



Clustering Validity Function Fusion Method of FCM Clustering Algorithm Based on Dempster–Shafer Evidence Theory

Hong-Yu Wang¹ · Jie-Sheng Wang¹ · Guan Wang¹

Received: 16 March 2021 / Revised: 24 June 2021 / Accepted: 5 August 2021 / Published online: 10 October 2021
© Taiwan Fuzzy Systems Association 2021

Abstract With the deepening of the research on clustering algorithm, clustering validity has become an indispensable part of cluster analysis. However, due to the complexity of data structure and different attributes, any clustering validity function cannot be applied to all datasets, so clustering validity function has been bringing forth new ones. Therefore, this paper proposes a clustering validity function fusion model based on D–S evidence theory (DS-CVFFM), which adopts FCM clustering algorithm as the base algorithm, calculates the values of different validity functions, and then uses the values of different clustering validity functions as the evidence source to construct the basic probability assignment function (BPA). Finally, it integrates with the fusion rules of D–S evidence theory, and outputs the optimal clustering number according to the decision conditions. DS-CVFFM uses the information fusion of multiple clustering validity functions to judge the number of optimal clusters without the need to propose complex validity functions, and avoid the influence of expert factors in the weighted combination clustering validity evaluation method. Finally, 4 sets of artificial datasets and 14 sets of UCI datasets are selected to verify the effectiveness of the proposed model. The experimental results show that compared with the traditional clustering validity evaluation methods, the proposed fusion model has a significant improvement in the accuracy of judging the optimal number of clusters, and the stability is improved under different values of fuzzy exponent, which can

overcome the shortcomings of traditional clustering validity evaluation methods.

Keywords D–S evidence theory · FCM clustering algorithm · Clustering validity function · Information fusion

1 Introduction

Cluster analysis is an important research topic in the field of machine learning and pattern recognition. As an unsupervised learning method, cluster analysis is to divide similar samples into one category and dissimilar samples into different categories for data without prior knowledge, so that the samples of the same category are as similar as possible and the samples of different categories are as different as possible [1]. Clustering methods are mainly divided into hierarchy-based clustering, partition-based clustering, density-based clustering, model-based clustering and more [2–5]. FCM clustering algorithm [6], as one of the algorithms for partition clustering, introduces the concept of fuzzy set [8] into K-means clustering algorithm [7], so that hard clustering can be extended to fuzzy clustering. In this way, data samples that are similar between classes have a better basis for division, and the clustering results become better and more closer to the practical needs [9]. FCM clustering algorithm has become one of the most widely used clustering algorithms by virtue of its simple principle, fast calculation speed and wide range of problem solving [10]. However, FCM clustering algorithm needs to verify the effectiveness of clustering, so as to determine the optimal number of clusters and judge the quality of clustering results. From the current research point of view, the research of clustering validity can be roughly divided into

✉ Jie-Sheng Wang
wang_jiesheng@126.com

¹ School of Electronic and Information Engineering,
University of Science & Technology Liaoning,
Anshan 114044, China

the research of clustering validity function and the research of combined clustering validity evaluation method.

The research on clustering validity function mainly focuses on the following two aspects: (1) Fuzzy clustering validity function based on membership degree only. For example, Bezdek proposed the partition coefficient [11], which was the first index to describe the validity of fuzzy clustering by the coincidence degree of data objects. Referring to the partition coefficient and Shannon's information theory [12], the partition entropy [13] was proposed. Dave proposed the modified partition coefficient [14] to suppress the monotonous change of the partition entropy and partition coefficient with the number of clusters, and the P index [15] proposed by Chen and Links in 2004. (2) Fuzzy clustering validity function based on membership degree and geometric structure of data set. For example, Xie-Beni validity function [16], Fakuyame-Sugeno validity function proposed in 1989 [17], clustering validity function proposed by Knows [18], the clustering validity function proposed by Wang [19], the clustering validity function [20] proposed by Zhu and more. However, due to the complexity of data structure and different attributes, any clustering validity function cannot be applied to all datasets. Because of this, the validity function of fuzzy clustering has been bringing forth new ideas, which also leads to the more and more complex structure design of the fuzzy clustering validity function.

The research on the validity evaluation method of combinatorial clustering mainly focuses on the weighted combination method. Weighted sum validity function (WSVF) proposed by Sheng in 2005 is a combination of hard clustering validity functions [21]. Xu also studied the combination method of hard clustering validity function [22]. Dong proposed the fuzzy weighted sum validity function (FWSVF) based on the idea of Sheng [23]. After that, Wu changed the weighted method of FWSVF and proposed a dynamic weighted sum validity function (DWSVF) [24]. Li extended the index of FWSVF to the combination of 10 validity functions and proposed weighted sum type of cluster validity index (WSCVI) [25]. Wang proposed a hybrid weighted method combining subjective weighted and objective weighted, which balanced the influence of subjectivity and objectivity on the results, which is named as the hybrid weighted clustering validity function (HWCVF) [26]. Compared with the clustering validity functions, the combined clustering validity method only needs the weighted combination of the existing validity functions. Its form is simple and easy to understand, but the weighted factor is often difficult to determine.

Since the cluster analysis is to cluster the data set without labels, after clustering the data set using the FCM clustering algorithm, it is necessary to evaluate the validity

of the final result to obtain the best number of clusters and then judge the clustering results are good or bad. Single clustering validity function will fail with the number of data samples, the change of dimensions, and the influence of noise data and overlapping data, meanwhile, the clustering validity function will become more and more complex with the advent of the era of big data. Combined clustering validity evaluation method will be difficult to select the weighting method, and prone to over fitting problems. In order to improve the accuracy of the clustering results of the FCM clustering algorithm and the judgment of the optimal number of clusters, it is not feasible to rely only on a single validity function and a combined validity evaluation method. It is necessary to find a method that can be applied to a large amount of data which can be used to judge the final clustering result. For this reason, this paper starts from the perspective of fusing multiple validity functions to determine the optimal number of clusters and improving the stability and accuracy of cluster validity evaluation. A fusion model (DS-CVFFM) based on D-S evidence theory [27–29] is proposed to integrate the information of multiple clustering validity functions.

Based on the idea of D-S evidence theory, this paper proposes a fusion model (DS-CVFFM) based on the information of clustering validity functions. The main idea of DS-CVFFM is to fuse multiple validity functions using D-S evidence theory. Firstly, FCM clustering algorithm is used to cluster the data set, and then several validity functions are used to judge the validity of the clustering results. There are some differences in the optimal cluster number of the same data set by different validity functions, and then normalize the values of different validity functions. The normalized values are used as propositions to construct the basic probability assignment function (BPA) of the validity function. Finally, the BPA value of the validity functions is taken as the input, and the D-S evidence theory is used to judge the confidence. The number of clusters corresponding to the maximum confidence value is the optimal number of clusters.

Finally, the feasibility of DS-CVFFM is verified by experimental simulation. The artificial data sets and UCI data sets are selected for experiments, and the proposed fusion model DS-CVFFM is compared with the traditional validity evaluation methods. The experimental results show that the accuracy of DS-CVFFM is significantly better than the traditional validity evaluation methods under the same data sets and experimental parameters. The experimental results show that DS-CVFFM has best stability under different values of fuzzy weighted. Therefore, DS-CVFFM can effectively avoid the influence of human factors caused by subjective weighting and the problem of objective weighting selection. It can also avoid the problem that a

single cluster validity function is easy to fail in the face of datasets with different structures. At the same time, it does not need to design a complex structure validity function, while taking into account the advantages of multiple validity functions, which provides a new solution for the research of fuzzy clustering effectiveness evaluation method.

2 FCM Clustering Algorithm and Clustering Validity Functions

2.1 FCM Clustering Algorithm

The fuzzy C-means (FCM) clustering algorithm proposed by Bezdek is the most representative one of the fuzzy clustering methods, and is also the most widely used one of the clustering algorithms. FCM clustering algorithm divides n objects x_1, x_2, \dots, x_n of data X into c fuzzy clusters by finding the minimum objective function, which is defined as follows [30]:

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

where $J_m(U, V)$ is the square error clustering criterion and the minimum value of $J_m(U, V)$ is called the minimum square error stationary point, $V = \{v_1, v_2, \dots, v_n\}$ is the corresponding cluster center divided on the data set X , parameter c represents the number of clusters. The parameter m is the fuzzy weighted, which is used to control the degree of fuzziness within the scope $m \in (1, \infty)$, $x_j - v_i$ represents the European distance between the data object x_j and cluster center v_i , u_{ij} ($0 \leq u_{ij} \leq 1$) is the membership degree of data object x_j of cluster center v_i . In Eq. (1), $u_{ij} \in U$, where in U is a membership matrix with fuzzy division of $c \times n$ that must satisfy the following conditions, and U_0

is called the initial membership matrix: $\sum_{i=1}^c u_{ij} = 1 (j = 1, 2, \dots, n)$ and $0 < \sum_{j=1}^n u_{ij} < n (i = 1, 2, \dots, c)$.

The clustering validity is a problem of finding the optimal solution of c under the condition of J_m minimization. In fact, more attention must be paid to the clustering validity if cluster analysis is to make a great contribution to engineering applications. Since J_m decreases monotonically with the decrease of c , an effective partition evaluation criterion is needed. The process of FCM clustering algorithm is described as follows [31]:

Step 1: Fix the cluster parameter c and fuzzy weighted m (usually between 1.5 and 2.5). When $m = 1$, FCM clustering algorithm is equivalent to K-means clustering algorithm, when m approaches 1 indefinitely, FCM tends to be more and more clustering algorithm of hardening fraction. On the contrary, when m tends to be infinite, all data objects x_j and cluster center v_i will coincide, while data object x_j belongs to each cluster will have the same membership degree, with the value of $1/c$.

Step 2: Initialize fuzzy division membership matrix.

Step 3: According to Eq. (2), update the cluster centers $V = \{v_1, v_2, \dots, v_c\}$.

$$v_i = \frac{\sum_{j=1}^n u_{ij}^{(t)m} \cdot x_j}{\sum_{j=1}^n u_{ij}^{(t)m}} \quad (2)$$

Step 4: According to Eq. (3), update the fuzzy partition matrix $U = (u_{ij})_{c \times n}$;

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

For $i = 1, 2, \dots, c$ and $j = 1, 2, \dots, n$.

Step 5: Calculate $e = V_{t+1} - V_t$. If $e \leq \varepsilon$ (ε is a threshold value usually from 0.001 to 0.01), the algorithm is stopped and the final clustering result is calculated, otherwise $t = t + 1$ and repeat Step 2.

FCM clustering algorithm has been widely concerned by scholars since it came out, and now it has formed a very mature theory. However, it also has many defects, such as the complex calculation process, the need to calculate all the samples in the data set, and facing the complex datasets easy to fail and so on. Therefore, scholars improve and optimize FCM clustering algorithm from different perspectives. In 1994, Park combined gradient descent with FCM and proposed a Gradient based fuzzy c-means (GBFCM) algorithm (GBFCM) [32], which can speed up the convergence of the algorithm. Wu proposed a new FCM clustering algorithm based on Mercer kernel function (FKCM) [33], which improved the stability of FCM algorithm. Sanchez proposed the Fuzzy Granular Gravitational Clustering Algorithm (FGGCA) in 2014 [34], which uses Newtonian gravity and granular computing to analyze the similarity of data set samples. Based on Wu's idea, Ding proposed Gauss kernel based FCM (KFCM) [35] in 2016 using Gauss kernel function and genetic algorithm to update membership and optimize FCM. Rubio used type-2 fuzzy logic techniques to optimize fuzzy Possibilistic C-means (FPCM) algorithm [36] in 2017 to improve the

accuracy of PFCM. Kuo proposed a new clustering algorithm by combining heuristic algorithm with kernel function [37]. A design methodology for a Mamdani based Interval Type-2 Fuzzy Logic System (MAM-IT2FLS) with Center-Of-Sets defuzzification was proposed by Moreno in 2020[38], which adopts the descriptive statistics and granular computing theory to better define the limits of uncertainty within the Interval Type-2 Membership Functions (IT2MF) as extracted from available data.

2.2 Clustering Validity Functions

The clustering validity functions can be roughly divided into two categories. One is based on membership degree only and the other is based on membership degree and geometric structure of data set. The clustering validity functions based on membership degree lack the connection with the geometric structure of the data set, so the final results are often one-sided and the accuracy needs to be improved. At present, clustering validity functions based on membership degree and geometric structure is a popular research direction. This paper lists some of the more common clustering validity functions (the name of the function, the definition formula, and the criteria to determine the optimal number of c) as shown in Tables 1 and 2.

For the clustering validity functions shown in Tables 1 and 2, u_{ij} indicates the fuzzy membership matrix, n indicates the number of data samples, c indicates clustering number, m indicates fuzzy index, x_j indicates data samples, v_i indicates the clustering center, $v_i - x_j$ indicates the distance between clustering center and the sample data, $\bar{v} = \sum_{i=1}^c v_i/c$, $u_{mj} = \min_{1 \leq i \leq c} u_{ij}^2$, $\beta_T = \frac{1}{c} \sum_{i=1}^c v_i - \bar{v}^2$, $median v_i - v_k^2$ represents the median of the distance between two cluster centers.

In addition to the clustering validity function listed above, the validity function of fuzzy clustering has always been a hot topic in the study of cluster validity evaluation. For example, Feng proposed a validity function based on

compactness and overlap degree in 2016 [43]. The HF validity function proposed by Haouas in 2017 [44] with the improvement of V_{WL} . Geng also defined the overlap degree in 2019, and proposed a new validity function for three variables of compactness and variability [45]. Ouchica proposed an ECS validity function to standardize separation and compactness in 2020 [46]. Liu proposed a new validity function, which is the combination of V_{WL} and HF combination [47]. From the above study of validity function, the single validity function is more and more complex because of the different data set structure. Most of the validity functions are based on compactness, separation degree and overlap degree, but only different calculation methods. Because of the diversity of datasets, there is no one validity function that can be used for all datasets, which is also the reason why fuzzy clustering validity functions emerge endlessly.

2.3 Combination Clustering Validity Evaluation Methods

Sheng proposed WSVF [21] composite clustering effectiveness evaluation method for hard clustering algorithm is defined as follows Eq. (4).

$$\max WSVF = \sum_{i=1}^m w_i f_i(x) \tag{4}$$

where $m = 6$, $\sum_{i=1}^m w_i = 1, w_1 = w_2 = w_3 = w_4 = w_5 = w_6 = \frac{1}{6}$, $f_i(x)$ is consist of $1/V_{DB}, V_{SIL}, V_D, V_{33}, V_{CH}$ and V_{PBM} .

