Wise Construction Method of FCM Clustering Validity Function Based on Subjective and Objective Weighting

Jia-Xu Liu¹ · Jie-Sheng Wang¹ · Guan Wang¹ · Xiao-Rui Zhao¹ · Hong-Yu Wang¹ · Di Jin²

Received: 31 May 2022 / Revised: 29 July 2022 / Accepted: 16 August 2022 / Published online: 26 September 2022
© The Author(s) under exclusive licence to Taiwan Fuzzy Systems Association 2022

Abstract The cluster validity function is used to evaluate the quality of the cluster results, and giving the exact number of initial cluster categories will rationalize the cluster results. Most single cluster validity functions and combined cluster validity functions generally have strong subjective problems, which also increases the burden on decision analysts and have great limitations in applications. To overcome the shortcomings of these clustering validity functions and improve the accuracy of the optimal cluster category classification for the datasets, based on the clustering performance evaluation components, a validity functional component construction method based on the exponential and log form was proposed. The weighting method adopts the combination of expert empowerment and standard separation method to combine the five weights so as to obtain 52 different fuzzy clustering validity functions. Then, based on the fuzzy C-mean (FCM) clustering algorithm, the performance analysis are carried out by using multiple data sets. Experimental simulation of these functions are proceeded on six commonly used UCI datasets. A clustering validity function with the simplest structure and the best classification effect was selected by comparison. Finally, this function is compared with 8 typical single clustering validity functions and four common clustering validity combination evaluation methods on 8 UCI data sets. Through experimental

simulation, the proposed validity function is compared in processing data sets, but also has strong scientific theoretical basis. Thus, the feasibility and effectiveness of the proposed clustering validity function construction method are proved.

Keywords Fuzzy c-means clustering algorithm · Clustering validity function · Subjective and objective weighting · Component-wise design

1 Introduction

Clustering is accompanied by the emergence and development of human society. People in the process of understanding and mastering objective things, always distinguish different things and recognize the similarity between things. Therefore, the research of cluster analysis is not only of great theoretical significance, but also has important engineering application and humanistic values. With its theory development, clustering has been widely used in many fields, such as speech recognition, face recognition, radar target recognition, biological information analysis [1, 2], image segmentation [3–6], edge detection [7], image compression [8], curve fitting, target detection and tracking, mobile robot positioning, traffic flow video detection [9, 10], and model identification and fuzzy rule establishment [11, 12]. Clustering learning is one of the earliest methods used in pattern recognition and data mining tasks, and is used to study large databases in various applications. Therefore, the clustering algorithm for big data has attracted more and more attention.

In recent years, with the development of computing theory and technology, many clustering methods have been proposed. According to the implementation ideas of

✉ Jie-Sheng Wang
wang_jiesheng@126.com

¹ School of Electronic and Information Engineering,
University of Science & Technology Liaoning,
Anshan 114051, China

² Inspection and Certification Center Liaoning,
Anshan 114051, China

clustering algorithms, they can be divided into hierarchical clustering algorithm, partitioned clustering algorithm, density based clustering algorithm, grid based clustering algorithm, and model-based clustering algorithm. Goldberger et al. Proposed a hierarchical clustering algorithm based on classical Hungarian method [13]. Typical clustering methods include k-means algorithm, k-medoids algorithm, and fuzzy c-means algorithm [14]. Macqueen proposed K-means clustering algorithm [15] in 1967, which has become one of the most classic clustering algorithms. Yodern et al. proposed a semi-supervised K-means clustering algorithm in 2017 [16]. Gengzhang et al. proposed a DC k-means algorithm in 2018 [17]. Hiep proposed a differential privacy preserving K-Modes algorithm in 2018 [18]. So far, many clustering algorithms have been put forward for different types of clustering, which can meet the different clustering requirements. However, many existing clustering algorithms need to specify the number of clusters in order to obtain the optimal clustering partition of the target datasets before performing the clustering task. The clustering validity index is used to evaluate the partition result of clustering algorithm, which are modeled by mathematical knowledge and can evaluate the effectiveness of clustering partition results. Through the mathematical evaluation of the clustering results for datasets, the clustering algorithm can also obtain the best clustering results under the premise of unable to achieve the given optimal number of clusters. From the present point of view, the research on clustering validity can be roughly divided into the study of single cluster validity function and the research of combined clustering validity evaluation method. The research on single clustering validity function focuses on the following two aspects.

- (1) The fuzzy clustering validity function based on membership degree. Partition coefficient (V_{PC}) defined by Bezdek is used to measure the overlap between clusters [19]. Bezdek will also proposed partition entropy (V_{PE}) used to measure the fuzziness of clustering partition [20]. This index is similar to V_{PC} . Bezdek proved that for all probabilistic cluster partitions, the structure of V_{PC} and V_{PE} is simple and the amount of calculation is small, but they will change monotonously with the number of clusters. An improved partition coefficient (V_{MPC}) is revised on V_{PC} about the existed monotone decreasing trend problem, but other aspects of the defects have not been improved [21]. In 2004, Chen and links proposed an effective index in the form of subtraction (V_P), which is an effective function that only focuses on membership [22]. In 2013, Jiashun proposed a clustering validity function (V_{CS}), which can effectively suppress noise data [23]. Joopudi

used the maximum membership degree and the minimum membership degree to measure the data overlap and proposed a clustering validity function (V_{GD}) [24].

- (2) The fuzzy clustering validity function based on geometric structure. Xie and Beni proposed a clustering validity function (V_{XB}) based on proportion operation in 1991 [25], which is the first clustering validity function which takes into account the structure of the data set. It is the ratio of the compactness within the cluster and the separation between the clusters. V_K is an validity index proposed by Kwon. The method of adding penalty items to the numerator of the index effectively restrained the trend of decreasing monotonously of V_{XB} . V_{PACES} index is a clustering validity index proposed by Wu and Yang in 2005. It describes the compactness and separation of clustering by fuzzy membership function and the relative value of the center distance of an exponential type structure [26]. Chi-Hung Wu proposed a clustering validity function (V_{WL}) in 2015 [27]. It considers all clusters and the overall compactness separation ratio of each cluster. Chi Yun proposed a validity function (V_{FM}) in 2007 [28], which takes the partition entropy and fuzzy partition factor into account, and defines the compactness and separation of clustering, but its performance on noisy data sets is poor. Zhu proposed a new clustering validity function (V_{ZLF}) in 2019 [29], which can divide high-dimensional data sets accurately. In 2021, Wang used the definition of compactness, separation, and overlap as reference, introduced a new concept to enhance the adaptability of the validity function, and thus proposed a clustering validity function (V_{HY}) [30]. Wang proposed a new clustering validity function (V_{WG}) in 2021, which can find the best clustering number of the noise, overlapping, and high-dimensional datasets [31].

The final clustering results will be directly affected by the performance of the clustering validity functions. Aiming at the shortcomings of the existing fuzzy clustering validity functions, this paper proposes a new fuzzy c-means clustering validity function based on multi clustering performance evaluation components by combining the subjective weighting method and standard deviation method. Two kinds of combination weighting methods in exponential and logarithmic forms are proposed, and then five FCM clustering performance evaluation components are continuously arranged and combined in weighted form. Several clustering validity functions based on combination weighting strategy are tested on UCI datasets. The

experimental results show that the two validity functions can obtain the correct clustering results on UCI datasets, which can overcome the defects of other clustering validity functions, and become a new direction to solve the problem of fuzzy clustering validity problem, and expand the theoretical system of constructing clustering validity functions based on components.

2 FCM Clustering Algorithm and Combined Clustering Validity Evaluation Method

2.1 FCM Clustering Algorithm

Fuzzy C-means (FCM) algorithm is a common soft clustering algorithm, and it is also the most representative fuzzy clustering algorithm. It is widely used in pattern recognition and clustering analysis. Set the target data set $X = \{x_1, x_2, \dots, x_n\}$ composed of n samples, the sample data $x_j = [x_{1j}, x_{2j}, \dots, x_{sj}]^T$, where x_{kj} is the k property value of x_j . For a given sample set X , the cluster analysis of X is divided into the c clusters. The minimum objective function is found by iteration, which is defined in Eq. (1).

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

where $J_{FCM}(U, V)$ represents the square error clustering criterion, and its minimum value is called as the stationary point of least square error. $V = \{v_1, v_2, \dots, v_n\}$ represents the set of clustering centers, whose definition is shown in Eq. (2).

$$v_i = \frac{\sum_{j=1}^n \sum_{k=1}^c u_{kj}^m \cdot x_k}{\sum_{j=1}^n u_{ij}^m} \tag{2}$$

where, c represents the number of clusters; $m \in (1, \infty)$ is the fuzzy coefficient to control the fuzziness of membership degree of each group data the range; v_i is on behalf of the i -th clustering centers; $\|x_j - v_i\|$ represents the distance between the objects x_j to the cluster center v_i , which usually adopts the euclidean distance; $u_{ij} (0 \leq u_{ij} \leq 1)$ represents the membership degree of the data objects x_j belonging to the cluster center v_i ; $u_{ij} \in U$ and U is the membership matrix of fuzzy partition and meet the following conditions.

$$u_{ij} = \left[\sum_{k=1}^c \left(\frac{\|x_j - v_i\|^2}{\|x_j - v_k\|^2} \right)^{2m-1} \right]^{-1}, \tag{3}$$

$$\sum_{i=1}^c u_{ij} = 1, 0 \leq \sum_{j=1}^n u_{ij} \leq n$$

where, $1 \leq j \leq n, 1 \leq i \leq c$.

The FCM clustering algorithm process is described as follows:

Step 1: Set the clustering parameter c , fuzzy factor m , and convergence threshold ε .

Step 2: Initialize the clustering center matrix V and membership matrix U , and obtain U_0 and V_0 .

Step 3: Update the fuzzy partition matrix $U = (u_{ij})_{c \times n}$ according to Eq. (3).

Step 4: Update the clustering center $V = \{v_1, v_2, \dots, v_c\}$ according to Eq. (2).

Step 5: Calculate $e = \|u_{t+1} - u_t\|$. If $e \leq \varepsilon$ (ε is a threshold from 0.001 to 0.01), the algorithm stops and the final clustering result is calculated. Otherwise, $U_t = U_{t+1}$ and repeat from Step 2.

2.2 Combined Clustering Validity Evaluation Method

The clustering validity problem mainly lies in how to select a clustering validity function to determine the optimal number of clusters in datasets. The clustering validity functions can be roughly divided into external validity function, internal validity function, and relative validity function. Both the internal fuzzy clustering validity function and the relative fuzzy clustering validity function have developed very mature and the system is more and more perfect. At present, the fusion of clustering validity functions mostly uses the weighted combination. The typical weighted combination clustering validity evaluation methods are listed in Table 1.

3 Exponent and Logarithm Component-wise Construction Method of FCM Clustering Validity Function

3.1 Clustering Validity Evaluation Components

Based on the characteristics of FCM clustering algorithm and typical clustering validity functions, five clustering validity evaluation components (CP) are defined in this paper. These components are used to represent the compactness, similarity, variability within the datasets, and the degree of separation and overlap between datasets, which are shown in Table 2.