Dong proposed FWSVF by introducing the Sheng’s WSVF into the fuzzy clustering. The definition is described as Eq. (5).

$$\max FWSVF = \sum_{i=1}^r w_i f_i(x) \tag{5}$$

Table 1 Cluster validity function based on membership degree

Validity Index	Function description	Optimal c
Partition coefficient [11]	$V_{PC} = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2$	Max
Partition entropy [13]	$V_{PE} = -\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n [u_{ij} \log_a(u_{ij})]$	Min
Modification of partition coefficient [14]	$V_{MPC} = 1 - \frac{c}{c-1} (1 - V_{PC})$	Max
P-index [15]	$V_P = \frac{1}{n} \sum_{j=1}^n \max_i(u_{ij}) - \frac{1}{k} \sum_{i=1}^{c-1} \sum_{j=i+1}^c \left[\frac{1}{n} \sum_{k=1}^n \min(u_{ik}, u_{jk}) \right]$	Max

Table 2 Clustering validity function based on geometric structure and membership degree

Validity Index	Function description	Optimal c
Xie and Beni [16]	$V_{XB} = \frac{\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m v_i - x_j^2}{\min_{i \neq j} v_i - v_j^2}$	Min
Separation coefficient [39]	$V_{SC} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij} v_i - x_j^2}{\sum_{i=1}^c n_i \sum_{j=1}^c v_i - v_j^2}$	Min
Fukuyama and Sugeno [17]	$V_{FS} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m (x_j - v_i^2 - v_i - \bar{v}^2)$	Min
Kwons [18]	$V_K = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 x_j - v_i^2 + \frac{1}{c} \sum_{i=1}^c v_i - \bar{v}^2}{\min_{i \neq k} v_i - v_k^2}$	Min
Partition coefficient and exponential separation [40]	$V_{PCAES} = \sum_{i=1}^c \sum_{j=1}^n \frac{u_{ij}^2}{u_{mj}} - \sum_{i=1}^c \exp\left(\frac{-\min_{k \neq i} v_i - v_k^2}{\beta_T}\right)$	Max
Pakhira–Bandyopadhyay–Maulik [41]	$V_{PBMF} = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^2 = \left(\frac{\sum_{j=1}^n u_{1j} x_j - v_1 \times \max_{i,j=1,\dots,k} v_i - v_j}{k \times \sum_{i=1}^k \sum_{j=1}^n u_{ij} x_j - v_j}\right)^2$	Max
Chib-Hung Wu [42]	$V_{WL} = \frac{\sum_{i=1}^c \left(\frac{\sum_{j=1}^n u_{ij}^2 x_j - v_i^2}{\sum_{j=1}^n u_{ij}}\right)}{\min_{i \neq k} v_i - v_k^2 + \text{median} v_i - v_k^2}$	Min
Ling-Feng Zhu [20]	$V_Z = \frac{\text{comp}}{\text{sep}} = \frac{\sum_{j=1}^n \frac{1 - \min u_{ij}}{\sum_{i=1}^c x_j - v_i}}{\sum_{k=1}^c \sum_{i=1; i \neq k}^c v_i - \bar{v} / \frac{c(c-1)}{2}}$	Min
Hong-Yu Wang	$V_{HY} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 x_j - v_i^2 + \frac{1}{k} \sum_{i=1}^{c-1} \sum_{k=i+1}^c \left[\frac{1}{n} \sum_{j=1}^n \min(u_{ij}, u_{kj})\right]}{\min_{1 \leq i \leq c} \sum_{j=1}^n u_{ij}^2 + \min_{i \neq k} v_i - v_k^2}$	Min

where $r = 4$, $f_i(x)$ is consist of $1/V_{XB}$, $1/V_{PE}$, V_{PC} and V_{PBMF} , and w_i is consisted of different options based on different datasets.

WSCVI proposed by Li [25] is defined as Eq. (6).

$$\min WSCVI = \sum_{i=1}^N w_i CVI_i \tag{6}$$

where $N = 10$ and $\sum w_i = 1, 0 \leq w \leq 1$, where CVI_i is consisted of $1/V_{PC}$, $1/V_{NPC}$, V_{PE} , V_{NPE} , V_{XB} , V_K , $1/V_{PBMF}$, V_{FS} , V_T and V_{SC} .

DWSVF proposed by Wu can be defined as Eq. (7).

$$\min DWSVF = \sum_{i=1}^m w_i f_i(x) \tag{7}$$

where $w_i = \frac{\text{sum} f_i - f_i}{\text{sum} f_i}$, w_i is a dynamic weighting that can be weighted based on the values of different indicators, and $f_i(x)$ is composed of $1/V_{MPC}$, V_{XB} and V_{PBMF} .

Wang et al. proposed a hybrid weighted combination method HWCVF [26]. The definition of HWCVF is shown in Eq. (8).

$$\min HWCVF = \sum_{i=1}^m w_{\text{hybrid}} F_i(x) \tag{8}$$

where expert weighted and information entropy weighted are combined, that is to say that $w_{\text{hybrid}} = \theta w_{\text{object}} + (1 - \theta) w_{\text{subject}}$, where w_{object} is the expert weighted [48], w_{subject} is information entropy weighted

[49], θ is balance factor, and $F_i(x)$ are the clustering validity functions after standardization. The evaluation method is based on the idea of linear superposition of combined function with the V_{MPC} , V_{XB} , V_K , V_P , V_{PBMF} , V_{WL} , V_Z and V_{HY} .

At present, the research based on combinatorial validity function is not very deep, but from the above research status, most of the research on combinatorial validity function is based on weighted combination. Weighted is also a method of information fusion of validity functions, but the weighted factors are often difficult to determine. For example, the selection of subjective weighted is easily affected by the factors of decision-makers, and the objective weighted will ignore the accuracy of each validity function. Therefore, it is difficult to choose an appropriate weighting factor.

3 FCM Clustering Validity Function Fusion Based on D–S Evidence Theory

3.1 Basic Concepts and Combination Rules of D–S Evidence Theory

Dempster used the upper and lower limit probability to solve the problem of multi valued mapping at first [50]. After that, Shafer introduced the concept of trust function

to further develop evidence theory, and formed a set of mathematical methods to deal with uncertain reasoning problems, which also marks that evidence theory has officially become a complete theory to deal with uncertain problems. Then Yager proposed an improved evidence theory [51] by considering the problem of evidence conflict. Nowadays, evidence theory has been widely used in the field of information fusion [52–55]. This section will introduce the basic concepts and combination rules of D–S evidence theory [56]. In D–S evidence theory, set is used to represent proposition. If Θ is a finite set, then Θ is called the recognition framework, which contains c mutually exclusive elements. 2^Θ is the power set of Θ , which is defined as Eq. (9)

$$\begin{cases} 2^\Theta = \{A|A \subseteq \Theta\} \\ \Theta = \{\theta_1, \theta_2, \dots, \theta_c\} \end{cases} \quad (9)$$

The basic probability assignment function (BPA) on the recognition framework is a function m that maps from 2^Θ to $[0, 1]$. Under the condition of Eq. (10), m is called the mass function of the recognition framework.

$$\begin{cases} \sum_{A \subseteq \Theta} m(A) = 1 \\ m(\emptyset) = 0 \end{cases} \quad (10)$$

where $m(A)$ indicates the degree to which a subset belongs to a set A . If $m(A) > 0$, the set A is called focus element, and specially $m(\emptyset) = 0$ indicates that the truth value does not belong to the trust degree of the recognition framework Θ . The $m(\Theta)$ indicates the degree of trust that truth value belongs to Θ , but it can't determine which element it is. Given the mass function m , the corresponding belief function (Bel) and plausibility function (PL) can be obtained. Bel and PL definitions are respectively shown as Eq. (11 and 12):

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad (11)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad (12)$$

$[Bel(A), Pl(A)]$ indicate the belief interval and the degree of trust in each hypothesis. There are one-to-one correspondence among the m , Bel and Pl. For $\forall A \subseteq \Theta$ and n numbers of $m(m_1, m_2, \dots, m_n)$ on Θ , the D–S fusion rule is shown in Eq. (13).

$$(m_1 \oplus m_2 \oplus \dots \oplus m_n)(A) = \frac{1}{k} \sum_{A_1 \cap A_2 \cap \dots \cap A_n = A} m_1(A_1) \cdot m_2(A_2) \cdot \dots \cdot m_n(A_n) \quad (13)$$

where $A_1, A_2, \dots, A_n \subseteq \Theta$, and k is a probability of conflict, which is defined as Eq. (14).

$$\begin{aligned} k &= (m_1 \oplus m_2 \oplus \dots \oplus m_n)(\emptyset) \\ &= \sum_{A_1 \cap A_2 \cap \dots \cap A_n = \emptyset} m_1(A_1) \cdot m_2(A_2) \cdot \dots \cdot m_n(A_n) \end{aligned} \quad (14)$$

D–S evidence theory can be widely concerned and applied, which is inseparable from the advantages of data analysis. Its main advantages include the ability to deal with uncertain and incomplete information efficiently. The information of the target and the subset of the recognition framework is corresponding and transformed into a mathematical model, which is convenient for data fusion. The support of evidence can be described by the evidence interval, and the unclear problems can be well expressed. When the evidence data are fused, the target prior knowledge is not needed. With the continuous expansion of the application of D–S evidence theory, it is found that the evidence theory also has some shortcomings that cannot be ignored. For example, the problem of information source independence, the problem of calculation explosion, the method of obtaining basic probability and the inability to deal with evidence conflict [57].

3.2 Construction of BPA Matrix of Clustering Validity Function

In the process of cluster validity fusion, for the judgment of the best cluster number of the same data set under m cluster validity functions, there are m results, that is to say that there are m possible types of targets. The value of c is generally $[2, \sqrt{nmax}]$, where $nmax$ is the number of samples in the data set, so let $n = \sqrt{nmax}$, the validity function is set up for this purpose, and the BPA matrix is defined as $M_{n \times m}$, as shown in Eq. (15).

$$M = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1m} \\ m_{21} & m_{22} & \dots & m_{2m} \\ \vdots & \vdots & \dots & \vdots \\ m_{n1} & m_{n2} & \dots & m_{nm} \end{bmatrix} \quad (15)$$

where m_{11} is the confidence degree of the first cluster validity function to $c = 2$, when $n = 1$. And so on, m_{nm} is the confidence degree of the m -th cluster validity function for $c = n$ when $n = \sqrt{nmax}$. The sum of the confidence degree of the identification result of the number of clusters by the same cluster validity function should be 1, that is $\sum_{i=2}^n m_{ij} = 1$, where $j = 1, 2, \dots, m...$

Different cluster validity functions vary greatly in the range of judging the number of clusters. Therefore, the cluster validity functions should be normalized and standardized to make the interval fall into $[0, 1]$, which is realized by Eq. (16).

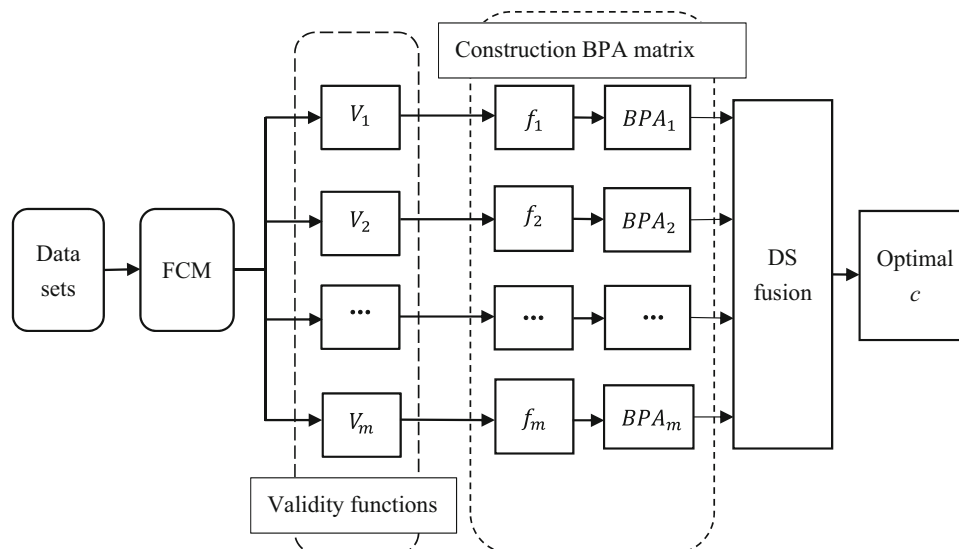


Fig. 1 D–S evidence theory-based clustering validity function fusion model (DS-CVFFM)

$$f_i = \frac{F_i - F_i^{\min}(x)}{F_i^{\max}(x) - F_i^{\min}(x)}. \quad (16)$$

where F_i is the value of validity function, $F_i^{\min}(x)$ is the minimum value of the validity function, $F_i^{\max}(x)$ is the maximum value of the validity function, f_i is the value of validity function after normalization.

however, directly inputting the value of the normalized validity function into the confidence matrix cannot satisfy the $\sum_{i=1}^n m_{ij} = 1$, and clustering validity functions are maximum efficiency and some are minimum efficiency. When the cluster validity function takes the maximum value of c , the BPA function is $m_{ij} = \frac{f_i}{\sum f_i}$. When the cluster validity function takes the minimum value of c as efficient, the minimum value of c should be transformed into the maximum validity, so the BPA function is $m_{ij} = \frac{1-f_i}{\sum 1-f_i}$. After the transformation, the sum of confidence degree results is 1.

A new matrix $A_{m \times m}$ can be obtained by multiplying one row of M with another row, and the definition of $A_{m \times m}$ is shown in Eq. (17).

$$A = M_i^T \times M_j = \begin{bmatrix} m_{i1} \times m_{j1} & m_{i1} \times m_{j2} & \dots & m_{i1} \times m_{jm} \\ m_{i2} \times m_{j1} & m_{i2} \times m_{j2} & \dots & m_{i2} \times m_{jm} \\ \vdots & \vdots & \vdots & \vdots \\ m_{im} \times m_{j1} & m_{im} \times m_{j2} & \dots & m_{im} \times m_{jm} \end{bmatrix} \quad (17)$$

where the main diagonal element is the cumulative confidence of the i -th and j -th clustering validity function for target recognition with the best number of clusters, which

defined as $Z = m_{ip} \times m_{jq} (p = q)$, the sum of non-principal diagonals constitutes the normalization factor k of evidence, which is defined as Eq. (18).

$$k = \sum_{p \neq q} m_{ip} \times m_{jq}, q = (1, 2, 3, \dots, m) \quad (18)$$

3.3 Clustering Validity Function Ensemble Model and Algorithm Flowchart

As shown in Fig. 1, DS-CVFFM is a fusion model of clustering validity functions proposed in this paper. In the research process of clustering validity, the clustering validity functions are independent of each other, and they have their own advantages and disadvantages for different types of datasets. In order to apply the clustering validity method to more datasets as much as possible, DS-CVFFM regards each clustering validity function as an independent evidence. Firstly, the clustering data set are input, and the value of clustering validity function is obtained by FCM clustering algorithm. Then the values of cluster validity functions are normalized, and the BPA is constructed by the results of cluster validity functions. Finally, the information from different cluster validity functions is fused using D–S evidence theory and the optimal number of clusters c is given using decision module. On the basis of FCM clustering algorithm, considering the advantages of adapting to different data, the fusion strategy of multiple clustering validity functions can better adapt to different types of data.

When designing the decision module for DS-CVFFM, set $A_i (i = 1, 2, \dots, n)$ is the number of clusters, A_ω is the target type, and after get the $m_{ij}(\theta)$ about A_i that in Θ , the design of decision module should comply with the

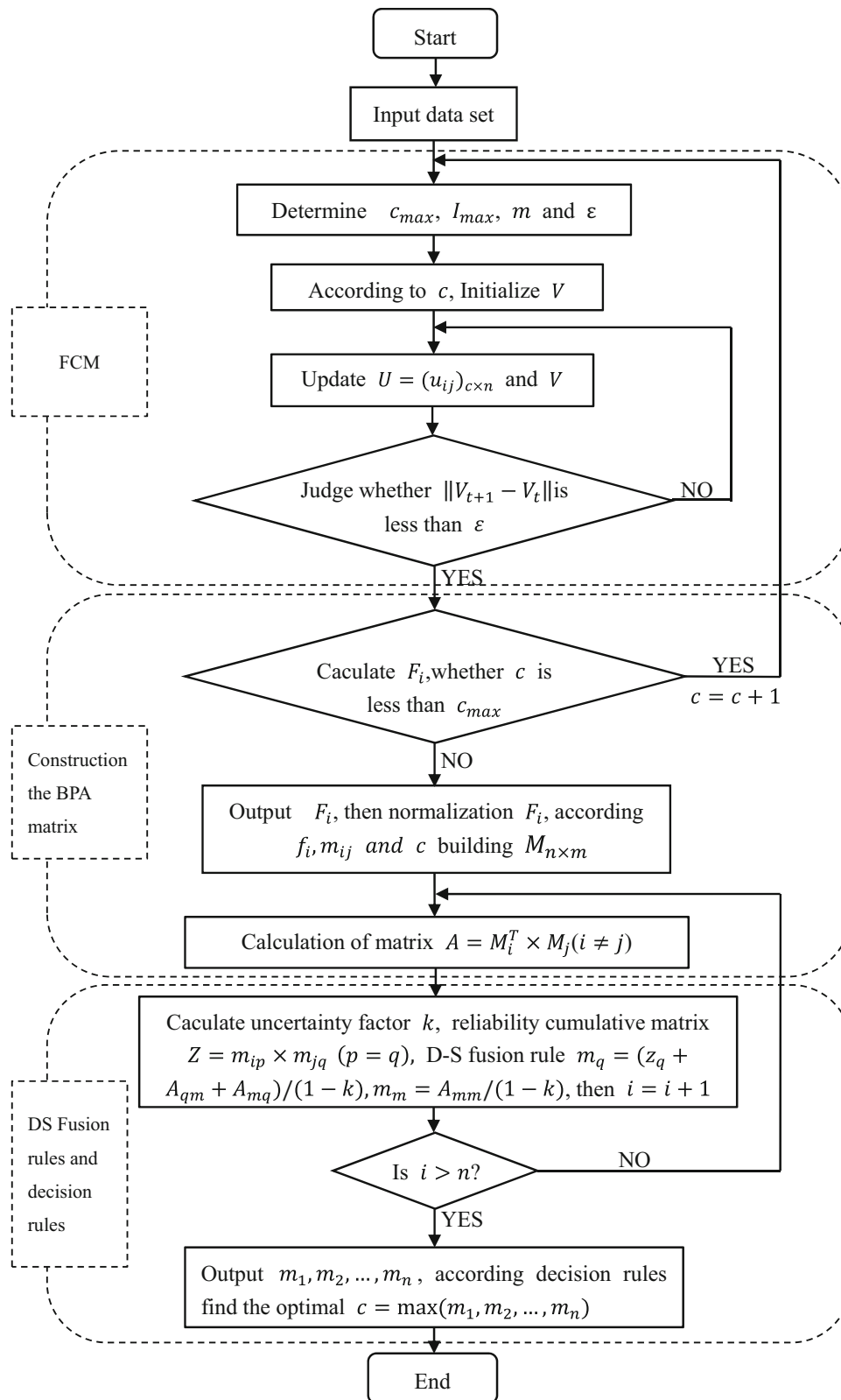
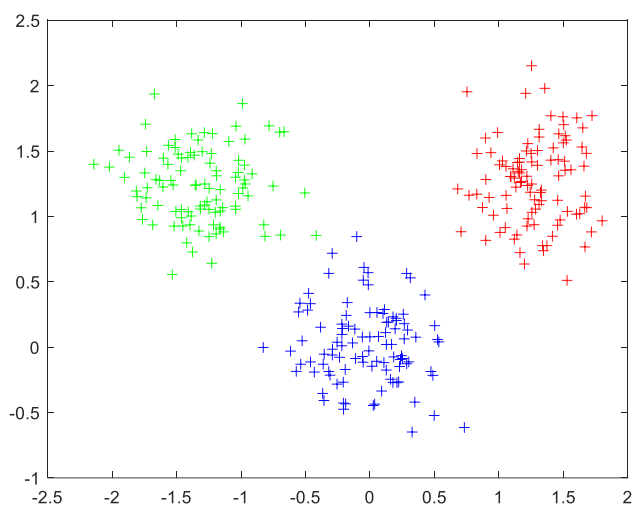
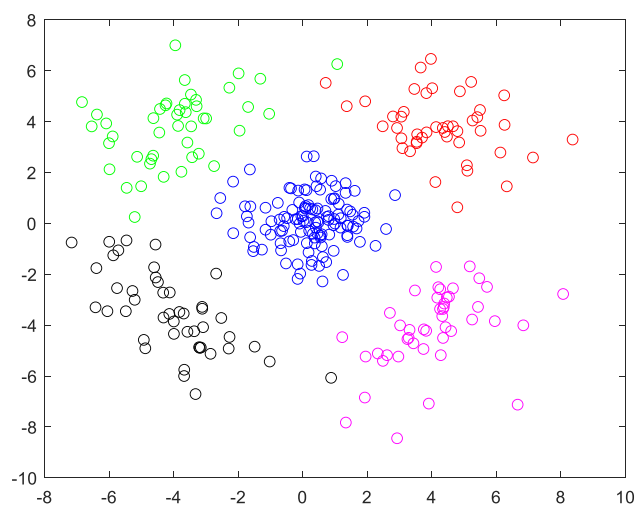


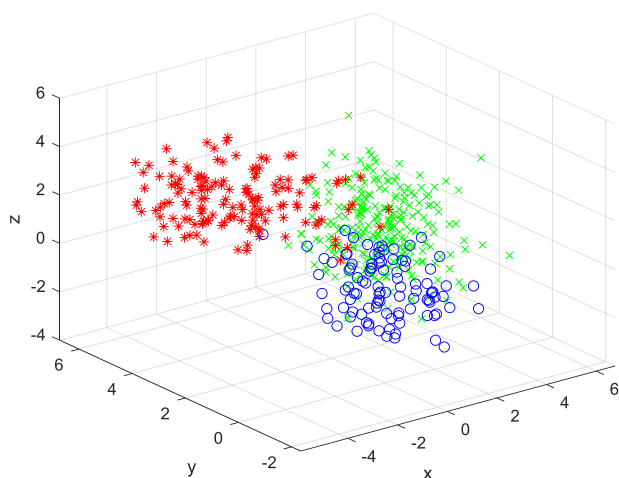
Fig. 2 Algorithm flowchart of DS-CVFFM



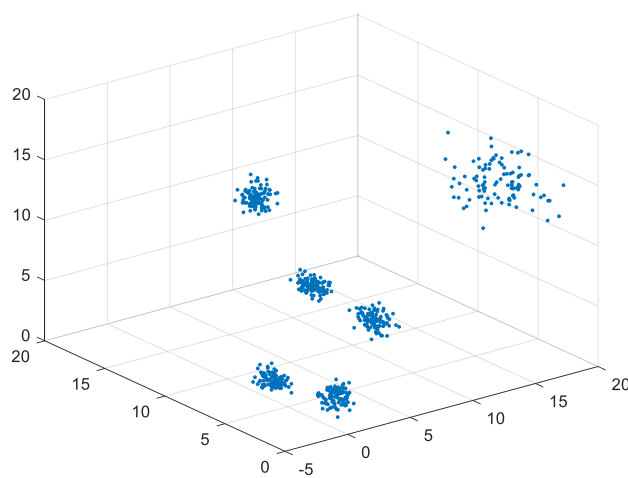
(a) data_2_3



(b) data_2_5



(c) data_3_3



(d) data_3_6

Fig. 3 Scatter plot of artificial datasets

following rules: (1) $m(A_\omega) = \max(m(A_i))$, that is, the target type (the best number of clusters) with maximum confidence. (2) $m(A_\omega) - m(A_i) > \varepsilon$, ($\varepsilon > 0$), that is, the reliability difference between the target class and other classes must be greater than a set threshold. The flowchart of DS-CVFFM is shown in Fig. 2.

It can be seen from Fig. 2 that DS-CVFFM uses FCM algorithm as the base algorithm to judge the clustering center V and membership matrix U , and then calculates multiple fuzzy clustering validity functions F_i under different clustering number c value conditions. By calculating different F_i , in this case, F_i has judged the validity of FCM clustering results, but the value of the best cluster number is different. Therefore, BPA function is constructed and

different F_i is normalized so that the optimal number of clusters corresponding to F_i is the maximum effective, that is, the maximum probability of BPA function corresponds to the optimal c . At this time, DS-CVFFM has judged the probability of the optimal c for F_i under different number of clusters. However, when judging the optimal c , it is often based on the change trend and peak value of the validity function.

Sometimes the local minimum or maximum value corresponds to the optimal c , so that the next highest probability value corresponds to the optimal c . Therefore, the D–S evidence theory is used to fuse the probabilities of different F_i under different cluster numbers. In this way, even if the individual validity function is second maximum

Table 3 Information introduction of UCI data set

Datasets	Instances	Attributes	Classes	Source
iris	150	4	3	http://archive.ics.uci.edu/ml/datasets/iris
Seeds	210	7	3	http://archive.ics.uci.edu/ml/datasets/seeds
Heart	270	13	2	http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
Ionosphere	351	34	2	http://archive.ics.uci.edu/ml/datasets/Ionosphere
wdbc	198	34	2	http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
Balance	625	4	3	http://archive.ics.uci.edu/ml/support/balance+scale
bupa	345	6	2	http://archive.ics.uci.edu/ml/datasets/Liver+Disorders
Breast	286	9	2	http://archive.ics.uci.edu/ml/datasets/Breast+Cancer
hfcv	299	13	4	http://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records#
German	1000	20	2	http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
Satimage	6435	36	7	http://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29
Led7	500	7	10	http://archive.ics.uci.edu/ml/datasets/LED+Display+Domain
Segment	2310	19	7	http://archive.ics.uci.edu/ml/datasets/Image+Segmentation
Pageblocks	5473	10	5	http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification

Table 4 Feasibility of DS-CVFFM under different datasets

<i>c</i>	data_2_3	data_2_5	data_3_3	data_3_6	iris	seeds
2	0.019609	0.004253	0.10901	0.006952	0.085826	0.061925
3	0.679854	0.04754	0.473281	0.091551	0.404787	0.58405
4	0.045133	0.211176	0.18756	0.17827	0.121695	0.224173
5	0.067632	0.39839	0.136499	0.181614	0.115372	0.070289
6	0.042748	0.049924	0.05367	0.267893	0.11509	0.018198
7	0.034925	0.043026	0.018572	0.072134	0.026748	0.007646
8	0.03592	0.047484	0.011697	0.131049	0.034771	0.002865
9	0.034293	0.025086	0.005653	0.024006	0.048921	0.016006
10	0.01926	0.069768	0.000924	0.029795	0.031844	0.010854
11	0.006622	0.04409	0.002287	0.00623	0.011494	0.002533
12	0.013626	0.059262	0.000786	0.009299	0.002889	0.00139
13	0.000379	0	6.04E-05	0.001206	0.000563	7.21E-05
14	0	0	0	0	0	0
Conclusion	Right	Right	Right	Right	Right	Right

Bold font indicates the validity function value corresponding to the optimal cluster number

validity, it will be determine the optimal number of clusters because other validity functions are modified at the maximum validity. In the Sect. 4, the paper will carry out simulation experiments on DS-CVFFM to prove the feasibility of the scheme.