3.2 Exponent and Logarithm Component-Wise Construction Method

In order to explore the combinatorial weighting component construction method of clustering validity evaluation, the above five components are normalized and standardized to

Table 1 Weighted combination clustering validity evaluation methods

Proposer	Definition	Weighting method	Selection of validity functions
Sheng [32]	$\max WSVF = \sum_{i=1}^m w_i f_i(x)$	Average weighting	$1/DB, V_{SIL}, V_D, V_{33}, V_{CH}, V_{PBM}$
Hong-bin Dong [33]	$\max FWSVF = \sum_{i=1}^r w_i f_i(x)$	Average weighting	$1/V_{XB}, 1/V_{PE}, V_{PC}, V_{PBMF}$
Wu [34]	$\min DWSVF = \sum_{i=1}^m w_i f_i(x)$	Dynamic weighting	$1/V_{MPC}, V_{PE}, V_{XB}, V_{PBMF}$
Wang [35]	$\min HWCVF = \sum_{i=1}^m w_i f_i(x)$	Mixed weighting	$V_{MPC}, V_{XB}, V_K, V_P, V_{PBMF}, V_{WL}, V_{ZLF}, V_{HY}$

Table 2 Clustering validity evaluation components

Component name	Define formula	Features of components
CP_1 (Compactness)	$CP_1 = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^2 \ x_j - v_i\ ^2$	The smaller the value of CP_1 is, the higher the similarity degree of data in the class is, and the closer the data in the class is
CP_2 (Variability)	$CP_2 = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n \exp\left(-\frac{\ x_j - v_i\ ^2}{\varepsilon}\right)$	The smaller the value of CP_2 is, the lower the variability of the data, and the more stable the intra class data. ε represents the average sum of distances between the data in the class and the mean value of the data sample. The smaller the value, the more stable the data
CP_3 (Overlap)	$CP_3 = \min_{i \neq j} \left(\frac{1}{n} \sum_{j=1}^n (1 - u_{ik} - u_{jk}) \right)$	CP_3 is proposed based on the membership degree to represent the degree of overlap between data sets. The closer the difference $ u_{ik} - u_{jk} $ will be to 1, the smaller the $1 - u_{ik} - u_{jk} $ value. Therefore CP_3 with a small value indicates that the classification effect is good
CP_4 (Similarity)	$CP_4 = \min_{1 \leq i \leq c} \sum_{j=1}^n (u_{ij})^2$	CP_4 represents the sum of the squares of membership of data points x_j belonging to the cluster center v_i . The higher the value of CP_4 , the more similar the data within the class
CP_5 (Separation)	$CP_5 = \min_{i \neq k} \ v_i - v_k\ ^2$	$CP_5 = \min_{i \neq k} \ v_i - v_k\ ^2$ is used to define the components of overlapping degree between classes, which represents the minimum distance between any two clusters. The higher the value CP_5 , the better the separation between classes

make them in the same dimension range. Then, these components are arranged and combined in the form of combination weight. The new exponential and logarithmic clustering validity functions are constructed. Their construction rules are shown in Eq. (4) and (5).

$$\min V = \sum_{i=1}^m w_{hybrid} e^{CP_i} \tag{4}$$

$$\min S = \sum_{j=1}^n w_{hybrid} \ln(1 + CP_j) \tag{5}$$

where, $m, i, n,$ and j are random positive integers from 1 to 5.

This paper proposed a clustering validity function constructed in the form of logarithm and exponent. The exponential function $y = e^x$ and the logarithm function $y = \ln x$ are monotonically increasing functions. In the interval

$0 + \infty$, the ordinate value of the exponential function is always greater than 0, while the value of the logarithmic function is less than 0 in the interval, and it needs to be greater than 0 in the constructor. Therefore, this problem can be solved by expressing it with the logarithmic form $y = \ln(1 + x)$. Then V and S are applied to the FCM clustering algorithm, and the algorithm flowchart for obtaining the optimal number of clusters is shown in Fig. 1. The flowchart of the FCM clustering algorithm based on V and S are described as follows.

The design method draws lessons from the linear superposition idea and the characteristics of exponential and logarithmic functions based on the previous combined clustering validity methods. However, the components selected in this paper have the problem that they are effective under different extreme conditions. Therefore, the components with the maximum effective value in Eq. (4) is

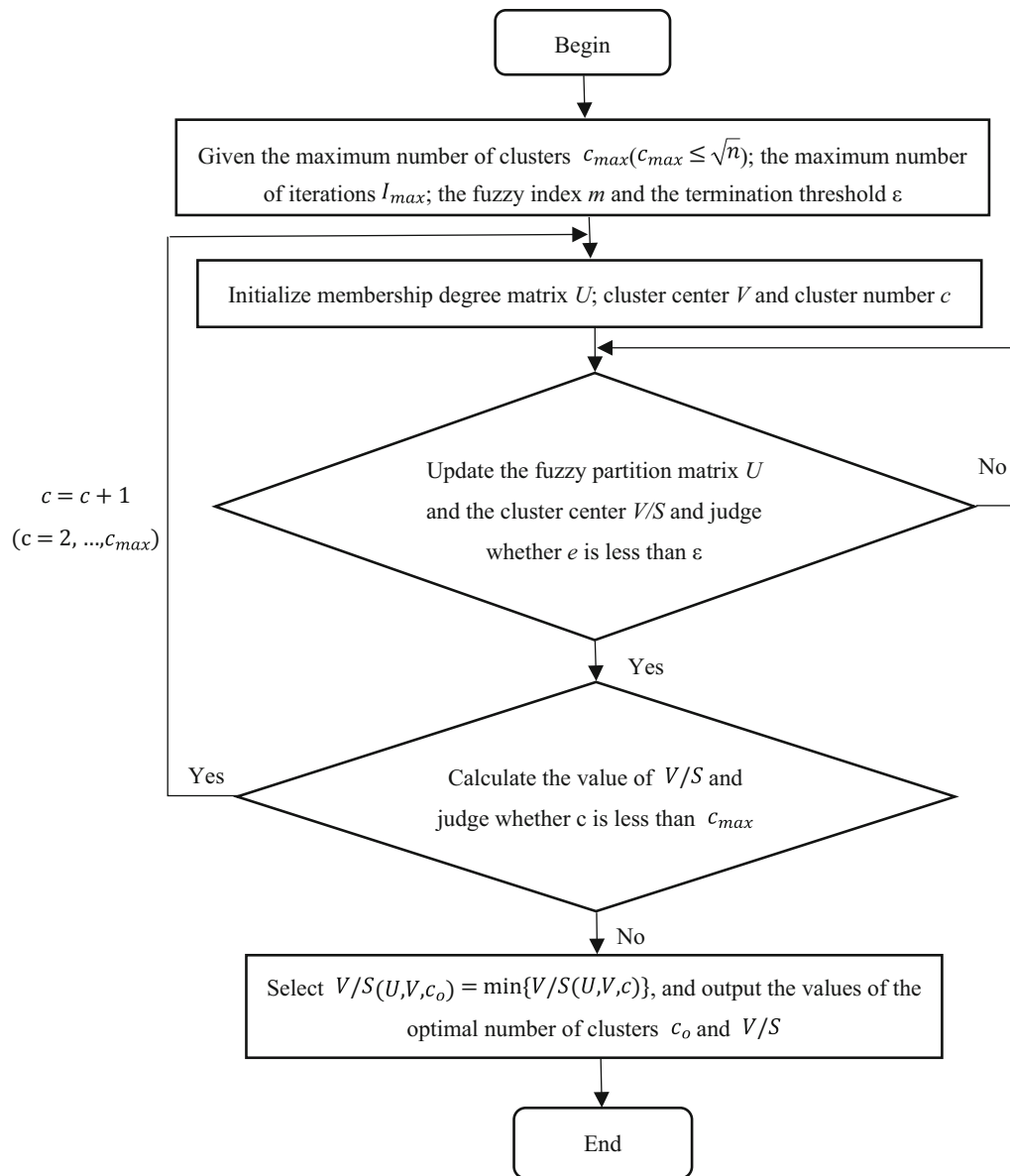


Fig. 1 Flowchart of FCM clustering algorithm based on the proposed validity function

used as $(1 - CP_i)$. In Eq. (5), the component is dealt with $(1 + CP_j)$ so as to obtain $CP_j = \{(1 + CP_1), (1 + CP_2), (1 + CP_3), (1 + CP_4), (1 + CP_5), (1 + CP_6)\}$. In this way, the format of each validity function can be guaranteed to be unified. In addition, for different components, the amount of information is different, but a component corresponds to a weight, which is independent of the position of the component in the clustering validity function. So w_{hybrid} based on five components can be expressed as $w_{hybrid} = \{w_1, w_2, w_3, w_4, w_5, w_6\}$. In this way, the clustering validity function is constructed as V and S . It is more convenient to carry out the simulation and contrast tests in the following paper by using similar structure mode and all the minimum values are effective.

3.3 Subjective and Objective Weighting Strategy

The selection of weight is very important in the combination evaluation process of clustering validity function. Different weight will bring different influence. The weight types can be mainly divided into the following two types.

- (1) Subjective weighting method. It is a method to determine the attribute weight according to the subjective attention of decision makers (experts). Common methods include Delphi method, analytic hierarchy process (AHP), fuzzy analysis method, linked ratio method, correlation tree method, set-valued iteration method, and eigenvalue method. The subjective weighting method is that the experts

reasonably determine the weight of each attribute according to the actual decision-making problems and their own knowledge and experience. The determination of the weight is generally in line with the reality, so it is highly explanatory. But the decision-making and evaluation results have strong subjectivity and randomness, so its objectivity is weak and has great limitations in application.

- (2) Objective weighting method. It is mainly based on the degree of connection between indicators, the amount of information provided by each index, and the impact on other indicators. Therefore, the obtained weight is observable and does not increase the burden of decision makers. This method has strong mathematical theoretical basis. Common objective weighting methods include the principal component analysis method, multi-objective programming method, entropy weight method, CRITIC method, and standard deviation method.

In this paper, the hybrid weighting method of subjective and objective combination is used to eliminate subjective deviation and objective one-sided, and show the subjective and objective information while determining the weight so that it can truly and objectively reflect the actual situation of single cluster validity function. The hybrid weighting method can be defined as follows:

$$w_{\text{hybrid}} = \delta w_{\text{object}} + (1 - \delta)w_{\text{subject}} \quad (6)$$

where, w_{object} is the subjective empowerment, w_{subject} is the objective empowerment, w_{hybrid} is the hybrid empowerment, and δ is the adjustment coefficient ($\delta \in [0, 1][0, 1]$).

When $\delta = 0$, $w_{\text{hybrid}} = w_{\text{object}}$ becomes the subjective weight. When $\delta = 1$, $w_{\text{hybrid}} = w_{\text{subject}}$ and the hybrid weight becomes the objective weight. When $\delta = 0.5$, we can conclude the objective index w_{object} and the subjective index w_{subject} have the same influence on w_{hybrid} . δ may be revised according to the importance of the index and the characteristics of the data set so as to improve the classification accuracy. The value of w_{object} is determined by the decision maker and is taken as $w_{\text{object}} = 1/n$ without special instructions, where n is the number of clustering validity functions. w_{subject} is determined by information entropy weighting.

The standard deviation method is used to determine the index weight. First, the standard deviation of the index is calculated, and then the weight is determined based on the mean square deviation of the index. In the multi-index comprehensive evaluation, because the dimension of the index or the quantity of the datasets may be different, there is no comparability between the indicators. So the first step is to process the data and make the indexes comparable. There are many ways to deal with it, such as normalization, standardization, and some other dimensionless methods.

For the convenience of the following description, it is advisable to set the index set as $G = \{G_1 G_2 \dots G_m\}$, the sample set (or scheme set) is $A = \{A_1 A_2 \dots A_n\}$, and the corresponding sample point is X_{ij} ($i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$). The weight vector of evaluation index is $W = (w_1 w_2 \dots w_m)^T$, which satisfies $\sum w_i = 1$. After dimensionless or standardized treatment, the matrix $X = (X_{ij})$ is changed to the matrix $Z = (Z_{ij})$. The standard deviation determines the weight of the index. The principle is that if the standard deviation of an index is smaller, the variation degree of the index is smaller, the amount of information provided is smaller, and the role played in the comprehensive evaluation is smaller, so the weight of the index is smaller. On the contrary, the greater the weight. The characteristic of standard deviation weighting method is that the index weight reflects the amount of data information or the change of index data.

The specific calculation steps of subjective and objective weighting are described as follows.

Step 1: The original data matrix X_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$) is carried out the dimensionless treatment. The extreme method is generally used for dimensionless treatment. The positive indices are $Z_{ij} = (X_{ij} - \min_j\{X_{ij}\}) / (\max_j\{X_{ij}\} - \min_j\{X_{ij}\})$. For the inverse index $Z_{ij} = (\max_j\{X_{ij}\} - X_{ij}) / (\max_j\{X_{ij}\} - \min_j\{X_{ij}\})$, obtain the matrix Z'_{ij} .