4 Experimental Simulation and Result Analysis

4.1 Selection of Datasets and Experimental Parameters

In order to verify the effectiveness of DS-CVFFM, this paper selects 4 sets of artificial datasets and 14 sets of UCI

datasets for experiments. The number of clusters selected in the experiment is $2 \leq c \leq 14$, and the value of fuzzy weighted m is different according to different designs. The threshold of D-S decision module is $\varepsilon = 0.1$. The experimental environment is Windows 10 operating system and the experimental platform is Matlab 2018a. The artificial datasets are shown in Fig. 3a–d, where data_2_3 is a set of 2-D and 3-class data set which obey Uniform distribution, data_2_5 is a set of 2-D and 5-class data set which obey Gaussian distribution, data_3_3 is a set of 3-D and 3-class data set which obey Gaussian distribution, and data_3_6 is a set of 3-D and 6-class data set with Uniform distribution. UCI datasets are shown in Table 3. The first list shows the name of the data set, the second list shows the number of

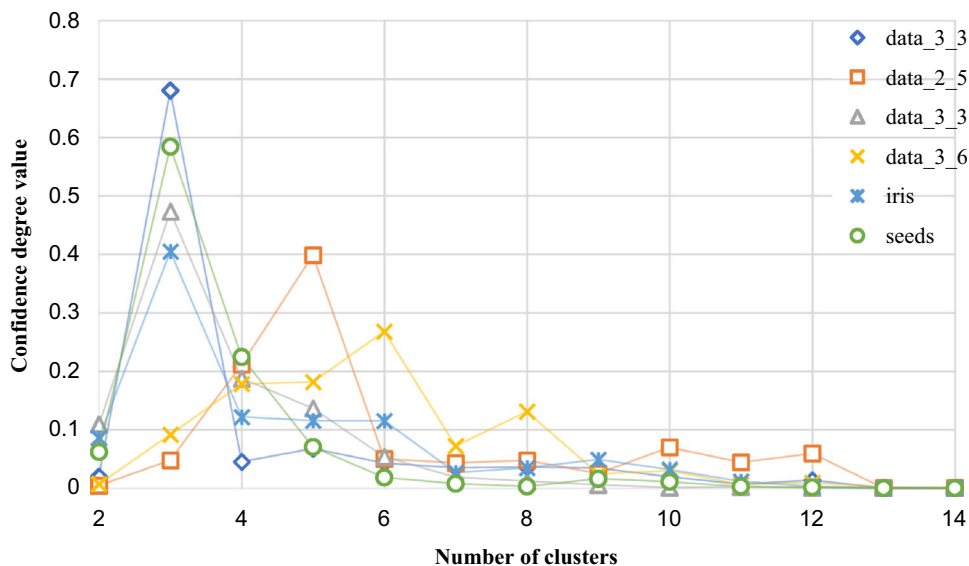


Fig. 4 Feasibility of DS-CVFFM verified by different datasets

Table 5 Confidence values of c for DS-CVFFM under different datasets and $V = 4$

c	iris	seeds	heart	balance	german	breast	bupa	sonar
2	0.118051	0.09895	0.571641	0.569742	0.519773	0.522242	0.494763	0.429013
3	0.445434	0.524318	0.23172	0.218145	0.235701	0.222788	0.229007	0.373859
4	0.15599	0.22895	0.086798	0.129179	0.1031	0.154714	0.15615	0.034793
5	0.165534	0.088572	0.036622	0.043722	0.086778	0.051521	0.06899	0.070132
6	0.067119	0.026893	0.056513	0.018868	0.0352	0.030379	0.025825	0.051773
7	0.019101	0.013852	0.006658	0.0154	0.012365	0.009591	0.013026	0.024536
8	0.017506	0.009156	0.002398	0.001879	0.004327	0.005216	0.008133	0.007085
9	0.006593	0.005552	0.000647	0.002566	0.001756	0.00194	0.002261	0.005219
10	0.002875	0.002271	0.005278	0.000136	0.000634	0.000721	0.00122	0.002339
11	0.001325	0.001162	0.00054	3.51E-04	0.000333	0.000651	0.000513	0.000925
12	0.000472	0.00032	0.001166	1.38E-05	3.19E-05	0.000235	0.000107	0.000319
13	1.23E-06	3.47E-06	1.93E-05	0.00E + 00	8.36E-08	2.34E-06	4.51E-06	5.30E-06
14	0	0	0	0	0	0	0	0
Conclusion	Right	Right	Right	False	Right	Right	Right	False

Bold font indicates the validity function value corresponding to the optimal cluster number

samples of the data set, the third list shows the attributes of the data set, the fourth list shows the number of classification of the data, and the fifth list shows the source of the data set.

4.2 Verify the feasibility of DS-CVFFM

In this section, 4 artificial datasets and 2 UCI datasets (iris and seeds) are selected to test the feasibility of DS-CVFFM. The structure and attributes of artificial data set are relatively simple, and it is easy to identify the best number of clusters. Iris and seeds datasets are the most frequently used datasets to verify the effectiveness of

clustering. Four clustering validity functions ($V_{MPC}, V_P, V_Z, V_{HY}$) are selected to fusion. The fuzzy weighted $m = 2.0$ is selected, and the simulation results are shown in Table 4 and Fig. 4.

It can be found in Table 4 that the confidence degree is the highest for data_2_3 when $c = 3$, the confidence degree is the highest for data_2_3 when $c = 5$, the confidence degree is the highest for data_3_6 when $c = 6$, and the confidence degree is the highest for iris and seeds when $c = 3$. The confidence of the six selected datasets is the highest when the number of clusters is the optimal, the results show that DS-CVFFM is feasible and can find the correct number of clusters. In order to intuitively see the

Table 6 Confidence values of c for DS-CVFFM under different datasets and $V = 6$

c	Iris	Seeds	Heart	Balance	German	Breast	bupa	Sonar
2	0.112914	0.055718	0.492762	0.225519	0.480163	0.312745	0.197208	0.609621
3	0.735986	0.61812	0.267742	0.355748	0.211172	0.342495	0.253516	0.067375
4	0.005908	0.243411	0.114723	0.2497	0.052527	0.160215	0.398797	0.093264
5	0.030801	0.048586	0.054873	0.13038	0.194641	0.083613	0.119198	0.086497
6	0.043273	0.025496	0.026331	0.014908	0.052471	0.05129	0.017226	0.081266
7	0.03938	0.005072	0.011969	0.011993	0.007242	0.026787	0.008924	0.0405
8	0.019045	0.00196	0.014186	0.007659	0.000801	0.013902	0.004723	0.006918
9	0.008986	0.000679	0.007559	0.002138	0.00045	0.005413	0.00025	0.008743
10	0.002998	0.000775	0.007669	0.001138	0.000185	0.002256	6.36E-05	0.003847
11	0.000689	0.000154	0.001859	0.000555	0.000313	0.00107	3.64E-05	0.001556
12	2.14E-05	2.99E-05	0.000302	0.000263	3.37E-05	0.000213	5.74E-05	0.00041
13	0	0	2.39E-05	0	6.93E-08	1.64E-06	8.17E-08	1.98E-06
14	0	0	0	0	0	0	0	0
Conclusion	Right	Right	Right	Right	Right	False	False	Right

Bold font indicates the validity function value corresponding to the optimal cluster number

Table 7 Confidence values of c for DS-CVFFM under different datasets and $V = 8$

c	Iris	Seeds	Heart	Balance	German	Breast	bupa	Sonar
2	0.102221	0.07317	0.547626	0.2137	0.573361	0.44406	0.242389	0.399016
3	0.707233	0.710988	0.278987	0.521442	0.168691	0.35104	0.251651	0.199162
4	0.020025	0.181549	0.100799	0.138887	0.052028	0.13741	0.377052	0.260894
5	0.160836	0.023132	0.041663	0.064831	0.162643	0.01921	0.114704	0.110647
6	3.46E-05	0.008472	0.017139	0.03555	0.038409	0.02804	0.005195	0.024567
7	0.000941	0.001921	0.001544	0.019151	0.004098	0.01186	0.006233	0.002366
8	0.007419	0.000278	0.001973	3.98E-03	0.00051	0.00568	0.002713	0.002306
9	0.001244	0.000134	0.004431	0.001792	1.41E-05	0.00209	4.80E-05	0.000718
10	2.08E-05	0.000315	0.003353	0.000432	0.000195	0.00046	9.56E-06	0.000211
11	3.73E-06	3.71E-05	0.001981	0.000173	4.69E-05	6.84E-05	5.19E-06	7.92E-05
12	2.22E-05	3.32E-06	0.000502	5.92E-05	3.40E-06	6.97E-05	1.07E-06	3.36E-05
13	0	8.17E-10	3.54E-07	5.06E-08	4.57E-09	3.96E-08	1.24E-10	0
14	0	0	0	0	0	0	0	0
Conclusion	Right	Right	Right	Right	Right	Right	False	Right

Bold font indicates the validity function value corresponding to the optimal cluster number

confidence change of each data set to c under the judgment of DS-CVFFM, the data in Table 4 is plotted as a scatter diagram, as shown in Fig. 4.

4.3 Influence of Different Number of Clustering Validity Functions on DS-CVFFM

DS-CVFFM can obtain the final number of clusters by fusion multiple cluster validity functions. Therefore, the selection of the number of cluster validity functions will affect the final clustering results. For this reason, this section selects 4, 6, 8, 10 cluster validity functions for experiments to observe the influence on the final clustering

results when selecting different number of cluster validity functions. Because the structure of artificial datasets are simple, only 8 sets of UCI datasets are used in this experiment, and the fuzzy weighted $m = 2$. The experimental results are shown in Tables 5, 6, 7, and 8.

Table 5 shows that when the number of cluster validity functions is 4 ($V_{PC}, V_{PE}, V_{MPC}, V_P$), the confidence result of c . The results show that DS-CVFFM cannot judge the correct number of clusters for balance and sonar when $V = 4$. For sonar, although the maximum confidence value is found at $c = 2$ which is in line with the selection of the best number of clusters, the difference between the confidence value and the maximum confidence value at $c = 3$ is

Table 8 Confidence values of c for DS-CVFFM under different datasets and $V = 10$

c	iris	seeds	heart	balance	german	breast	bupa	sonar
2	0.125069	0.083453	0.55746	0.07793	0.73571	0.234987	0.259001	0.154502
3	0.713212	0.733263	0.265808	0.578955	0.11986	0.146249	0.254292	0.309436
4	0.06989	0.141669	0.094624	0.160863	0.003477	0.388979	0.354511	0.165942
5	0.079517	0.030642	0.019203	0.109562	0.122668	0.076552	0.11497	0.040994
6	0.004986	0.008805	0.010865	0.041234	0.017823	0.08892	0.008195	0.007835
7	0.00037	0.001573	0.003475	0.027199	0.000401	0.04851	0.006349	0.309342
8	0.006559	0.000236	0.005911	0.002779	2.27E-06	0.002611	0.002518	0.000984
9	0.000142	0.000313	0.014853	0.000695	2.61E-05	0.012365	0.000141	0.00053
10	0.00017	8.07E-07	0.001452	0.000686	2.55E-05	7.06E-05	2.14E-05	0.000357
11	8.19E-05	4.07E-05	0.006811	6.40E-05	5.26E-06	0.000732	1.18E-06	0.002367
12	3.06E-06	3.97E-06	0.019539	3.26E-05	1.40E-06	2.47E-05	1.01E-06	0.00771
13	0	0	1.69E-07	0.00E + 00	0	2.87E-09	2.14E-09	2.46E-08
14	0	0	0	0	0	0	0	0
Conclusion	Right	Right	Right	Right	Right	False	False	False

Bold font indicates the validity function value corresponding to the optimal cluster number

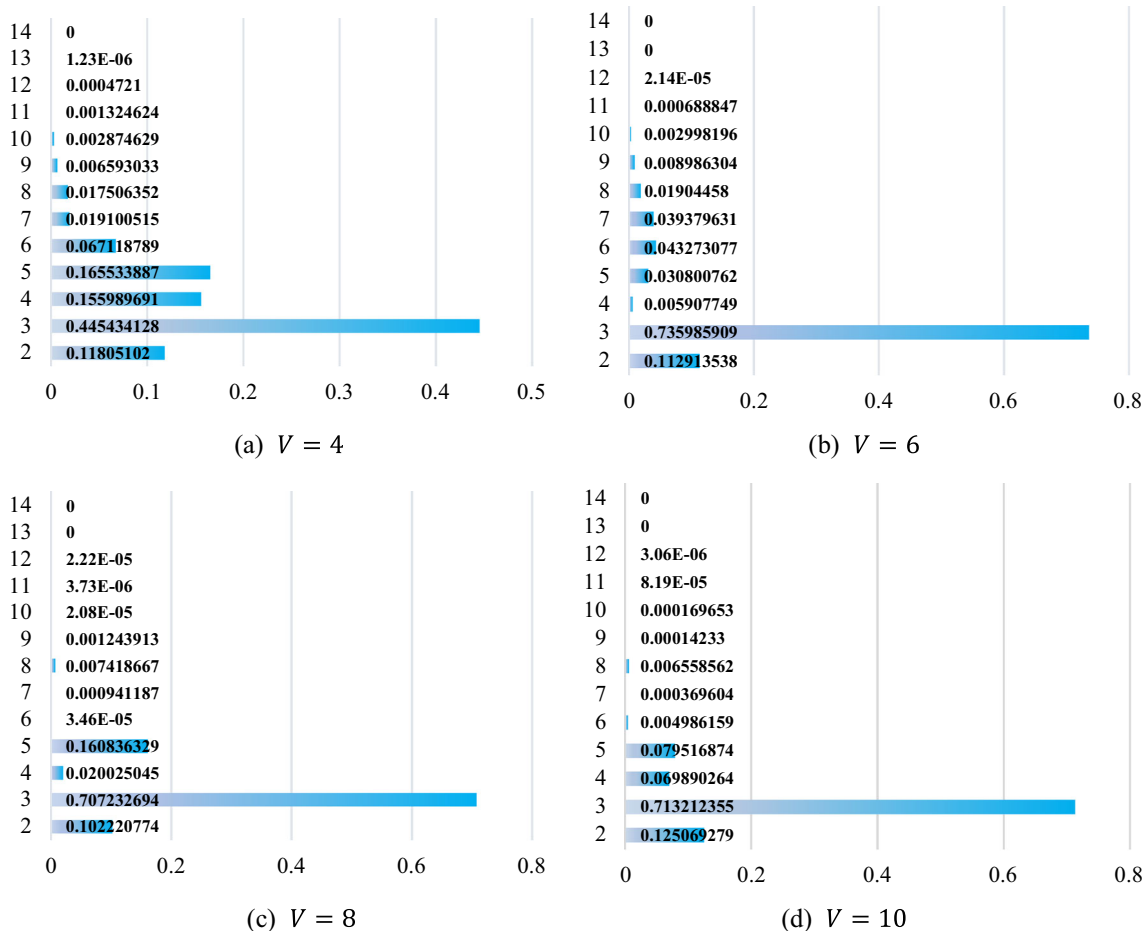


Fig. 5 Confidence values of c with the increase of the number of validity functions for iris data set