Step 2: Calculate the mean value of random variables by $\bar{Z}_j = \frac{1}{n} \sum_{i=1}^n Z_{ij}$;

Step 3: Calculate the mean square error of index j by $\sigma_j = \sqrt{\sum_{i=1}^n (Z_{ij} - \bar{Z}_j)^2}$;

Step 4: Calculate the weight of index j by $w_j = \sigma_j / \sum_{j=1}^p \sigma_j$;

Step 5: According to the multi-index weighted comprehensive evaluation model $D_i = \sum_{j=1}^m W_j \cdot Z_{ij}$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$), calculate the comprehensive evaluation value, where W_j is the weight of the j the index.

4 Simulation Experiment and Result Analysis

According to the composition rules shown in Eq. (4) and Eq. (5), 52 different clustering validity functions can be formed by permutation and combination of five components in the form of exponent and logarithm. In order to facilitate the experiment, these 52 validity functions were divided into six groups (eight in four groups and nine in two groups), which are listed in Tables 3, 4, 5, 6, 7, and 8, respectively. These tables show the names and the simplified function form of the 52 validity functions. According to the prior knowledge, the fuzzy index can be determined as $1.5 \leq m \leq 2.5$ and the number of clusters is

selected as $2 \leq c \leq \sqrt{n}$. This paper chooses $m = 2$ and $2 \leq c \leq 14$. Then whether the classification by using each clustering validity function is accurate is judged on different data sets. In this paper, we select the UCI datasets to carry out simulation experiments, such as Iris, Seeds, Balance, Hfcr, Glass, and Cooking. The amount of data, categories, and attributes of the adopted UCI datasets are listed in Table 9. In order to better observe the changing trend of six groups of clustering validity functions in the

face of UCI datasets, the function values of their experimental results are placed in the normalized coordinate system, as shown in Figs. 1, 2, 3, 4, 5, and 6. In this way, the clustering effect of each validity function can be compared more intuitively. Finally, for different UCI datasets, the optimal number of clusters for each cluster validity function is listed. The results are listed in Tables 10, 11, 12, 13, 14, and 15.

Table 3 Exponent clustering validity functions (Group 1)

Function name	Shorthand function description
V_1	$w_1e^{CP_1} + w_2e^{CP_2}$
V_2	$w_1e^{CP_1} + w_3e^{CP_3}$
V_3	$w_1e^{CP_1} + w_4e^{(1-CP_4)}$
V_4	$w_1e^{CP_1} + w_5e^{(1-CP_5)}$
V_5	$w_2e^{CP_2} + w_3e^{CP_3}$
V_6	$w_2e^{CP_2} + w_4e^{1-CP_4}$
V_7	$w_2e^{CP_2} + w_5e^{1-CP_5}$
V_8	$w_3e^{CP_3} + w_4e^{1-CP_4}$
V_9	$w_3e^{CP_3} + w_5e^{1-CP_5}$

Table 4 Exponent clustering validity functions (Group 2)

Function name	Shorthand function description
V_{10}	$w_4e^{1-CP_4} + w_5e^{1-CP_5}$
V_{11}	$w_1e^{CP_1} + w_2e^{CP_2} + w_3e^{CP_3}$
V_{12}	$w_2e^{CP_2} + w_3e^{CP_3} + w_5e^{(1-CP_5)}$
V_{13}	$w_1e^{CP_1} + w_2e^{CP_2} + w_5e^{(1-CP_5)}$
V_{14}	$w_1e^{CP_1} + w_3e^{CP_3} + w_4e^{(1-CP_4)}$
V_{15}	$w_1e^{CP_1} + w_3e^{CP_3} + w_5e^{(1-CP_5)}$
V_{16}	$w_1e^{CP_1} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{17}	$w_2e^{CP_2} + w_3e^{CP_3} + w_4e^{(1-CP_4)}$
V_{18}	$w_2e^{CP_2} + w_3e^{CP_3} + w_5e^{(1-CP_5)}$

Table 5 Exponent clustering validity functions (Group 3)

Function name	Shorthand function description
V_{19}	$w_2e^{CP_2} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{20}	$w_3e^{CP_3} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{21}	$w_1e^{CP_1} + w_2e^{CP_2} + w_3e^{CP_3} + w_4e^{(1-CP_4)}$
V_{22}	$w_1e^{CP_1} + w_2e^{CP_2} + w_3e^{CP_3} + w_5e^{(1-CP_5)}$
V_{23}	$w_1e^{CP_1} + w_2e^{CP_2} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{24}	$w_1e^{CP_1} + w_3e^{CP_3} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{25}	$w_2e^{CP_2} + w_3e^{CP_3} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$
V_{26}	$w_1e^{CP_1} + w_2e^{CP_2} + w_3e^{CP_3} + w_4e^{(1-CP_4)} + w_5e^{(1-CP_5)}$

Table 6 Logarithmic clustering validity functions (Group 1)

Function name	Shorthand function description
S_1	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2)$
S_2	$w_1 \ln(1 + cp_1) + w_3 \ln(1 + cp_3)$
S_3	$w_1 \ln(1 + cp_1) + w_4 / \ln(1 + cp_4)$
S_4	$w_1 \ln(1 + cp_1) + w_5 / \ln(1 + cp_5)$
S_5	$w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3)$
S_6	$w_2 \ln(1 + cp_2) + w_4 / \ln(1 + cp_4)$
S_7	$w_2 \ln(1 + cp_2) + w_5 / \ln(1 + cp_5)$
S_8	$w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4)$
S_9	$w_3 \ln(1 + cp_3) + w_5 / \ln(1 + cp_5)$

Table 7 Logarithmic clustering validity functions (Group 2)

Function name	Shorthand function description
S_{10}	$w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{11}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3)$
S_{12}	$w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_5 / \ln(1 + cp_5)$
S_{13}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_5 / \ln(1 + cp_5)$
S_{14}	$w_1 \ln(1 + cp_1) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4)$
S_{15}	$w_1 \ln(1 + cp_1) + w_3 \ln(1 + cp_3) + w_5 / \ln(1 + cp_5)$
S_{16}	$w_1 \ln(1 + cp_1) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{17}	$w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4)$
S_{18}	$w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_5 / \ln(1 + cp_5)$

Table 8 Logarithmic clustering validity functions (Group 3)

Function name	Shorthand function description
S_{19}	$w_2 \ln(1 + cp_2) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{20}	$w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{21}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4)$
S_{22}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_5 / \ln(1 + cp_5)$
S_{23}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{24}	$w_1 \ln(1 + cp_1) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{25}	$w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$
S_{26}	$w_1 \ln(1 + cp_1) + w_2 \ln(1 + cp_2) + w_3 \ln(1 + cp_3) + w_4 / \ln(1 + cp_4) + w_5 / \ln(1 + cp_5)$

Table 9 UCI data sets (component experiments)

Data sets	Data numbers	Attributes	Classes
Iris	150	4	3
Seeds	210	7	3
Balance	635	4	3
Hfcr	299	13	4
Glass	214	9	6
Cooking	4900	3	7

5 Simulation Results and Analysis on Exponent Validity Functions (Group 1)

As can be seen from Fig. 2a–f, the six clustering validity functions of the first group of indices are defined as $V_1, V_5, V_6, V_7, V_8,$ and V_9 , which cannot classify any group of UCI datasets. Each group of data in UCI data set can be classified by V_2 . It can be seen from Fig. 2a–c that Iris, Seeds and Balance data sets can be correctly divided into three categories by V_3 , and the other data sets are classified incorrectly. Finally, it can be seen from Fig. 2b and d, V_3 only can the optimal number of clusters for the Seeds and Hfcr datasets. It can be found in the comparison experiment of the first group of exponent clustering validity functions. Three UCI datasets can be distinguished correctly by V_3 ; Only two sets of real datasets can be correctly divided by V_4 ; $V_1, V_5, V_6, V_7, V_8,$ and V_9 unable to accurately cluster any set of UCI datasets. The validity function V_2 has the best clustering performance, which can be achieved by dividing all UCI datasets.

6 Simulation Results and Analysis on Exponent Validity Functions (Group 2)

From the experimental results in Fig. 3a–d, it can be concluded that when processing Glass and Cooking datasets, the second group of clustering validity function cannot correctly classify them. It can be seen from Fig. 3a–b that the samples

in Iris and Seeds datasets can be accurately divided into three categories by $V_{11}, V_{12},$ and V_{14} . It can be found from Fig. 3b and d that Seeds and Hfcr datasets can be effectively classified by V_{13} . As shown in Fig. 3c, Balance can be accurately divided into three categories by V_{12} . Through the comparative experiments on the second group of exponential validity functions, it can be found that, $V_{10}, V_{15}, V_{16}, V_{17},$ and V_{18} are unable to successfully distinguish any of the selected UCI datasets. Two UCI datasets can be distinguished successfully by $V_{11}, V_{13},$ and V_{14} . The best clustering number of three UCI datasets can be obtained by V_{12} .

7 Simulation Results and Analysis on Exponent Validity Functions (Group 3)

In Fig. 4d–f, it can be found that none of the eight validity functions can classify Hfcr, Glass, and Cooking data sets. As can be seen from Fig. 4c, in addition to the validity function V_{21} , the other seven validity functions can classify the Balance data set into three categories and the classification performance is excellent. Finally, according to Fig. 4a and b, Iris and Seeds datasets can be correctly divided into three categories by V_{21} , while other seven validity functions cannot effectively classify these two datasets. Based on the comparative experiment results by adopting the third group of exponential validity functions, we can draw the conclusion that V_{21} can successfully separate two datasets, while the other seven validity functions cannot classify the other datasets except the Balance data set. The classification performance of the eight clustering validity functions is relatively poor and is excluded from the selection range.

8 Simulation Results and Analysis on Logarithm Validity Functions (Group 1)

According to Fig. 5a–f, the six datasets in UCI datasets can be classified successfully by S_2 . Seen from Fig. 5c, Balance can be divided into three categories by S_5 , and the optimal