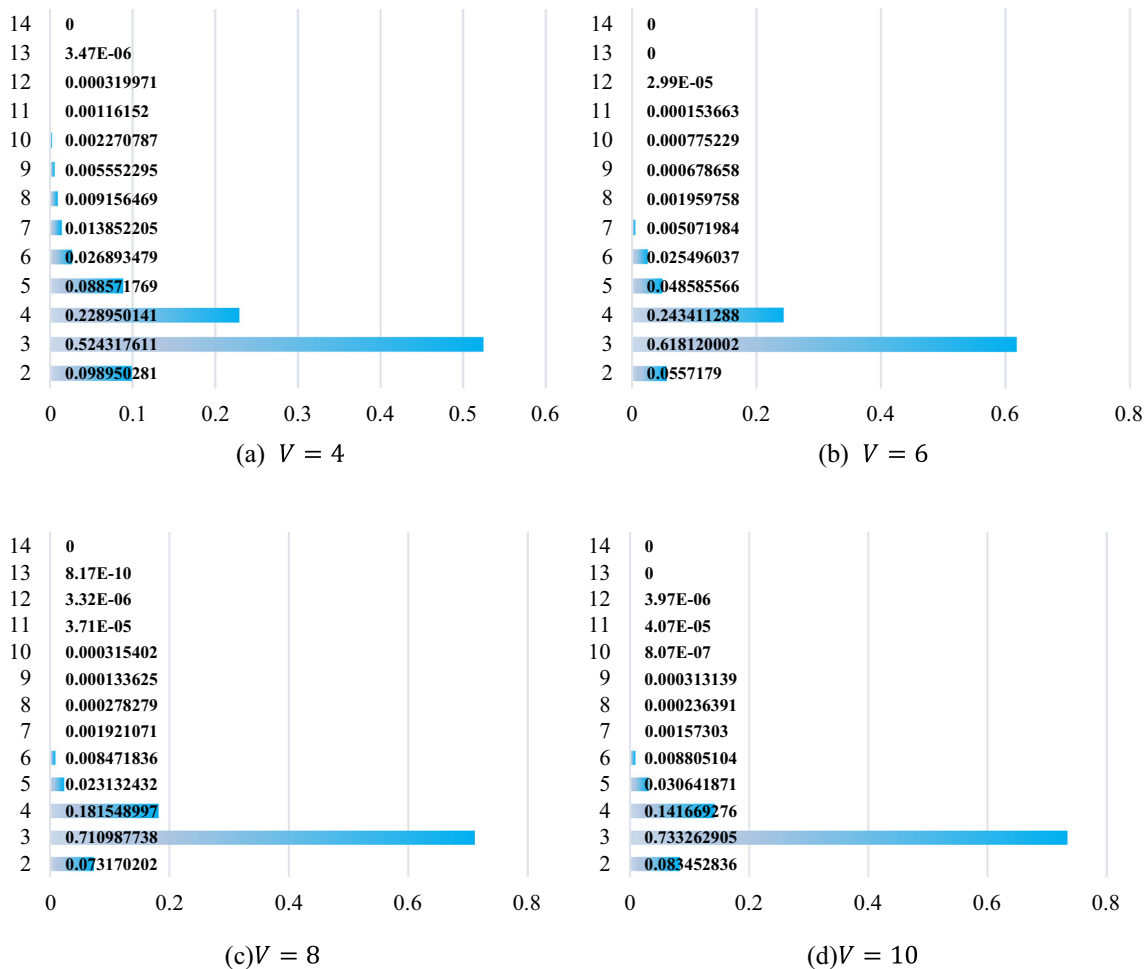


Fig. 6 Confidence values of c with the increase of the number of validity functions for seeds data set

less than the threshold value, so the correct number of clusters cannot be output. Therefore, under the condition of $V = 4$, DS-CVFFM can find the optimal number of clusters for 6 datasets. Table 6 shows that when the number of cluster validity functions is 6 ($V_{PC}, V_{PE}, V_{MPC}, V_P, V_{XB}, V_{FS}$), the confidence result of c . The results show that DS-CVFFM cannot judge the correct number of clusters for bupa and breast under the condition of $V = 6$. Therefore, under the condition of $V = 6$, DS-CVFFM can find the optimal number of clusters for 6 datasets. Table 7 shows that when the number of cluster validity functions is 8 ($V_{PC}, V_{PE}, V_{MPC}, V_P, V_{XB}, V_{FS}, V_Z, V_{HY}$), the confidence result of c . The results show that DS-CVFFM cannot judge the correct number of clusters for bupa only under the condition of $V = 8$, so DS-CVFFM can find the best number of clusters for 7 datasets under the condition of $V = 8$. Table 8 shows that when the number of cluster validity functions is 10 ($V_{PC}, V_{PE}, V_{MPC}, V_P, V_{XB}, V_{FS}, V_Z, V_{HY}, V_{WL}, V_{PCAES}$), the

confidence result of c . The results show that DS-CVFFM cannot judge the correct number of clusters for breast, bupa and sonar under the condition of $V = 10$, so DS-CVFFM can find the best number of clusters for 5 datasets under the condition of $V = 10$.

According to the experimental results listed in Tables 5, 6, 7, and 8, it can be found that DS-CVFFM has the highest accuracy in judging the optimal number of clusters for different datasets under the condition of $V = 8$, with only one error, and the lowest accuracy under the condition of $V = 10$. Under the condition of $V = 4$ and $V = 6$, there are two errors. Compared with $V = 8$, the variety of the correct rate of judging the optimal number of clusters is not very significant. In order to distinguish the influence of $V = 4, V = 6$ and $V = 8$ on DS-CVFFM, the number of V is determined by judging the difference between the maximum and the second largest value of c . when the difference between the maximum and the second largest value is larger, the value of c corresponding to the maximum confidence of c can be regarded as the optimal number of

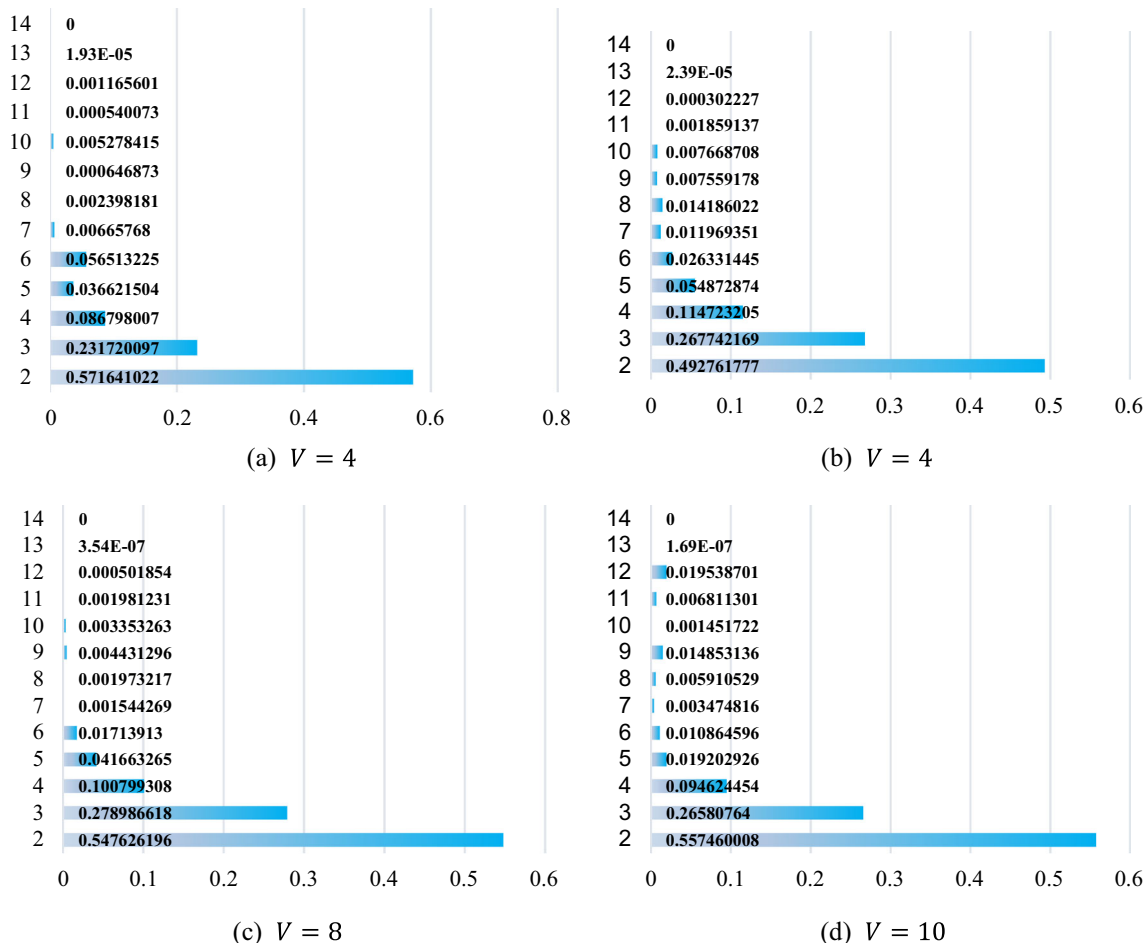


Fig. 7 Confidence values of c with the increase of the number of validity functions for heart data set

clusters. For this reason, 4 experimental datasets (iris, seeds, heart, german) which can still obtain the optimal number of clusters after fusing the clustering validity functions with different numbers are selected from the above four groups of experiments for comparison, as shown in Figs. 5, 6, 7, and 8a–d.

As shown in Fig. 5a–d, the selected experimental data set is iris. When $V = 4$, the difference between the maximum and the second largest value of c confidence is 0.2799 (approximate to four decimal places). Similarly, when $V = 6$, the difference between the maximum and the second largest value of c confidence is 0.6231, $V = 8$ is 0.5664, and $V = 10$ is 0.5881. When the selected data set is seeds, the experimental results are shown in Fig. 6a–d. When V is 4, 6, 8 and 10, respectively, the corresponding differences are 0.2944, 0.3747, 0.5295 and 0.5916. When the selected data set is heart, the experimental results are shown in Fig. 7a–d. When V is 4, 6, 8 and 10, respectively, the corresponding differences are 0.3404, 0.2251, 0.2696 and 0.2917. When the selected data set is german, the experimental results are shown in Fig. 8a–d. When V is 4,

6, 8 and 10, the corresponding differences are 0.2814, 0.2690, 0.4108 and 0.6130. The difference summary is shown in Table 9. When $V = 10$, there are three datasets with the largest difference, and $V = 6$ has 1 data set with the largest difference. However, when $V = 8$, although there is no maximum difference, the difference is second only to $V = 10$. According to the comprehensive experimental results in Tables 5, 6, 7, 8, and 9, $V = 8$ is the most appropriate number of clustering effectiveness functions.

4.4 Influence of Fuzzy Weighted on DS-CVFFM

The selection of fuzzy weighted m has a certain impact on the final clustering results. In the past research process, m is usually taken as $m = 2$, but it is not suitable for all datasets and clustering algorithms. In order to improve the ability of DS-CVFFM to find the optimal number of clusters, select multiple m for the same group of cluster validity functions, and then fusion the values of the final validity functions to make the final results more reliable. Therefore, different values of m are changed to observe the fusion results of

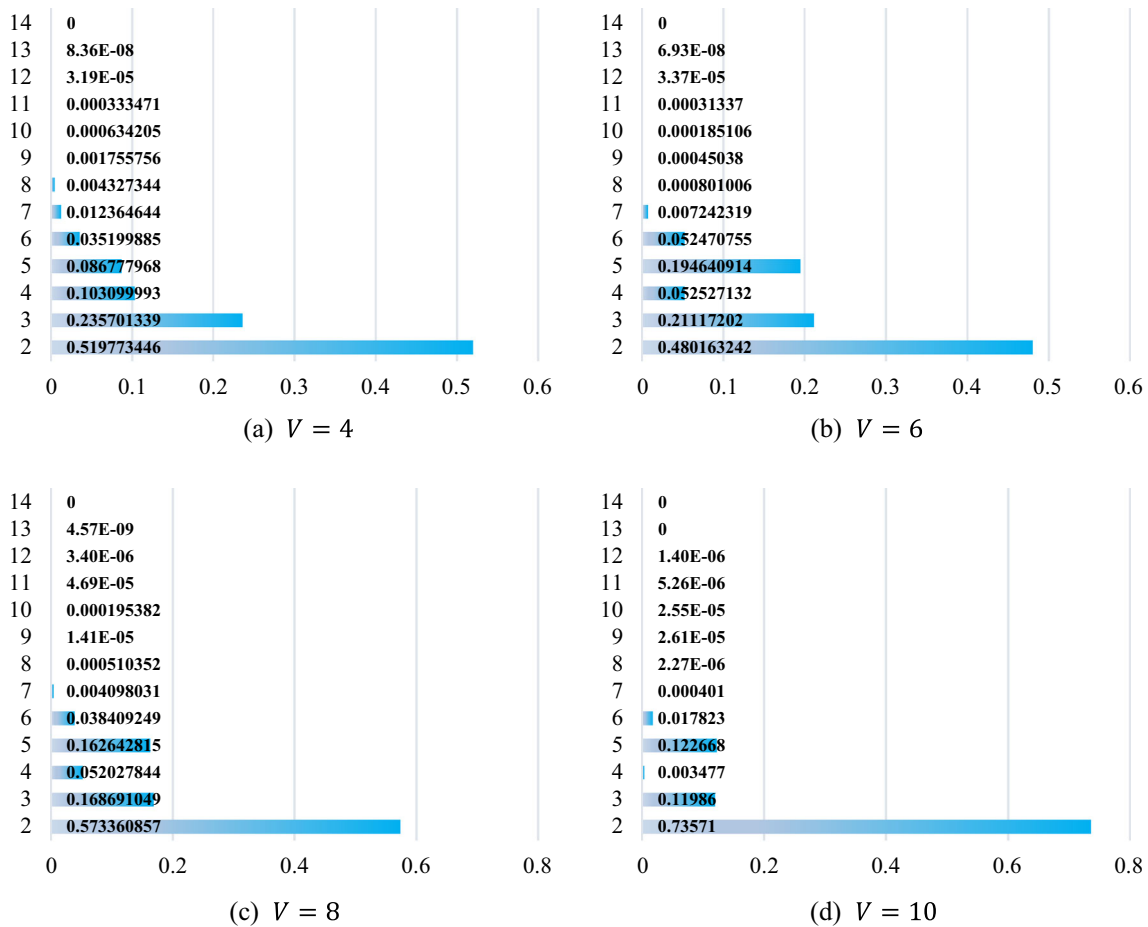


Fig. 8 Confidence values of c with the increase of the number of validity functions for german data set

Table 9 Difference between maximum value and second maximum value of the confidence for different c when selecting different numbers of cluster validity functions

	iris	seeds	heart	german
$V = 4$	0.2799	0.2944	0.3404	0.2814
$V = 6$	0.6231	0.3747	0.2251	0.2690
$V = 8$	0.5664	0.5295	0.2696	0.4108
$V = 10$	0.5881	0.5916	0.2917	0.6130

DS-CVFFM on different datasets. The experimental results are shown in Tables 10, 11, and 12.