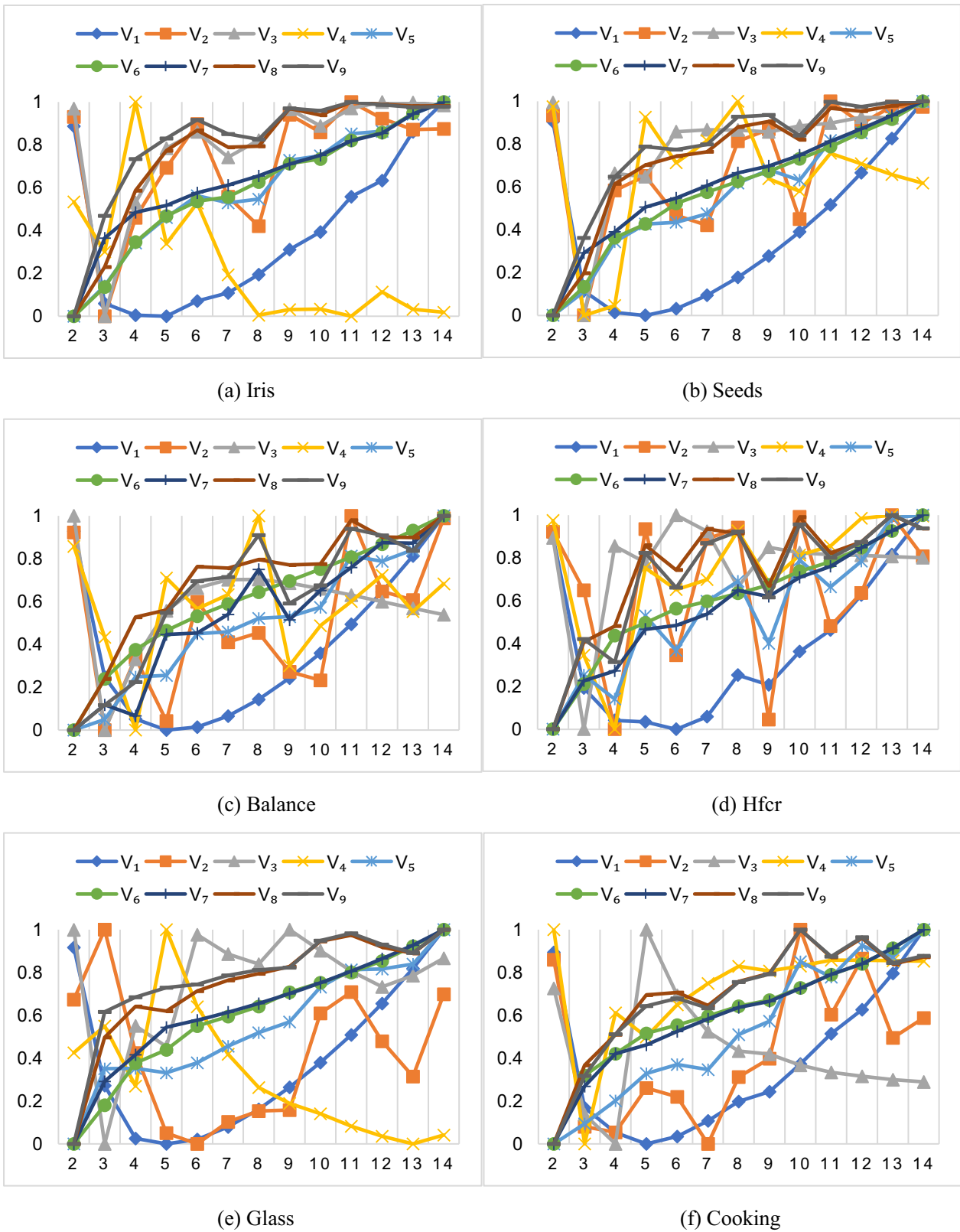


Fig. 2 Variation trend of normalized exponential clustering validity functions (Group 1)

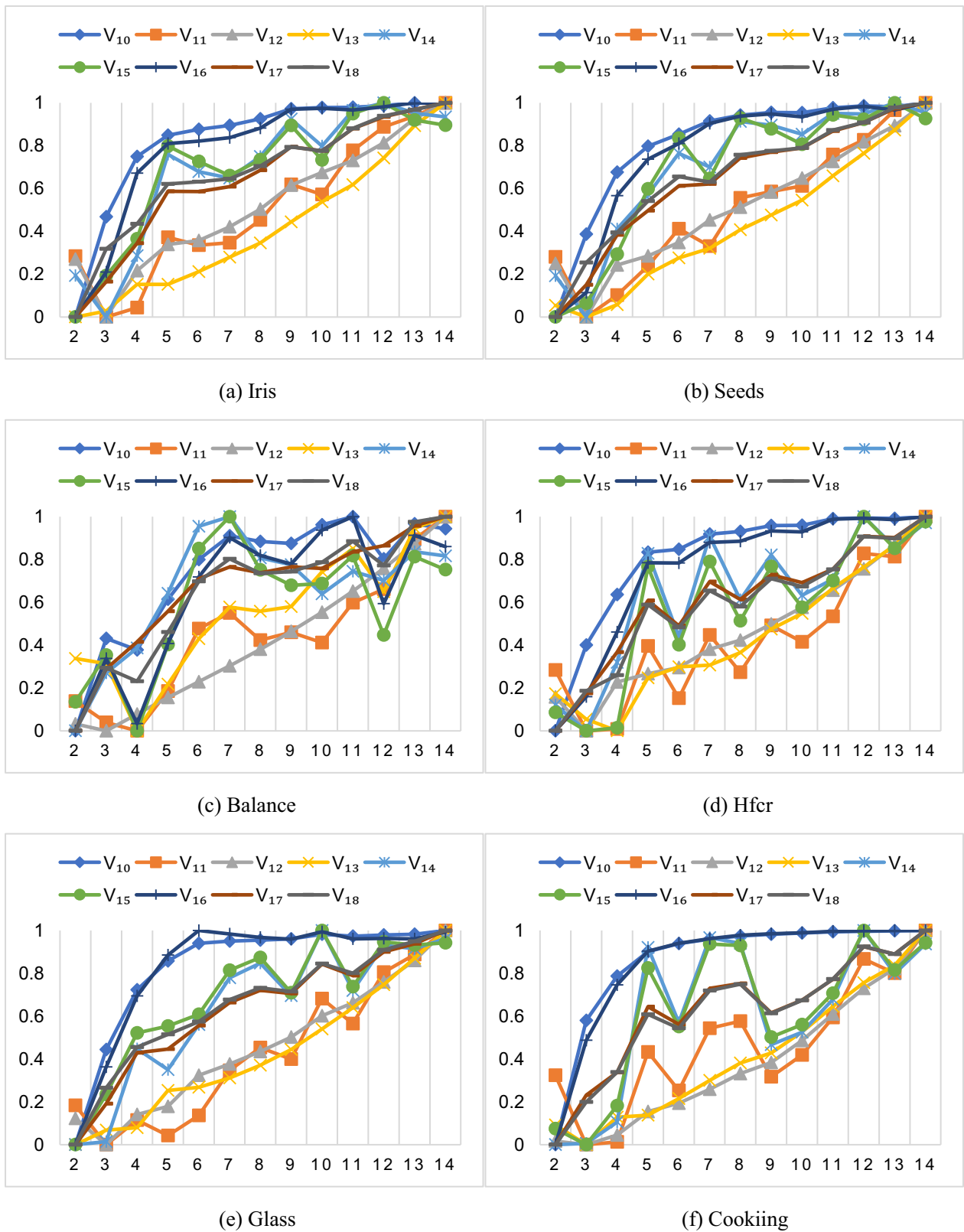


Fig. 3 Variation trend of normalized exponential clustering validity functions (Group 2)

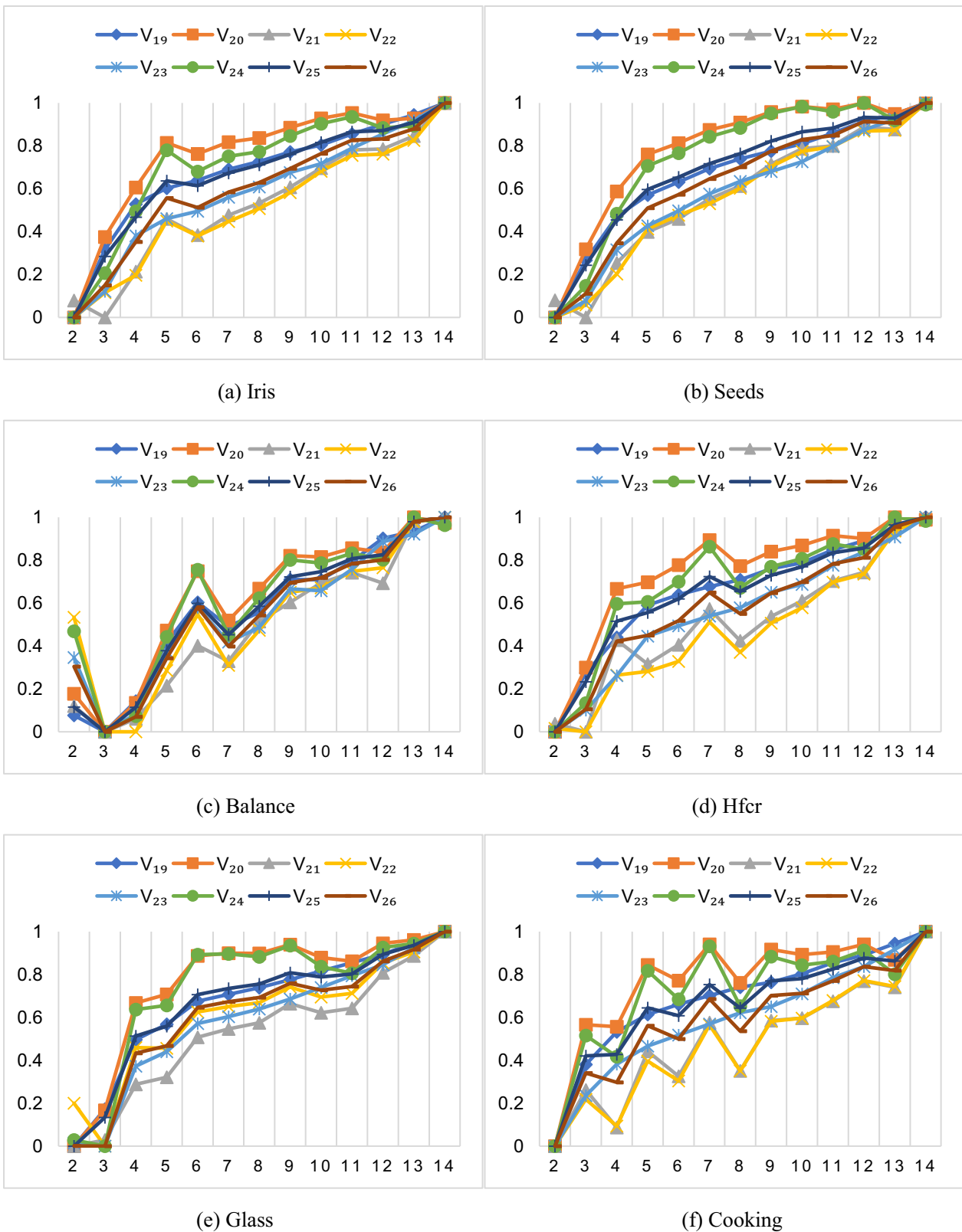


Fig. 4 Variation trend of normalized exponential clustering validity functions (Group 3)

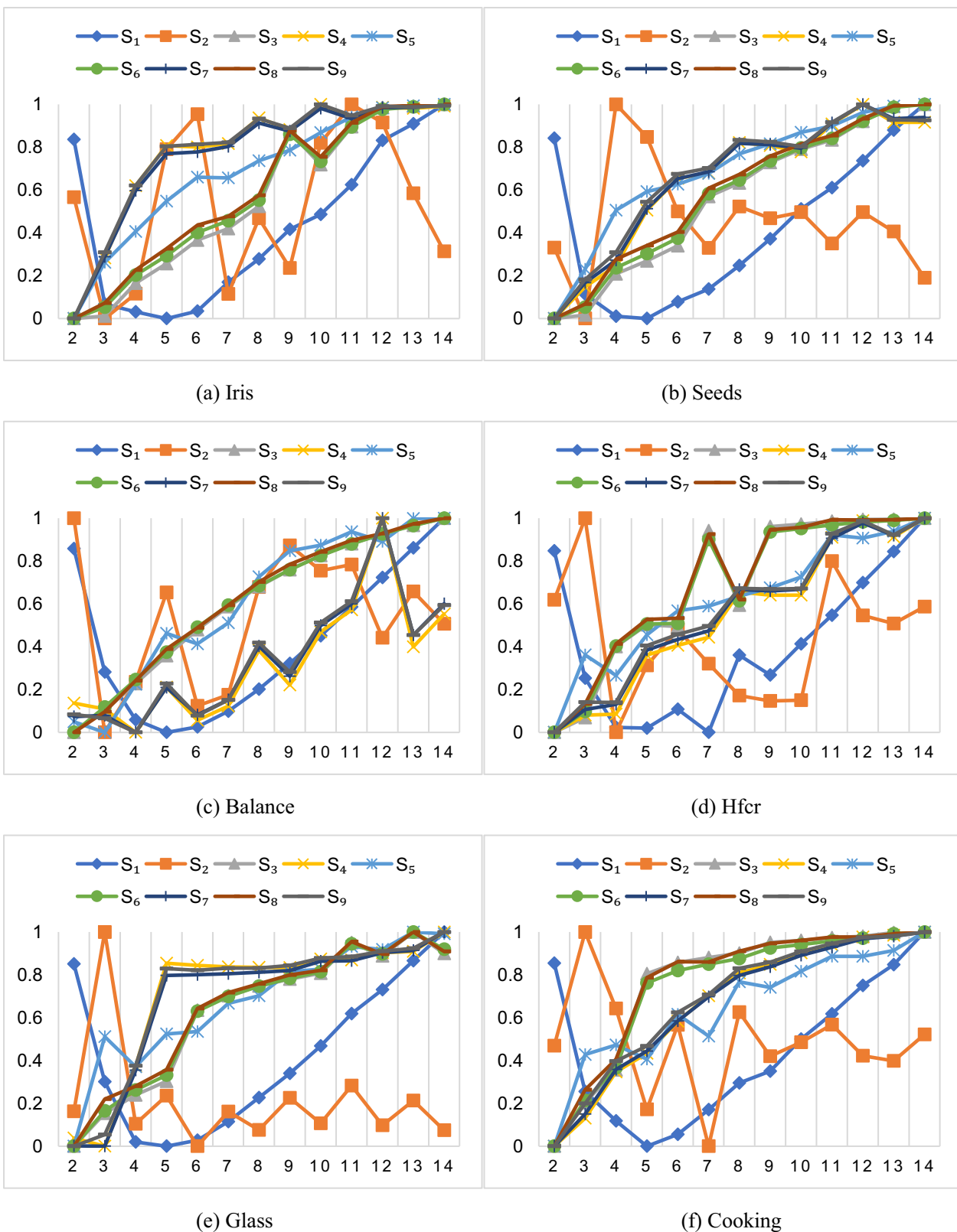


Fig. 5 Variation trend of normalized logarithmic clustering validity functions (Group 1)

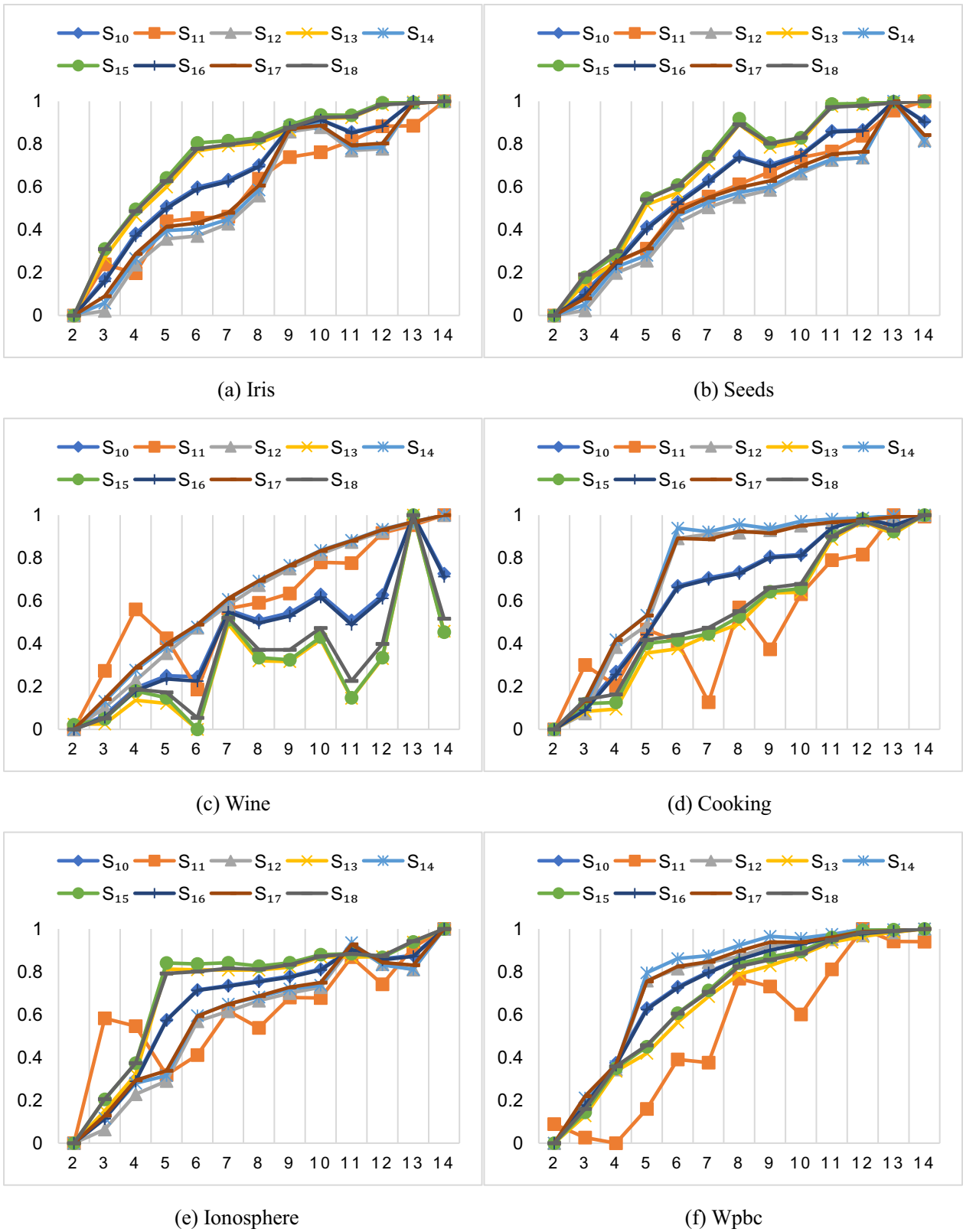


Fig. 6 Variation trend of normalized logarithmic clustering validity functions (Group 2)

cluster number of Balance is $c = 3$. However, the rest of the log validity functions in the first group cannot distinguish any selected UCI datasets correctly. The experimental results are similar to those of the first exponential group. According to the contrast experiment results on the first group of logarithmic validity functions, it can be concluded that S_5 can obtain the perfect clustering number of data sets accurately. The clustering performance of S_2 is very good and it can accurately and effectively classify all adopted UCI datasets, so it is included in our selection range.

9 Simulation Results and Analysis on Logarithm Validity Functions (Group 2)

From Fig. 6a–f, it can be observed that the performance of the nine logarithmic validity functions in the clustering and classification process on UCI data sets is very poor, and it is impossible to classify any datasets successfully. Therefore, it shows that the nine logarithmic validity functions in the simulation experiments cannot be used to classify UCI data sets, it is necessary to select and compare other logarithmic validity functions composed of components.

Table 10 The best cluster number of the exponential validity functions for UCI datasets (Group 1)

Data	Optimal c	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
Iris	3	5	3	3	11	2	2	2	2	2
Seeds	3	5	3	3	3	2	2	2	2	2
Balance	3	5	3	3	4	2	2	2	2	2
Hfcr	4	6	4	3	4	2	2	2	2	2
Glass	6	5	6	3	13	2	2	2	2	2
Cooking	7	5	7	4	3	2	2	2	2	2

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

Table 11 The best cluster number of the exponential validity functions for UCI datasets (Group 2)

Data	Optimal c	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	V_{16}	V_{17}	V_{18}
Iris	3	2	3	3	2	3	2	2	2	2
Seeds	3	2	3	3	3	3	2	2	2	2
Balance	3	2	4	3	4	2	4	2	2	2
Hfcr	4	2	3	3	4	3	3	2	2	2
Glass	6	2	3	3	2	2	2	2	2	2
Cooking	7	2	3	3	3	2	3	2	2	2

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

10 Simulation Results and Analysis on Logarithm Validity Functions (Group 3)

It can be observed from Fig. 7a–f that the simulation results of this group are similar to those of the second group of logarithmic clustering functions. These eight logarithmic validity functions also have no way to distinguish UCI datasets effectively. As can be seen from Fig. 7e, Glass dataset can only be divided into three categories by S_{22} , but it is still not the optimal number of clusters. The other validity functions in this group can only divide each data set into two categories, and none of them

Table 12 The best cluster number of the exponential validity functions for UCI datasets (Group 3)

Data	Optimal c	V_{19}	V_{20}	V_{21}	V_{22}	V_{23}	V_{24}	V_{25}	V_{26}
Iris	3	2	2	3	2	2	2	2	2
Seeds	3	2	2	3	2	2	2	2	2
Balance	3	3	3	4	3	3	3	3	3
Hfcr	4	2	2	3	3	2	2	2	2
Glass	6	2	2	2	3	3	3	2	3
Cooking	7	2	2	2	2	2	2	2	2

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

Table 13 The best cluster number of the logarithmic validity functions for UCI datasets (Group 1)

Data	Optimal c	S_1	S_1	S_1	S_1	S_1	S_1	S_1	S_1	S_1
Iris	3	5	3	2	2	2	2	2	2	2
Seeds	3	5	3	2	2	2	2	2	2	2
Balance	3	5	3	2	4	3	2	4	2	4
Hfcr	4	7	4	2	2	2	2	2	2	2
Glass	6	5	6	2	3	2	2	3	2	2
Cooking	7	5	7	2	2	2	2	2	2	2

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

Table 14 The best cluster number of the logarithmic validity functions for UCI datasets (Group 2)

Data	Optimal c	S ₁₀	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇	S ₁₈
Iris	3	2	2	2	2	2	2	2	2	2
Seeds	3	2	2	2	2	2	2	2	2	2
Balance	3	2	2	2	6	2	6	2	2	2
Hfcr	4	2	2	2	2	2	2	2	2	2
Glass	6	2	2	2	2	2	2	2	2	2
Cooking	7	2	4	2	2	2	2	2	2	2

Table 15 The best cluster number of the logarithmic validity functions for UCI datasets (Group 3)

Data	Optimal c	S ₁₉	S ₁₉	S ₁₉	S ₁₉	S ₁₉	S ₁₉	S ₁₉	S ₁₉
Iris	3	2	2	2	2	2	2	2	2
Seeds	3	2	2	2	2	2	2	2	2
Balance	3	2	2	2	2	2	2	2	2
Hfcr	4	2	2	2	2	2	2	2	2
Glass	6	2	2	2	3	2	2	2	2
Cooking	7	2	2	2	2	2	2	2	2

can be classified successfully. Therefore, the performance of these eight logarithmic validity functions in this group is still very poor and is not in the selected range.

The validity functions V and S are the best clustering validity evaluation criteria when both the minimum value c is taken. From Fig. 2 to Fig. 7, 52 validity functions constructed in this paper can clustering the selected data sets into several categories, which is the improved quantitative result. In practice, if a data is given, in the cluster number $2 \leq c \leq \sqrt{n}$, the two optimal validity functions V_2 and S_2 compared with the 52 validity functions are used to determine how much is the optimal cluster number.

11 Simulation Comparison with Single Clustering Validity Functions and Combined Clustering Validity Methods

11.1 Simulation Comparison with Single Clustering Validity Functions

Through the simulation comparison on the above six groups of clustering validity functions, six UCI datasets can only be divided successfully by the exponential validity function V_2 and the logarithmic validity function S_2 . In order to fully reflect the clustering performance of the validity function V_2 and S_2 , this paper selected the eight common clustering validity functions (V_{MPC} , V_{XB} , V_{PCAES} , V_{WL} , V_{FM} , V_{ZLF} , V_{HY} , and V_{WG}) and 8 kinds of commonly used UCI datasets to carry out the comparative experiments

with V_2 and S_2 . The function description and optimal cluster number of the eight typical clustering validity functions are listed in Table 16, where $\bar{v} = \sum_{i=1}^c v_i/c$ represents the mean value of the cluster centers. The geometric meaning of $u_{mj} = \min_{1 \leq i \leq c} \sum_{j=1}^n (u_{ij})^2$ can refer to the component CP_4 shown in Sect. 3. 1 and $median\|v_i - v_k\|^2$ represents the median distance between two cluster centers.

Table 17 lists the number of samples, attributes, and categories of UCI datasets selected in this experiment. Vehicle and Led7 datasets are added in the comparison process to improve the experiment integrity by adopting the 10 normalized fuzzy validity functions (V_{MPC} , V_{XB} , V_{PCAES} , V_{WL} , V_{FM} , V_{ZLF} , V_{HY} , V_{WG} , V_2 , and S_2) in the same coordinate system. The simulation results are shown in Fig. 7a–h. Finally, for different UCI datasets, the optimal number of clusters for each cluster validity function is listed in Table 18.