As shown in Table 10, When $m = 2.0$, set the fusion period $T = 1$, and select the number of clustering validity functions $V = 8$. DS-CVFFM cannot get the correct number of clusters for haberman and sonar datasets and the correct number of clusters can be obtained for the other 6 datasets. When $m = 1.5, 2.0$ and 3.0 , as shown in Table 11, set the fusion period $T = 3$, and select the number of clustering validity functions $V = 8$. DS-CVFFM cannot get

the correct number of clusters for haberman, blast, bupa and sonar datasets and the correct number of clusters can be obtained for the other 4 datasets. When $m = 1.5, 1.75, 2.0, 2.5$ and 3.0 , as shown in Table 12, set the fusion period $T = 5$ and select the number of clustering validity functions $V = 8$. DS-CVFFM can get the correct number of clusters for 8 datasets.

According to Tables 10, 11, and 12, it can be found that when $m = 2.0$ is indeed the most appropriate value of fuzzy weighted. When $m = 1.5, m = 3.0$, there will be some interference to the final fusion results, resulting in some data cannot find the correct number of clusters. However, when we continue to expand the value condition of m , it can find that taking $m = 1.75$ and $m = 2.5$ in the left and right fluctuation of $m = 2$ can improve the accuracy of the final clustering results. And when $m = 1.5, 1.75, 2.0, 2.5$ and 3.0 , DS-CVFFM gets the highest number of datasets. In order to better observe the change of confidence of DS-CVFFM under different m values, Tables 10, 11, and 12 can also be shown in the form of Figs. 9, 10, and 11.

Table 10 Confidence values of c for DS-CVFFM under different datasets and $m = 2.0$

c	Haberman	wdbc	Balance	German	Ionosphere	Breast	bupa	Sonar
2	0.05225	0.589444	0.07793	0.73571	0.55966	0.234987	0.259001	0.154502
3	0.116626	0.251063	0.578955	0.11986	0.274062	0.146249	0.254292	0.309436
4	0.506713	0.109726	0.160863	0.003477	0.08923	0.388979	0.354511	0.165942
5	0.175709	0.028946	0.109562	0.122668	0.04393	0.076552	0.11497	0.040994
6	0.111474	0.014355	0.041234	0.017823	0.018519	0.08892	0.008195	0.007835
7	0.034224	0.0045	0.027199	0.000401	0.003257	0.04851	0.006349	0.309342
8	1.67E-04	0.001372	0.002779	2.27E-06	0.007729	0.002611	0.002518	0.000984
9	1.83E-03	0.000419	0.000695	2.61E-05	0.002646	0.012365	0.000141	0.00053
10	0.000759	0.000145	0.000686	2.55E-05	0.000837	7.06E-05	2.14E-05	0.000357
11	2.51E-04	2.82E-05	6.40E-05	5.26E-06	8.91E-05	0.000732	1.18E-06	0.002367
12	1.28E-06	2.71E-06	3.26E-05	1.40E-06	4.15E-05	2.47E-05	1.01E-06	0.00771
13	1.45E-09	0	0.00E + 00	0	6.30E-09	2.87E-09	2.14E-09	2.46E-08
14	0	0	0	0	0	0	0	0
Conclusion	False	Right	Right	Right	Right	Right	Right	False

Table 11 Confidence values of c for DS-CVFFM under different datasets and $m = 1.5, 2.0$ and 3.0

c	haberman	wdbc	balance	german	ionosphere	breast	bupa	sonar
2	0.188451	0.705052	0.277337	0.924333	0.674215	0.25727	0.146694	0.001555
3	0.050535	0.260343	0.629261	0.067085	0.295807	0.486345	0.799251	0.006106
4	0.729413	0.033566	0.082375	2.67E-05	0.022039	0.253178	0.046615	0.000659
5	0.026825	0.000877	0.010416	0.008469	0.007642	0.002534	0.00744	9.12E-06
6	0.004718	0.000151	0.000579	8.66E-05	0.000269	0.00058	7.02E-08	1.64E-10
7	3.26E-05	1.06E-05	1.09E-05	1.91E-07	1.93E-05	8.13E-06	4.25E-08	0.991671
8	2.63E-05	7.09E-07	2.11E-05	1.48E-09	6.27E-06	8.29E-05	6.98E-08	1.51E-08
9	2.21E-09	3.24E-08	4.74E-07	1.29E-10	2.75E-06	8.75E-07	5.05E-09	1.05E-13
10	8.29E-08	4.94E-09	1.28E-06	2.44E-12	1.40E-07	2.50E-07	1.68E-12	9.72E-14
11	9.89E-11	1.41E-10	8.13E-10	1.94E-12	2.41E-09	1.51E-09	1.34E-13	6.44E-13
12	1.83E-10	2.33E-11	2.72E-11	8.41E-15	1.32E-11	4.04E-09	1.87E-18	4.92E-16
13	1.43E-20	0	0	0	0	7.44E-22	0	0
14	0	0	0	0	0	0	0	0
Conclusion	False	Right	Right	Right	Right	False	False	False

4.5 Comparison of DS-CVFFM and Traditional Clustering Validity Evaluation Methods

In order to verify that the result of DS-CVFFM is more accurate than that of clustering validity functions and combined validity evaluation methods in judging the optimal number of clusters of data set, DS-CVFFM is selected to compare with clustering validity functions (V_{MPC} , V_{XB} , V_K , V_P , V_{WL} , V_Z , V_{HY}) and combined clustering validity evaluation methods (DWSVF, FWSVF, WSCVI, HWCVF). The comparative experimental results

of DS-CVFFM and clustering validity functions are shown in Tables 13 and 14, and the comparative experimental results of DS-CVFFM and combined clustering validity evaluation methods are shown in Figs. 12 and 13.

Select DS-CVFFM and V_{MPC} , V_{XB} , V_K , V_P , V_{WL} , V_Z , V_{HY} was compared in 14 UCI datasets, and the experimental results are shown in Tables 13 and 14. Table 13 shows the experiment of selecting the data set with the best number of clusters $c < 5$, and Table 14 shows the experiment of selecting the data set with the best number of clusters $5 < c < 10$. The table only intercepts the value of

Table 12 Confidence values of c for DS-CVFFM under different datasets and $m = 1.5, 1.75, 2.0, 2.5$ and 3.0

c	Haberman	wdbc	Balance	German	Ionosphere	Breast	bupa	Sonar
2	0.64333	0.812528	0.055149	0.977629	0.902627	0.122977	0.87066	0.911098
3	0.000434	0.166291	0.931496	0.02134	0.096661	0.750553	0.105597	0.053815
4	0.248068	0.020616	0.012826	5.56E-07	0.000655	0.122287	0.023212	0.030756
5	0.106904	6.13E-06	0.000525	0.00103	3.68E-05	0.003907	0.000531	0.00012
6	0.001263	0.000558	4.39E-06	6.58E-07	1.77E-06	9.58E-05	3.51E-08	1.06E-09
7	6.50E-09	3.21E-07	3.72E-07	1.43E-11	1.36E-05	0.000179	2.11E-13	0.004211
8	5.87E-10	2.41E-09	6.54E-09	1.69E-14	4.94E-06	4.99E-08	5.99E-14	3.79E-11
9	1.39E-11	1.61E-11	2.37E-11	3.32E-18	6.43E-12	1.01E-08	8.41E-18	2.33E-14
10	2.91E-14	1.29E-13	3.78E-12	4.00E-19	2.24E-09	5.53E-11	1.85E-20	1.58E-16
11	3.61E-18	4.00E-13	3.63E-14	2.43E-20	2.05E-17	1.95E-15	2.80E-30	1.92E-17
12	2.24E-19	6.08E-21	3.55E-18	1.95E-27	4.60E-23	3.86E-17	8.09E-27	5.14E-22
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
Conclusion	Right	Right	Right	Right	Right	Right	Right	Right

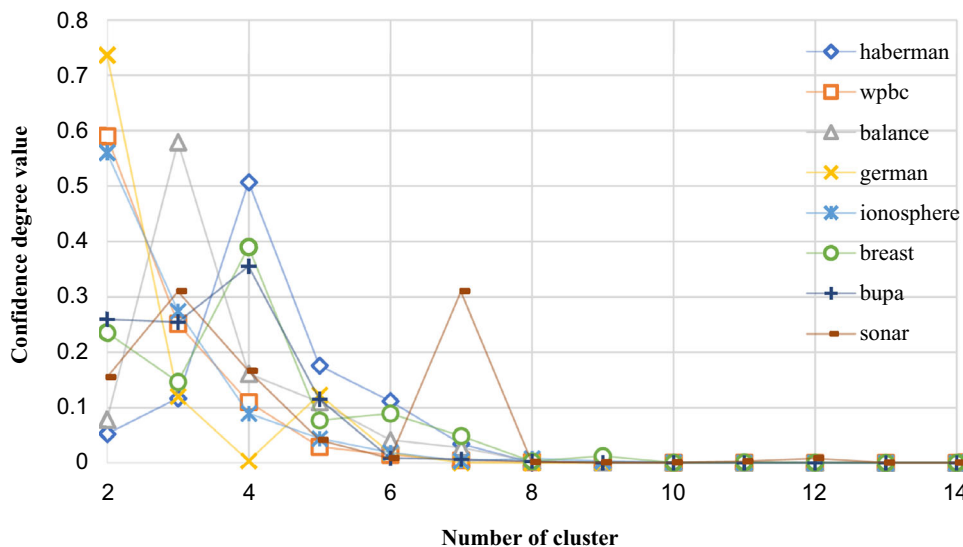


Fig. 9 Confidence values of c for DS-CVFFM under different datasets and $m = 2.0$

c , and reserves the data results that provide comparison for the experiment. As shown in Table 13 (a–j), for iris data set, DS-CVFFM and V_{HY} can judge the optimal number of clusters. For seeds data set, DS-CVFFM, V_{MPC} , V_P and V_{HY} can judge the optimal number of clusters. For heart, ionosphere, wdbc, bupa 4 data set, only V_{MPC} cannot judge the optimal number of clusters. For balance data set, DS-CVFFM and V_{MPC} can judge the optimal number of clusters. For sonar data set, DS-CVFFM, V_{XB} , V_K , V_Z and V_{HY} can judge the optimal number of clusters. For hfcr data set, DS-CVFFM, V_{XB} and V_Z can judge the optimal number of

clusters. As shown in Table 14 (a–d), satimage and Led7 datasets can be found the correct clustering number by DS-CVFFM, V_{WL} and V_{HY} . For segment dataset V_{XB} , V_K , V_{WL} and DS-CVFFM can judge the correct number of clusters. For the pageblocks dataset, V_{XB} , V_K , V_Z and DS-CVFFM can judge the correct number of clusters. According to the experimental results in Tables 13 and 14. The results show that DS-CVFFM can find the correct number of clusters for all 14 datasets, V_{HY} and V_{XB} can find 8 datasets, V_Z , V_K and V_{WL} can find 7 datasets, V_P can find 5 datasets, V_{MPC} can find 2 datasets. Thus, the performance of DS-CVFFM is

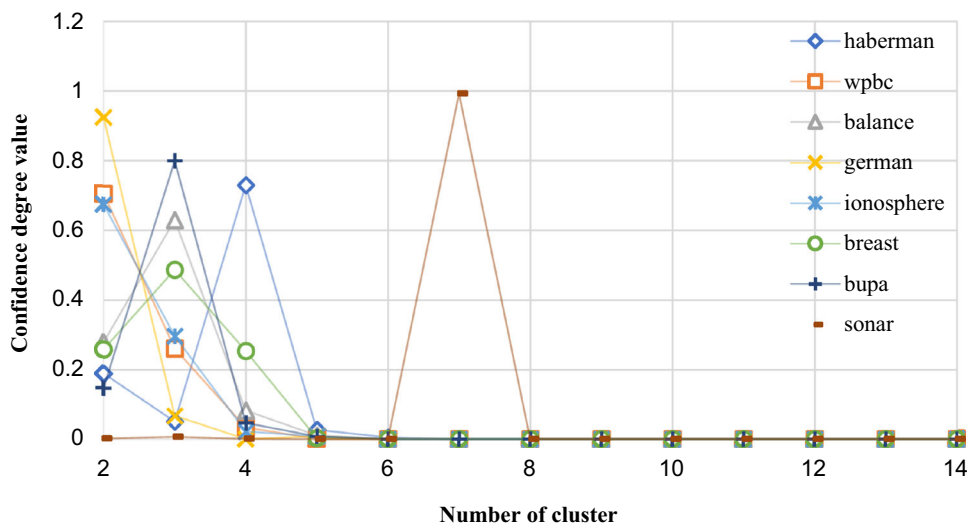


Fig. 10 Confidence values of c for DS-CVFFM under different datasets and $m = 1.5, 2.0$ and 3.0

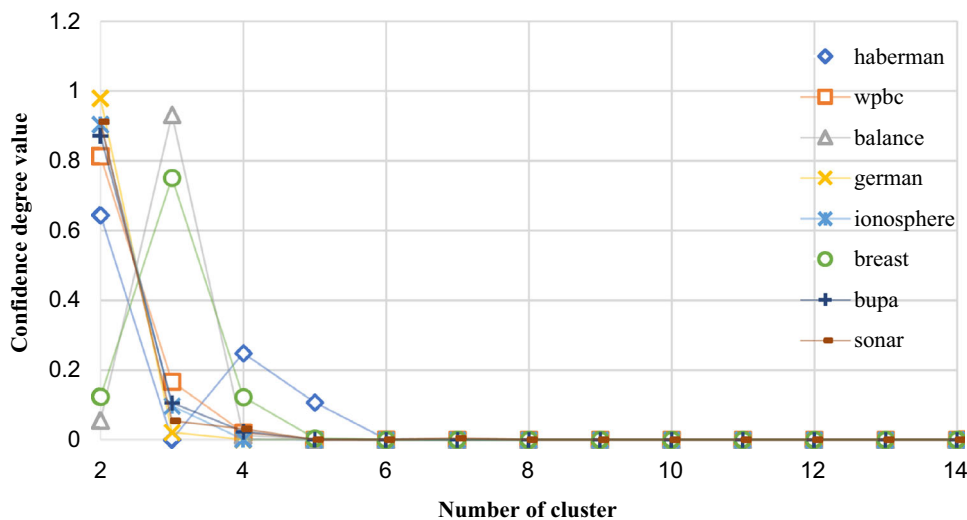


Fig. 11 Confidence values of c for DS-CVFFM under different datasets and $m = 1.5, 1.75, 2.0, 2.5$ and 3.0

significantly better than that of single clustering validity function.