As can be seen from Fig. 8a, b, Iris and Seeds data sets can be successfully divided into three categories by the four validity functions (V_{HY} , V_{WG} , V_2 , S_2), while other six validity functions could not get the correct cluster number. This shows that when dealing with Iris and Seeds datasets with complex structures, V_2 and S_2 have the better performance compared with some other typical clustering validity functions. As shown in Fig. 8e-h, only V_2 and S_2 can accurately distinguish Glass, Cooking, Vehicle and Led7 datasets. It can be observed from Fig. 8c and d, except that V_2 and S_2 , there are other validity functions that distinguish individual datasets from UCI datasets, such as V_{WL} can get the correct number of clusters in the Balance data set. The optimal clustering number of Hfcr data set can be correctly divided by V_{ZLF} , whose the best cluster number can be divided into four. As shown in Fig. 8a–h. The best classification number of all UCI datasets can only be found by V_2 and S_2 , which indicates that when faced with data sets with overlapping samples, noise data and higher dimensions, the classification number of UCI datasets can be found. The clustering result of V_2 and S_2 is better than other clustering validity functions.

This paper constructs 52 validity functions. The computational complexity of simulating these 52 validity

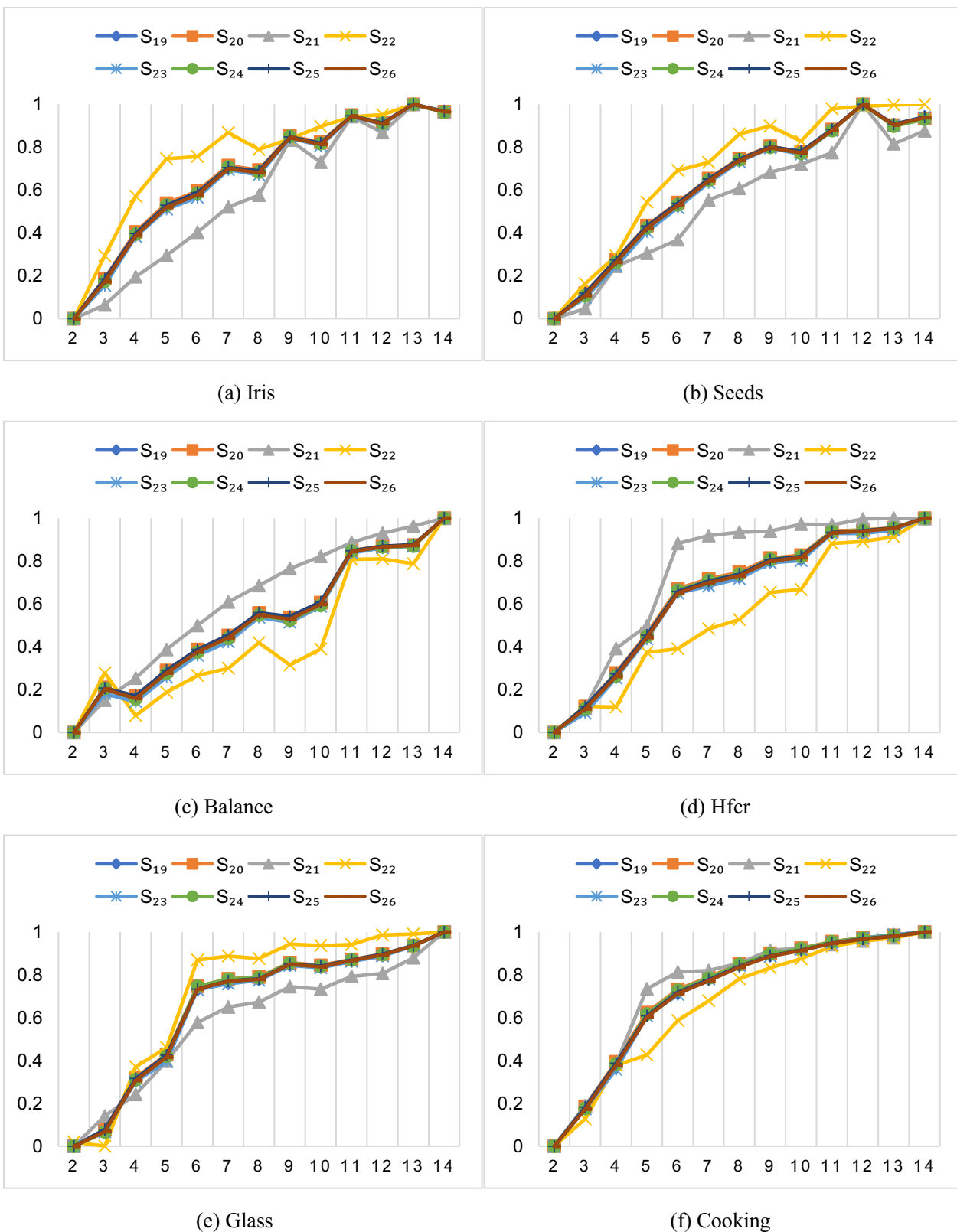


Fig. 7 Variation trend of normalized logarithmic clustering validity functions (Group 3)

Table 16 Typical clustering validity functions

Validity Index	Function description	Optimal c
Modification of partition coefficient	$V_{MPC} = 1 - \frac{c}{c-1} (1 - \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2)$	Max
Xie and Beni	$V_{XB} = \frac{\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \ v_i - v_j\ ^2}{\min_{i \neq j} \ v_i - v_j\ ^2}$	Min
Partition coefficient and exponential separation	$V_{PCAES} = \sum_{i=1}^c \sum_{j=1}^n \frac{u_{ij}^2}{u_{ij}} - \sum_{i=1}^c \exp\left(\frac{-\min_{k \neq i} \ v_i - v_k\ ^2}{\beta_T}\right)$	Max
Chih-Hung Wu	$V_{WL} = \frac{\sum_{i=1}^c \left(\frac{\sum_{j=1}^n u_{ij}^2 \ s_j - v_i\ ^2}{\sum_{j=1}^n u_{ij}} \right)}{\min_{i \neq k} \ v_i - v_k\ ^2 + \text{median} \ v_i - v_k\ ^2}$	Min
FM-index	$V_{FM} = \frac{\sum_{i=1}^c \sum_{j=1}^n (u_{ij} - \frac{1}{c})^2 \ v_i - v_j\ ^2}{n \min_{i \neq j} \ v_i - v_j\ ^2} \times (-\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n [u_{ij} \log_a(u_{ij})])$	Min
Zhu	$V_{ZLF} = \frac{\text{comp}}{\text{sep}} = \frac{\sum_{j=1}^n \sum_{i=1}^c \frac{1 - \min_{i \neq j} u_{ij}}{\ s_j - v_i\ }}{\sum_{k=1}^c \sum_{i=1, i \neq k}^c \frac{\ v_i - \bar{v}\ ^{c/(c-1)}}{2}}$	Min
Wang2020	$V_{HY} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \ x_j - v_i\ ^2 + \frac{1}{c} \sum_{i=1}^{c-1} \sum_{k=i+1}^c \left[\frac{1}{n} \sum_{j=1}^n \min(u_{ij}, u_{kj}) \right]}{\min_{1 \leq i \leq c} \sum_{j=1}^n u_{ij}^2 + \min_{i \neq k} \ v_i - v_k\ ^2}$	Min
Guan Wang	$V_{WG} = \frac{\sum_{i=1}^c \sum_{j=1}^n (u_{ij})^2 \ x_j - v_i\ ^2 + \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n \exp\left(-\frac{\ x_j - v_i\ ^2}{\epsilon}\right) + \min_{i \neq j} \left(\frac{1}{n} \sum_{j=1}^n (1 - u_{ik} - u_{jk})\right)}{\min_{1 \leq i \leq c} \sum_{j=1}^n u_{ij}^2 + \min_{i \neq k} \ v_i - v_k\ ^2 + \frac{1}{c} \sum_{i=1}^c \ v_i - \bar{v}\ ^2}$	Min
Jiaxu Liu	$V_2 = \frac{\sum_{i=1}^c \sum_{j=1}^n e^{(u_{ij})^2} \ s_j - v_i\ ^2 + \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n \exp\exp\left(-\frac{\ x_j - v_i\ ^2}{\epsilon}\right) + \min_{i \neq j} \left(\frac{1}{n} \sum_{j=1}^n \exp(1 - u_{ik} - u_{jk})\right)}{\min_{1 \leq i \leq c} \sum_{j=1}^n \exp(u_{ij}^2) + \min_{i \neq k} \exp(\ v_i - v_k\ ^2)}$	Min
Jiaxu Liu	$S_2 = \frac{\sum_{i=1}^c \sum_{j=1}^n \ln((u_{ij})^2) \ x_j - v_i\ ^2 + \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n \left(-\frac{\ x_j - v_i\ ^2}{\epsilon}\right) + \min_{i \neq j} \left(\frac{1}{n} \sum_{j=1}^n \ln(1 - u_{ik} - u_{jk})\right)}{\min_{1 \leq i \leq c} \sum_{j=1}^n 2 \ln(u_{ij}) + \min_{i \neq k} 2 \ln(\ v_i - v_k\)}$	Min

Table 17 UCI datasets (comparative experiments)

Data sets	Data numbers	Attributes	Classes
Iris	150	4	3
Seeds	210	7	3
balance	635	4	3
Hfcr	299	13	4
glass	214	9	6
Cooking	4900	3	7
Vehicle	846	18	4
Led7	500	7	10

Table 18 The best cluster number of the different validity functions for UCI datasets

Data	Optimal c	V_{MPC}	V_{XB}	V_{PCAES}	V_{WL}	V_{FM}	V_{ZLF}	V_{HY}	V_{WG}	V_2	S_2
Iris	3	2	2	11	2	2	2	3	3	3	3
Seeds	3	2	2	14	2	2	2	3	3	3	3
Balance	3	10	10	2	3	2	10	2	2	3	3
Hfcr	4	10	8	14	10	2	4	8	8	4	4
Glass	6	2	2	12	2	2	2	2	2	6	6
Cooking	7	2	2	13	2	2	2	2	2	7	7
Vehicle	4	2	2	14	2	2	2	2	2	4	14
Led7	10	12	2	14	2	2	2	2	2	6	10

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

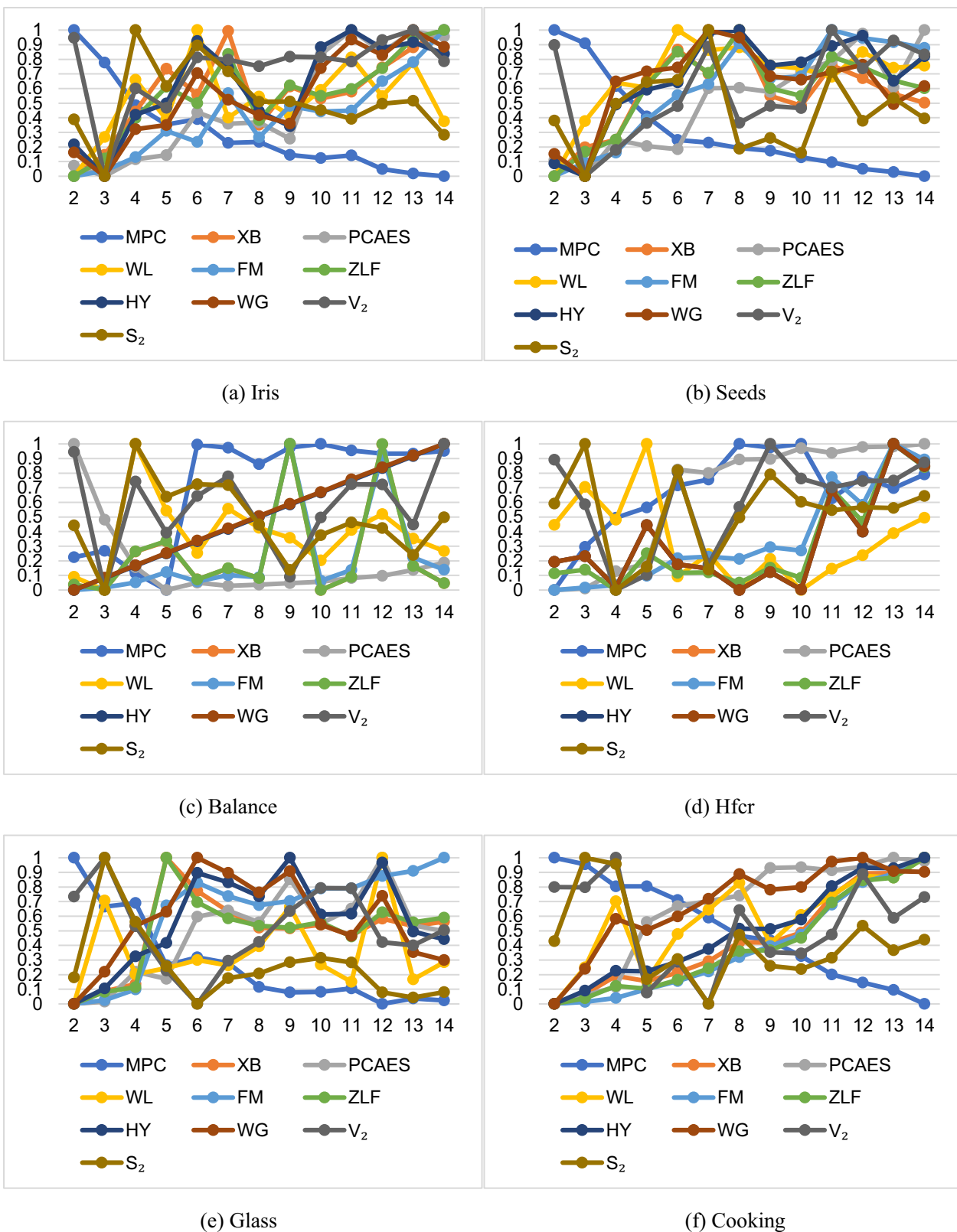


Fig. 8 Variation trend of normalized clustering validity functions

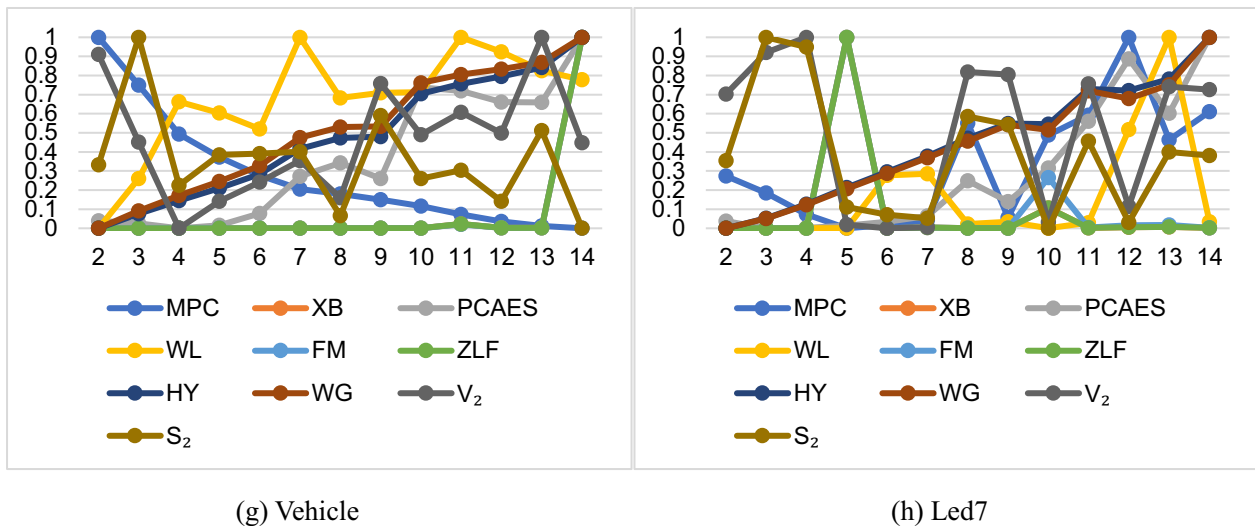


Fig. 8 continued

function experiments is the same as the calculation complexity of a single validity function. The single validity function calculation complexity takes time, and the time complexity is positively proportional to the number of validity function so that the time complexity is a positive linear relationship with the constructed validity function. The selected eight common clustering validity functions (V_{MPC} , V_{XB} , V_{PCAES} , V_{WL} , V_{FM} , V_{ZLF} , V_{HY} , and V_{WG}) are designed through subjective experience based on the basic concepts of FCM clustering algorithm and validity evaluation criteria. The improved partition coefficients V_{MPC} correct the V_{PC} existing monotonic reduction problem, but still lack a direct connection to the geometry of the data set. V_{XB} is the structure of the data set clustering validity function, but this calculation method often ignore the noise data. V_{PCAES} is a function of exponential operation proposed by Wu and Yang et al. V_{WL} has best classification effect among these eight validity functions. V_{FM} has poor performance on the data set with noise. On the other hand, the exponential validity function V_2 and the log validity function S_2 used for comparison are the two best clustering validity functions by the objective combination and then performance comparison. These two validity functions proposed in this paper not only avoid the strong subjective randomness of the subjective experience in designing the validity function, but also greatly reduce the limitations in real applications.

11.2 Simulation Comparison with Combined Clustering Validity Methods

In order to better highlight the advantages of the weighting method and validity function proposed in this paper compared with other traditional methods, so as to enhance its

persuasiveness, four combined clustering validity methods introduced in Sect. 2.2 (DWSVF, FWSVF, WSCVI, and HWCVF) and 8 UCI datasets are selected to carry out the simulation experiments. They are compared with V_2 and S_2 through simulation experiments. Then the six clustering validity evaluation methods (DWSVF, FWSVF, WSCVI, HWCVF, V_2 , and S_2) are placed in the normalized coordinate system, and the experimental results are shown in Fig. 9a–h. Finally, Table 19 lists the optimal number of clusters for each clustering validity evaluation method for different UCI datasets.

As can be seen from Fig. 8a–c, the Iris, Seeds and Balance datasets can be accurately classified into three subsets by DWSVF, WSCVI, V_2 , and S_2 . In Fig. 9d and h, it can be found that when processing Hfcr, Glass, Cooking, Vehicle, and Led7 datasets, only V_2 and S_2 can obtain the best cluster number of these six datasets. None of the other clustering combination evaluation methods can accurately divide these six datasets. Obviously, V_2 and S_2 are much better than other classical clustering combination evaluation methods. From the experimental results shown in Fig. 8a–h, the following conclusions can be drawn. For the above 8 commonly used UCI data sets, the classification performance of V_2 and S_2 is also better.

12 Conclusion

In this paper, a new combination empowerment method is defined through subjective empowerment and standard separation method empowerment, and a component construction method of FCM clustering validity function based on five cluster performance evaluation components was proposed. By using the UCI data sets, the best clustering

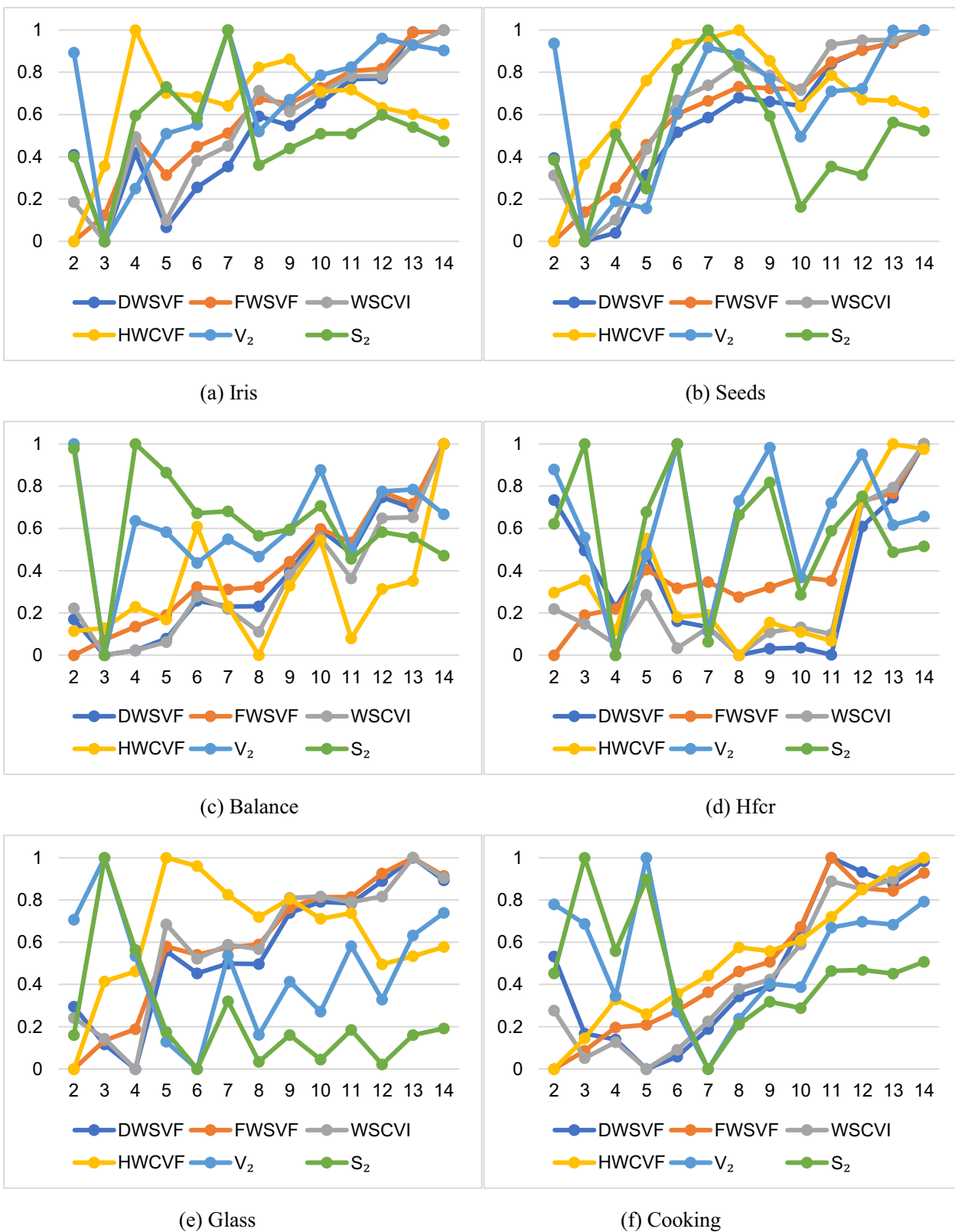


Fig. 9 Variation trend of normalized clustering combination evaluation methods

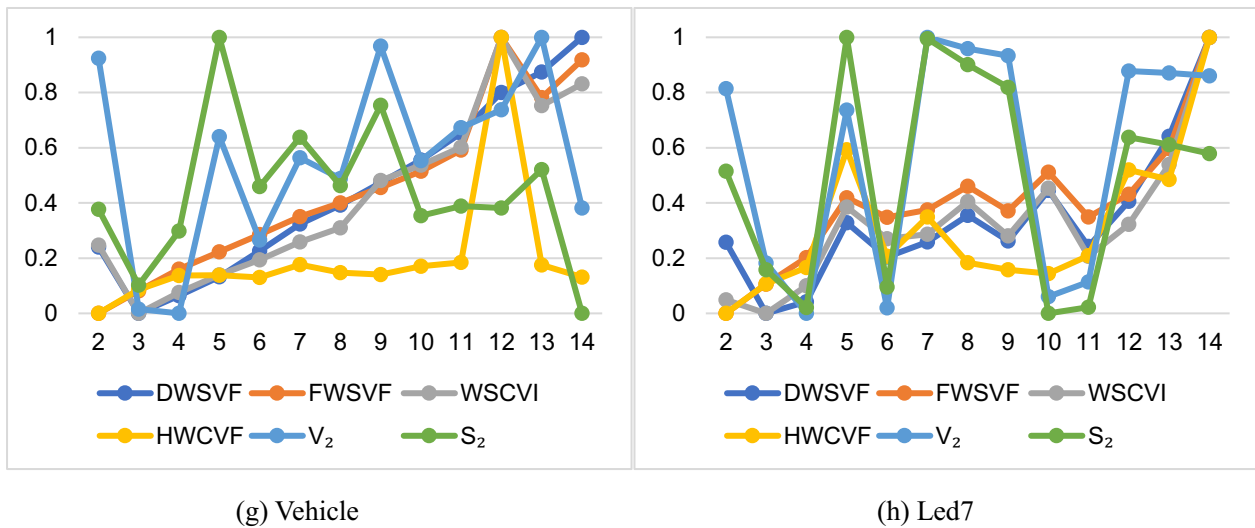


Fig. 9 continued

Table 19 The best cluster number of different clustering combination evaluation methods for UCI datasets