DS-CVFFM is selected to compare with DWSVF, FWSCF, WSCVI, HWCVF in 15 UCI datasets, and the results are shown in Figs. 12 and 13. Figure 12 shows the experiment of selecting the data set with the best number of clusters $c < 5$, and Fig. 13 shows the experiment of selecting the data set with the best number of clusters $5 < c < 10$. The abscissa is the number of clusters, and the ordinate is the value of each method. For iris, seeds and balance data set, only FWSCF cannot find the correct

optimal number of clusters. For heart, ionosphere and sonar datasets, only DWSVF cannot find the correct optimal number of clusters. For wpbc and bupa datasets, DWSVF and WSCVI cannot find the correct number of clusters. For hfer dataset, DWSVF and FWSCF cannot find the correct optimal number of clusters. For the german dataset, all methods can determine the optimal number of clusters. As shown in Fig. 13, only DS-CVFFM can find the correct optimal cluster number for the data set of satimage and led7. DWSVF, WSCVI and DS-CVFFM can judge the optimal clustering number for pageblocks dataset.

Table 13 Comparison between DS-CVFFM and single cluster validity function, when the selected data set cluster is $c < 5$

Datasets	c	MPC	XB	VK	VP	VWL	VZ	VHY	DS-CVFFM
Iris	2	- 1.41358	0.056458	8.593358	0.858337	0.052668	0.004407	2.037941	0.009093
	3	-1.05798	0.126879	19.60584	0.81187	0.090402	0.016634	1.593598	0.989377
	4	- 1.09369	0.269549	41.92238	0.751594	0.14561	0.044469	2.452269	0.00153
	5	- 1.09734	0.41106	65.54202	0.720105	0.107521	0.066233	2.611343	8.00E-08
Seeds	2	- 1.643508	0.102115	21.52627	0.734575	0.102528	0.0035714	7.551749	0.001205
	3	-1.249509	0.151415	32.02605	0.739776	0.166908	0.005858	6.885022	0.998587
	4	- 1.254923	0.164044	34.77424	0.692544	0.211557	0.0069858	10.62651	0.000207
	5	- 1.301568	0.264412	56.26118	0.654337	0.205666	0.0120681	11.39837	1.30E-07
Heart	2	- 2.25676	1.05217	319.427	0.31311	1.05276	0.52444	2.57002	0.930425
	3	-2.0374	163.8319	49.786.29	0.195901	1.713673	91.19391	3.80866	0.043352
	4	- 2.07519	4067.82	1,236,999	0.148596	3.10733	2369.782	5.309402	0.009787
	5	- 2.15117	3022.462	919,689.4	0.132415	17.45175	1775.245	6.857363	0.001279
Ionosphere	2	- 2.05485	0.73807	259.264	0.46948	0.76549	0.04329	17.7678	0.674215
	3	-1.83559	282,275.9	99,185,326	0.333412	1.473044	19,968.06	30.93041	0.295807
	4	- 1.8655	12.28319	4317.294	0.316001	3.570514	0.853357	40.50212	0.022039
	5	- 1.93885	6.36E + 09	2.24E + 12	0.271125	3.398134	4.62E + 08	52.06155	0.007642
wpbc	2	- 2.30723	2.00464	397.936	0.24316	2.00584	1.56284	1.88817	0.705052
	3	-2.09046	17.41653	3460.809	0.155392	4.42135	14.56786	2.784114	0.260343
	4	- 2.12829	295.4459	58,791.09	0.117321	6.619864	250.7506	3.777422	0.033566
	5	- 2.21278	2952.327	588,225	0.088789	4.970808	2552.701	4.683525	0.000877
bupa	2	- 1.576106	0.12612	43.5184	0.77606	0.18561	1.46E-05	43.3033	0.494763
	3	- 1.432179	0.589929	203.576	0.632274	0.264043	0.0001	200.5302	0.229007
	4	-1.39465	0.661272	228.2321	0.601669	0.311896	0.000115	225.5318	0.15615
	5	- 1.503332	0.488328	168.5624	0.545977	0.3976	9.62E-05	166.9205	0.06899
Balance	2	- 2.385692	273.8465	171,158.2	0.019745	273.8465	3.36E + 01	15.99203	0.055149
	3	-2.147072	249.9714	156,239.3	0.018687	266.9815	30.550503	23.99005	0.931496
	4	- 2.179824	244.271	152,680	0.018208	252.0499	29.779576	31.9939	0.012826
	5	- 2.258389	225.5505	140,981.4	0.016551	289.5852	2.75E + 01	40.01161	0.000525
Sonar	2	0.680532	0	0	0.829182	7.98E-05	0	0	0.053815
	3	1	1	1	1	0	1	0.082477	0.911098
	4	0.953789	7.82E-05	7.82E-05	0.625697	7.50E-05	7.93E-05	0.166112	0.030756
	5	0.850442	0.000701	0.000702	0.467349	0.012787	0.0007149	0.24943	0.00012
HFCR	2	- 1.59518	0.152988	45.74348	0.758688	0.170223	5.35E-12	45.74347	0.009408
	3	- 1.09735	0.16226	48.51561	0.784963	0.207819	5.71E-12	48.51561	0.004624
	4	- 0.93998	0.110531	33.04893	0.804485	0.175338	3.67E-12	33.04891	0.359076
	5	- 0.86632	0.214297	64.0749	0.807848	0.250148	7.37E-12	64.07487	0.012239
	6	- 0.79959	0.148822	44.49774	0.818496	0.119479	5.40E-12	44.49769	0.064041
	7	- 0.77483	0.139861	41.81854	0.816947	0.139168	5.36E-12	41.81847	0.207208
	8	- 0.76987	0.270992	81.02684	0.819283	0.23492	8.34E-12	81.02675	0.195032
	9	- 0.72467	0.094827	28.35322	0.823217	0.112558	4.18E-12	28.35314	0.098257
	10	-0.7195	0.110981	33.18348	0.823372	0.143777	5.22E-12	33.18337	0.050115
	11	- 0.7753	0.26122	78.10506	0.814801	0.128178	1.32E-11	78.10475	6.34E-14
	12	- 0.74477	0.359942	107.6232	0.817206	0.147883	1.84E-11	107.6227	4.50E-10

Table 14 Comparison between DS-CVFFM and single cluster validity function, when the selected data set cluster is $5 < c < 10$

Datasets	c	MPC	XB	VK	VP	VWL	VZ	VHY	DS-CVFFM
Satimiage	2	- 2.23695	1.298768	1930.182	0.283356	1.308576	10.70647	0.431999	0.000727
	3	- 2.02133	12.71356	18,910.63	0.223545	2.204941	112.9902	0.249214	0.064173
	4	-1.96662	6.404124	9544.446	0.322485	1.494781	50.39918	0.295718	0.040935
	5	- 2.02838	28.44829	42,465.52	0.248927	1.257219	225.9657	0.328284	0.101419
	6	- 2.12299	47.33898	70,743.22	0.215836	1.199033	388.8797	0.359071	0.074272
	7	- 2.22178	188.965	282,605.8	0.18313	1.166976	1604.076	0.171883	0.718276
	8	- 2.31655	183.1075	274,007.8	0.159437	5.89938	1596.255	0.509584	0.000199
	9	- 2.40437	996.2908	1,491,857	0.142911	5.425206	8842.979	0.594398	1.72E-07
	10	- 2.48299	3479.173	5,213,916	0.135062	6.67682	31,303.76	0.668938	2.43E-07
	Led7	2	- 2.29307	1.474364	737.8931	0.255255	1.477769	0.752648	2.961281
3		- 2.02304	8.191963	4103.64	0.230195	1.522349	4.141228	3.852237	0.00945
4		- 2.0597	109,978.2	55,120,209	0.142804	2.212272	61,220.3	5.145121	0.17031
5		-1.98764	31,967.1	16,046,002	0.2687	1.706431	14,731.3	6.710202	0.004175
6		- 2.21738	1,075,684	5.39E + 08	0.104607	42,932.42	633,950.3	8.065039	0.0022
7		- 2.15376	2,017,453	1.01E + 09	0.179595	1.924358	1,079,579	9.606212	8.22E-05
8		- 2.36186	19,434,138	9.75E + 09	0.093391	633.327	11,744,762	11.00071	0.007331
9		- 2.25836	3.09E + 08	1.55E + 11	0.181523	12.28297	1.68E + 08	12.72086	0.001034
10		- 2.32061	1.27E + 08	6.39E + 10	0.163481	2341.189	71,884,211	14.07261	0.79371
11		- 2.37533	6,184,500	3.11E + 09	0.160069	1590.15	3,539,687	15.57659	0.001034
12		- 2.27011	34,132,067	1.72E + 10	0.21244	69,608.56	18,578,239	15.42733	1.85E-04
segment		2	- 1.95535	0.359261	829.8965	0.565913	2395.608	0.380263	7.38E-06
	3	-1.54013	0.433397	1001.155	0.600328	870.2199	0.291297	8.46E-06	0.011671
	4	- 1.5688	0.56036	1294.445	0.537328	291.9352	0.48948	1.23E-05	0.052074
	5	- 1.65965	0.853424	1971.434	0.483214	118.4338	0.674654	1.99E-05	0.098813
	6	- 1.70729	0.568778	1313.897	0.478632	69.3486	0.639538	1.32E-05	0.085876
	7	- 1.68937	0.340593	786.7865	0.489912	42.5733	0.492979	7.36E-06	0.64186
	8	- 1.8046	0.504003	1164.27	0.469069	18.36192	0.794681	1.14E-05	0.000411
	9	- 1.83086	0.400353	924.836	0.472434	12.22728	0.727572	8.91E-06	0.000386
	10	- 1.8679	0.559687	1292.915	0.446245	6.844587	0.737305	1.28E-05	5.78E-05
	Pageblocks	2	- 1.85931	0.461255	1065.534	0.47292	7.778437	0.694348	1.01E-05
3		- 1.90357	0.484211	1118.566	0.469721	5.458754	0.70802	1.06E-05	7.90E-05
4		- 1.91058	0.270441	27.43066	0.614269	1.542048	0.286675	1.04E-02	0.020439
5		- 1.35218	0.145559	14.87953	0.666474	0.583084	0.159405	5.87E-03	0.953877
6		-1.12435	0.164623	16.9702	0.742137	0.123593	0.159605	6.26E-03	0.008686
7		- 1.13066	0.134844	13.96071	0.728711	0.064746	0.23902	5.09E-03	2.60E-05
8		- 1.15339	0.224578	23.48546	0.726273	0.034424	0.179744	8.83E-03	0.000169
9		- 1.15965	0.177088	18.65366	0.72846	0.025901	0.225359	6.78E-03	1.18E-06
10		- 1.25776	0.823846	86.96238	0.671386	0.015574	0.32561	3.41E-02	0.000571
11		- 1.2197	0.751351	80.44176	0.646945	0.006623	0.180493	3.31E-02	0.000411
12		- 1.13526	0.542651	58.17192	0.702824	0.007478	0.255017	2.24E-02	7.26E-06

HWCVF, WSCVI and DS-CVFFM can judge the optimal clustering number for segment data set. From Figs. 12 and 13, it can be seen that DS-CVFFM is more accurate than the weighted cluster validity evaluation method.

5 Conclusions

In this paper, a novel clustering validity evaluation method is proposed, which combines D-S evidence theory with multiple clustering validity functions. By standardizing the

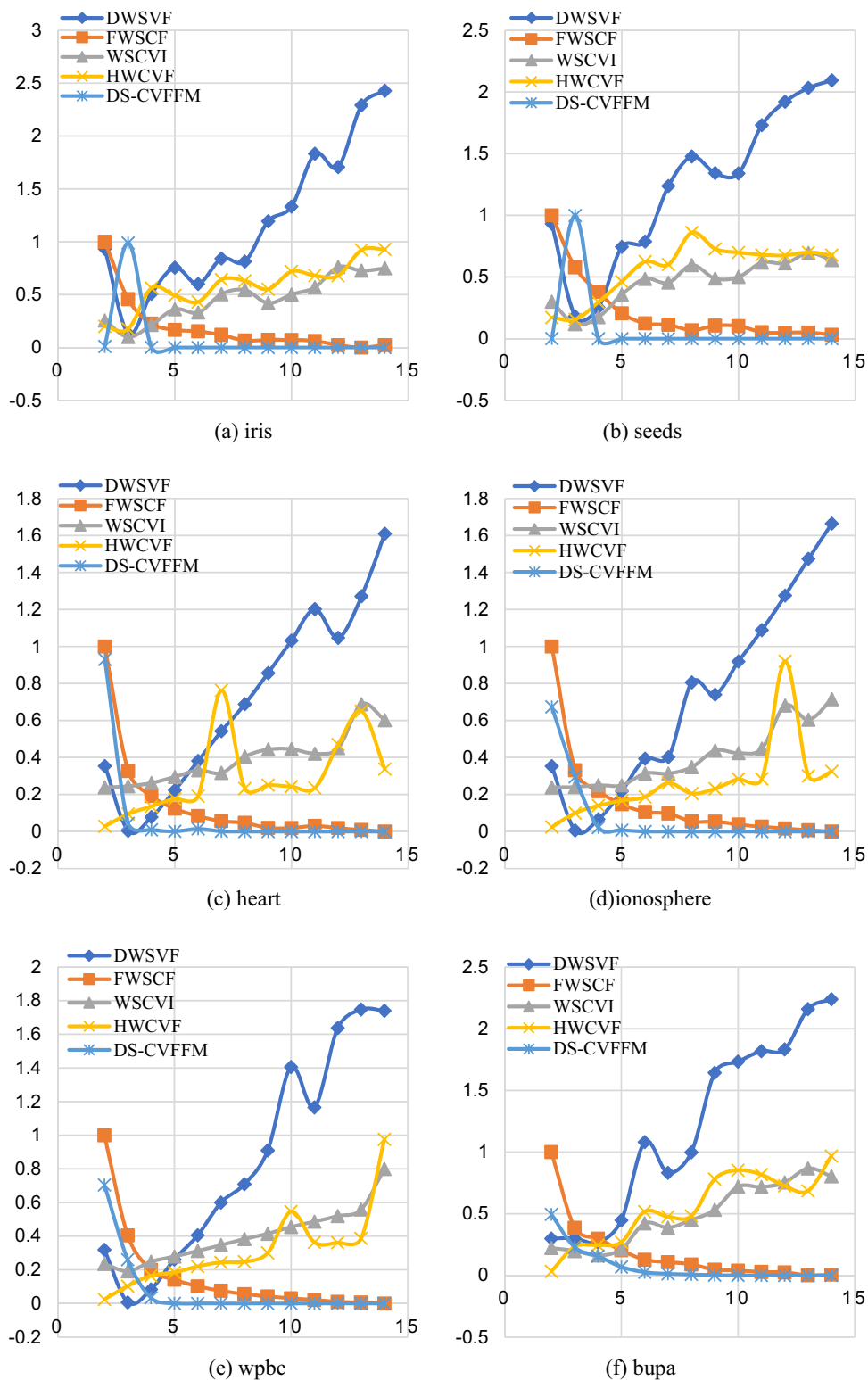


Fig. 12 Comparison between DS-CVFFM and combined validity evaluation methods, when the selected data set cluster is $c < 5$

values of clustering validity functions, the basic probability assignment function (BPA) is constructed, the integration is carried out according to the fusion rules of D–S evidence

theory, and the optimal number of clusters is output by decision module. Finally, the validity of the proposed model (DS-CVFFM) is verified by artificial data set and