Data	Optimal c	DWSVF	FWSVF	WSCVI	HWCVF	V_2	S_2
Iris	3	3	2	3	2	3	3
Seeds	3	3	2	3	2	3	3
Balance	3	3	2	3	8	3	3
Hfcr	4	8	2	8	8	4	4
Glass	6	4	2	4	2	6	6
Cooking	7	5	2	5	2	7	7
Vehicle	4	3	2	3	2	4	14
Led7	10	3	2	3	2	4	10

Bold indicates that the clustering validity function can obtain the best number of clusters for this dataset

validity function was selected by carrying out the simulation comparison. Finally, the eight commonly used single clustering validity functions and four typical combined clustering validity evaluation methods were simulated and verified on 8 UCI datasets. The simulation results prove that the proposed validity function based on this construction method is more complex in the data structure. The datasets with noise and overlapping data interference is better than other single or combination clustering validity functions. The screening and comparison of many experiments prove the proposed constructor more objective. With a strong scientific theoretical basis, it can reduce the one-sided aspect of proposing the validity function based on the subjective intention and improve the research depth of the clustering validity function. However, the construction method also has its own limitations, so some constructed superior validity functions will be selected to integrate the method, and comprehensively use multiple validity functions for evaluation. Therefore, the component integration for the clustering validity function is put into further study.

Acknowledgements This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. LJKZ0293), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Author Contributions JL participated in the data collection, analysis, algorithm simulation, and draft writing. JW participated in the concept, design, interpretation, and commented on the manuscript. GW, XZ, HW, and DJ participated in the critical revision of this paper.

Declarations

Conflict of interest The authors declare that there is no conflict of interests regarding the publication of this article.

References

1. Leung, S.H., Wang, S.L., Lau, W.H.: Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Trans. Image Process.* **13**(1), 51–62 (2004)
2. Lu, J., Yuan, X., Yahagi, T.: A method of face recognition based on fuzzy c-means clustering and associated sub-NNs. *IEEE Trans. Neural Netw.* **18**(1), 150–160 (2007)

3. Wang, Y.H., Zhao, H.C.: PolSAR image segmentation by mean shift clustering in the tensor space. *Acta Automatica Sinica* **36**(6), 778–806 (2010)
4. Hung, W.L., Yang, M.S., Chen, D.H.: Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation. *Pattern Recogn. Lett.* **29**(9), 1317–1325 (2008)
5. Maulik, U., Saha, I.: Automatic fuzzy clustering using modified differential evolution for image classification. *IEEE Trans. Geosci. Remote Sens.* **48**(9), 3503–3510 (2010)
6. Sulaiman, S.N., Isa, N.A.M.: Adaptive fuzzy-K-means clustering algorithm for image segmentation. *IEEE Trans. Consum. Electron.* **56**(4), 2661–2668 (2010)
7. Tung, F., Wong, A., Clausi, D.A.: Enabling scalable spectral clustering for image segmentation. *Pattern Recogn.* **43**(12), 4069–4076 (2010)
8. Yang, S.Y., Wu, R.X., Wang, M., Jiao, L.C.: Evolutionary clustering based vector quantization and SPIHT coding for image compression. *Pattern Recogn. Lett.* **31**(13), 1773–1780 (2010)
9. Wang, Z.M., Song, Q., Soh, Y.C., Sim, K.: Robust curve clustering based on a multivariate t-distribution model. *IEEE Trans. Neural Netw.* **21**(12), 1976–1984 (2010)
10. Liu, P.X., Meng, M.Q.H.: Online data-driven fuzzy clustering with applications to real-time robotic tracking. *IEEE Trans. Fuzzy Syst.* **12**(4), 516–523 (2004)
11. Celikyilmaz, A., Turksen, I.B.: Enhanced fuzzy system models with improved fuzzy clustering algorithm. *IEEE Trans. Fuzzy Syst.* **16**(3), 779–794 (2008)
12. Hyong-Euk, L., Kwang-Hyun, P., Bien, Z.Z.: Iterative fuzzy clustering algorithm with supervision to construct probabilistic fuzzy rule base from numerical data. *IEEE Trans. Fuzzy Syst.* **16**(1), 263–277 (2008)
13. Goldberger, J., Tamir, T.: A hierarchical clustering algorithm based on the Hungarian method. *Pattern Recogn. Lett.* **29**(11), 1632–1638 (2008)
14. Bezdek, J.C.: *Pattern Recognition with Fuzzy Object Algorithms*, pp. 54–57. Plenum Press, New York (1981)
15. MacQueen J.: Some Methods of Classification and Analysis of MultiVariate Observations. *Proc of Berkeley Symposium on Mathematical Statistics and Probability*, 281–297 (1967).
16. Yodern, J., Priebe, C.E.: Semi-supervised K-means++. *J. Stat. Comput. Simul.* **87**(13), 2597–2608 (2017)
17. Geng, Z., Chengchang, Z., Huayu, Z.: Improved K-means algorithm based on density Canopy. *Knowl.-Based Syst.* **145**, 289–297 (2018)
18. Hiep, N.H.: Privacy-preserving mechanisms for k-modes clustering. *Comput. Secur.* **78**, 60–75 (2018)
19. J. C. Bezdek, N. R. Pal.: Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B(Cybernetics)*, **28**(3), 301–315 (1998).
20. D. A. Simovici, S. Jaroszewicz.: An axiomatization of partition entropy. *IEEE Transactions on Information Theory*, **48**(7), 2138–2142 (2002).
21. Silva, L., Moura, R., Canuto, A.M.P., Santiago, R.H.N., Bedregal, B.: An Interval-Based Framework for Fuzzy Clustering Applications. *IEEE Trans. Fuzzy Syst.* **23**(6), 2174–2187 (2015)
22. Chen, M.Y., Linkens, D.A.: Rule-base self-generation and simplification for data-driven fuzzy models. *Fuzzy Sets Syst.* **142**(2), 243–265 (2004)
23. Chen J.S., Pi D. C.: A cluster validity index for fuzzy clustering based on non-distance. In 2013 International Conference on Computational and Information Science. *IEEE*, 880–883 (2013).
24. Joopudi, S., Rathi, S.S., Narasimhan, S., Rengaswamy, R.: A new cluster validity index for fuzzy clustering. *IFAC Proc.* Vol. **46**(32), 325–330 (2013)
25. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(6), 841–847 (1991)
26. Wu, K.-L., Yang, M.-S.: A cluster validity index for fuzzy clustering. *Pattern Recogn. Lett.* **26**(9), 1275–1291 (2005)
27. C. Wu, C. Ouyang, L. Chen and L. Lu.: A New Fuzzy Clustering Validity Index With a Median Factor for Centroid-Based Clustering. *IEEE Transactions on Fuzzy Systems*, **23**(3): 701–718 (2004).
28. Meng, L., Hu, C.: Cluster Validity Index Based on Measure of Fuzzy Partition. *Comput. Eng.* **33**(11), 15–17 (2007)
29. Zhu, L.F., Wang, J.S., Wang, H.Y.: A novel clustering validity function of FCM clustering algorithm. *IEEE Access* **7**, 152289–152315 (2019)
30. Wang, H.Y., Wang, J.S., Zhu, L.F.: A new validity function of FCM clustering algorithm based on the intra-class compactness and inter-class separation. *J. Intell. Fuzzy Syst.* **40**(6), 12411–12432 (2021)
31. Wang, G., Wang, J.S., Wang, H.Y.: Fuzzy C-means clustering validity function based on multiple clustering performance evaluation components. *Int. J. Fuzzy Syst.* **1**, 1–29 (2022)
32. Sheng, W.G., Swift, S., Zhang, L.: A weighted sum validity function for clustering with a hybrid niching genetic algorithm. *IEEE Trans. Syst. Man Cybern. B* **35**(6), 1156–1167 (2005)
33. Dong H.B, Hou W, Ying G.S.: An evolutionary clustering algorithm based on adaptive fuzzy weighted sum validity function. In: 2010 Third International Joint Conference on Computational Science and Optimization, *IEEE*, 357–361 (2010).
34. Wu Z. F, Huang H. K.: A dynamic weighted sum validity function for fuzzy clustering with an adaptive differential evolution algorithm. In: 2010 Third International Joint Conference on Computational Science and Optimization, *IEEE*, 362366 (2010).
35. Wang, H.Y., Wang, J.S.: Combination evaluation method of fuzzy C-means clustering validity based on hybrid weighted strategy. *IEEE Access* **9**, 27239–27261 (2021)

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Jia-Xu Liu received his B. Eng. degree in Automation from University of Science and Technology Liaoning, China in 2020. He is currently pursuing the M. Eng. degree with the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China. His main research interest is machine learning and deep learning.

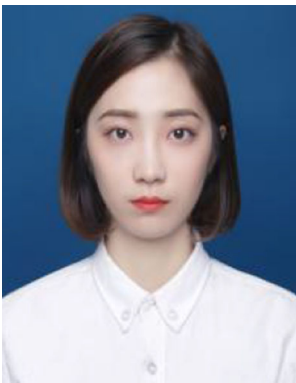


Jie-Sheng Wang received his B. Sc. And M. Sc. degrees in Control Science and Engineering from University of Science and Technology Liaoning, China in 1999 and 2002, respectively, and his Ph. D. degree in Control Science and Engineering from Dalian University of Technology, China in 2006. He is currently a professor and Doctor's Supervisor in School of Electronic and Information Engineering, University of Science and

Technology Liaoning. His main research interest is modeling of complex industry process, intelligent control and Computer integrated manufacturing.



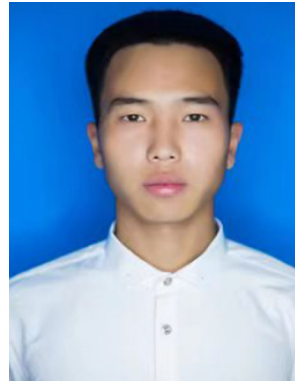
Guan Wang received the B. Eng. degree in Automation from University of Science and Technology Liaoning, China in 2020. He is currently pursuing the M. Eng. degree with the School of Control Science and Engineering, University of Science and Technology Liaoning, Anshan, China. His research interests include machine learning and intelligent optimization algorithm.



Xiao-Rui Zhao received her B. Eng. degree in Communication Engineering from University of Science and Technology Liaoning, China in 2020. She is currently pursuing the M. Eng. degree with the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China. Her main research interest is soft measurement and machine learning.



Hong-Yu Wang received his B. Eng. degree in Measuring Technology and Instrument from University of Science and Technology Liaoning, China in 2019. He is currently pursuing the M. Eng. degree with the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China. His main research interest is meta-heuristic algorithms and machine learning.



Di Jin received his B. Eng. degree in Measurement and Control Technology and Instruments from University of Science and Technology Liaoning, China in 2020. He is currently working in Inspection and Certification Center Liaoning. His main research interest is machine learning.