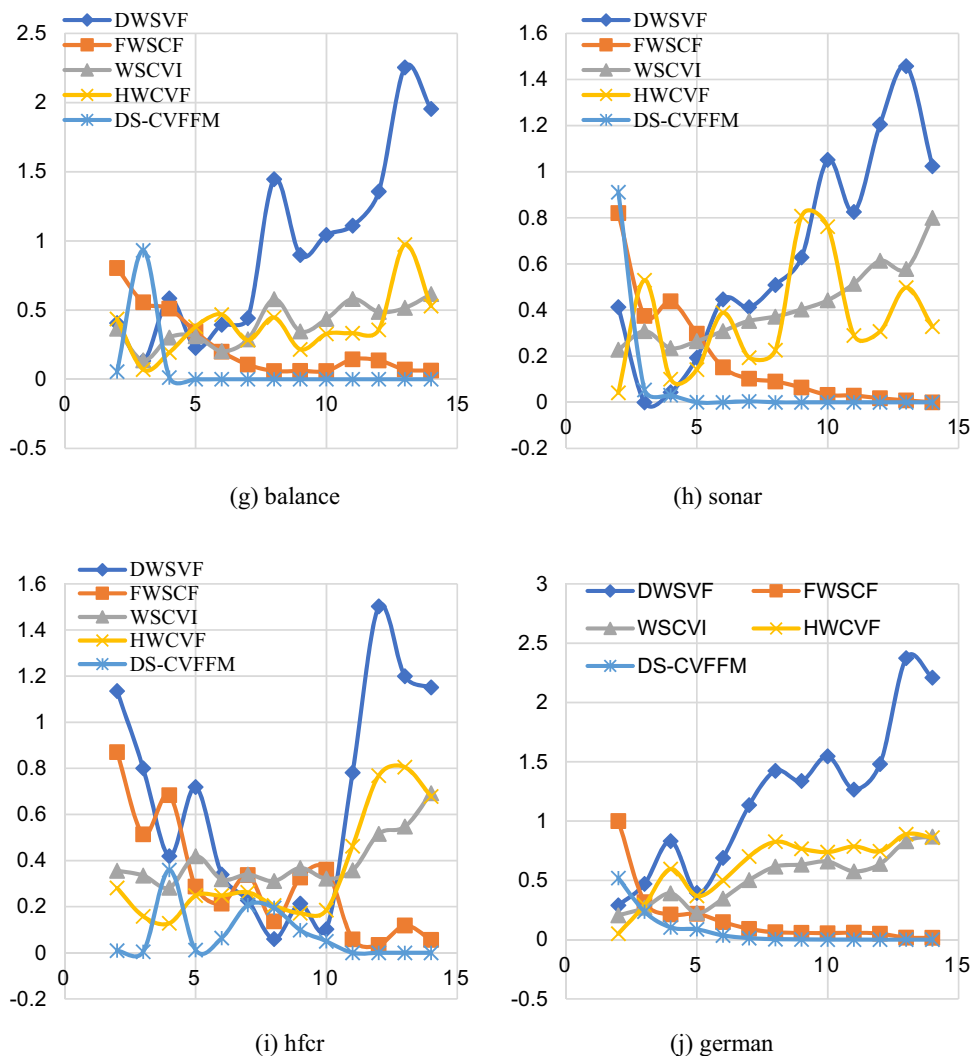


Fig. 12 continued

UCI data set. The accuracy of DS-CVFFM is judged by comparing DS-CVFFM with traditional validity evaluation methods, and then the stability of DS-CVFFM is judged by changing the fuzzy weight. The experimental results show that DS-CVFFM has better stability and accuracy than the traditional clustering validity evaluation methods under different datasets.

However, D–S evidence theory has some limitations, such as large amount of calculation, easy failure and evidence conflict in the face of independent information sources, which makes it impossible to determine the optimal number of clusters. Therefore, other optimization

techniques can be used to replace D–S evidence theory to avoid the above problems. Some clustering validity functions are only suitable for datasets with a small number of clusters. This also causes DS-CVFFM to fail to obtain the optimal number of clusters for datasets with a high number of clusters. We can preprocess the data set and then improve the accuracy of the final model. FCM algorithm also has some defects in updating the membership and clustering center. Therefore, in the future work, we will also consider using multiple FCM clustering algorithms combined with multiple validity functions to obtain the final optimal number of clusters, so as to optimize the DS-

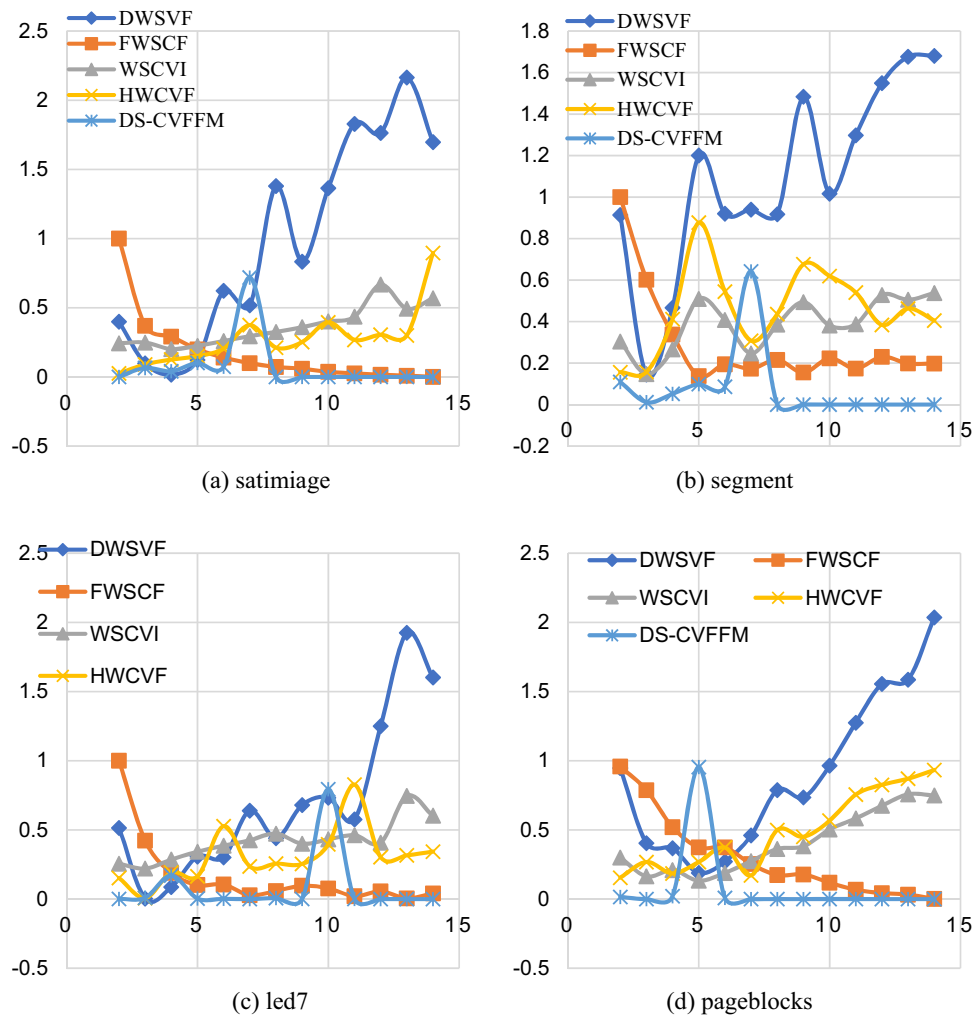


Fig. 13 Comparison between DS-CVFFM and combined validity evaluation methods, when the selected data set cluster is $5 < c < 10$

CVFFM model and improve the stability and accuracy of the model.

Acknowledgements This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. 2017FWDF10), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Author Contributions Hong-yu Wang participated in the data collection, analysis, algorithm simulation, and draft writing. Jie-Sheng Wang participated in the concept, design, interpretation and commented on the manuscript. Guan Wang 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. Frossyniotis, D., Likas, A., Stafylopatis, A.: A clustering method based on boosting. *Pattern Recogn. Lett.* **25**(6), 641–654 (2004)
2. Li, W.C., Zhou, Y., Xia, S.X.: A novel clustering algorithm based on hierarchical and K-means clustering. In: 2007 Chinese Control Conference, pp. 605–609. IEEE (2007)
3. Kriegel, H.P.: Density-based clustering. *Wiley Interdisciplinary Rev. Data Mining Knowledge Discovery* **1**(3), 231–240 (2011)
4. Gurrutxaga, I.: SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recogn.* **43**(10), 3364–3373 (2010)
5. Gertraud, M.W., Sylvia, F.S., Bettina, G.: Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26**(1–2), 303–324 (2016)
6. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10**(2–3), 191–203 (1984)
7. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C* **28**(1), 100–108 (1979)
8. Zadeh, L.A.: Fuzzy sets. In: *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: selected papers by Lotfi A Zadeh*. pp. 394–432 (1996)
9. Huang, H.: Brain image segmentation based on FCM clustering algorithm and rough set. *IEEE Access* **7**, 12386–12396 (2019)

10. Nayak, J., Naik, B., Behera, H.: Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. *Comput. Intell. Data Mining* **2**, 133–149 (2015)
11. Bezdek, J.C., Pal, N.R.: Some New indexes of cluster validity. *IEEE Trans. Syst. Man Cybernetics* **28**(Part B), 301–315 (1998)
12. Dos, S., Roberto, J.V.: Generalization of Shannon's theorem for Tsallis entropy. *J. Math. Phys.* **38**(8), 4104–4107 (1997)
13. Simovici, D.A., Jaroszewicz, S.: An axiomatization of partition entropy. *IEEE Trans. Inf. Theory* **48**(7), 2138–2142 (2002)
14. Silva, L.: An interval-based framework for fuzzy clustering applications. *IEEE Trans. Syst.* **23**(6), 2174–2187 (2015)
15. 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)
16. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(8), 841–847 (1991)
17. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of clusters for the fuzzy c-means method. In: 5th Fuzzy Systems Symposium, pp. 247–250 (1989)
18. Kwon, S.H.: Cluster validity index for fuzzy clustering. *Electron. Lett.* **34**(22), 2176–2177 (1998)
19. Wang, J.S.: A new clustering validity function for the fuzzy C-means algorithm. In: *IEEE Conference on Control & Decision.* IEEE (2008)
20. Zhu, L.F., Wang, J.S., Wang, H.Y.: A novel clustering validity function of FCM clustering algorithm. *IEEE Access* **7**, 152289–152315 (2019)
21. Sheng, W. G. A weighted sum validity function for clustering with a hybrid niching genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2005, 35.6: 1156–1167.
22. Xu, R.; Xu, J.; Wunsch, D. C. A comparison study of validity indices on swarm-intelligence-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, 42.4: 1243–1256.
23. Dong, H.B., Hou, W., Yin, 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, pp. 357–361. IEEE (2010)
24. 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, pp. 362–366. IEEE (2010)
25. Zhou, K.L.: Comparison and weighted summation type of fuzzy cluster validity indices. *Int. J. Comput. Commun. Control* **9**(3), 370–378 (2014)
26. Wang, H.Y., Wang, J.S.: Combination evaluation method of fuzzy c-mean clustering validity based on hybrid weighted strategy. *IEEE Access* **9**, 27239–27261 (2021)
27. Shafer, G.: *A Mathematical Theory of Evidence*, vol. 42. Princeton University Press, Princeton (1976)
28. Fixsen, D., Mahler, R.P.S.: The modified Dempster-Shafer approach to classification. *Syst. Man Cybernetics Part A* **27**(1), 96–104 (1997)
29. Fan, X., Guo, Y., Ju, Y., Bao, J.: Multi sensor fusion method based on the belief entropy and DS evidence theory. *J. Sens.* (2020). <https://doi.org/10.1155/2020/7917512>
30. Wu, Z.H., Wu, Z.C., Zhang, J.: An improved FCM algorithm with adaptive weights based on SA-PSO. *Neural Comput. Appl.* **28**(10), 3113–3118 (2017)
31. Lin, G.P., Qian, Y.H., Liang, J.Y.: An information fusion approach by combining multi granulation rough sets and evidence theory. *Inf. Sci.* **314**, 184–119 (2015)
32. Park, D.C., Dagher, I.: Gradient based fuzzy c-means (GBFCM) algorithm. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, pp. 1626–1631. IEEE (1994)
33. Wu, Z.D., Xie, W.X., Yu, J.P.: Fuzzy c-means clustering algorithm based on kernel method. In: *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*, pp: 49–54. IEEE (2003)
34. Sanchez, M.A.: Fuzzy granular gravitational clustering algorithm for multivariate data. *Inf. Sci.* **279**, 498–511 (2014)
35. Ding, Y., Fu, X.: Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm. *Neurocomputing* **188**, 233–238 (2016)
36. Rubio, E., Castillo, O., Valdez, F., Melin, P., Gonzalez, C.I., Martinez, G.: An extension of the fuzzy possibilistic clustering algorithm using type-2 fuzzy logic techniques. In: *Advances in Fuzzy Systems* (2017)
37. Kuo, R.J.: A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis. *Appl. Soft Comput.* **67**, 299–308 (2018)
38. Moreno, J.E., Sanchez, M.A., Mendoza, O., Rodriguez-Diaz, A., Castillo, O., Melin, P., Castro, J.R.: Design of an interval Type-2 fuzzy model with justifiable uncertainty. *Inf. Sci.* **513**, 206–221 (2020)
39. Wu, K.L., Yang, M.S.: A cluster validity index for fuzzy clustering. *Pattern Recogn. Lett.* **26**(9), 1275–1291 (2005)
40. Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P., Silbiger, M.L., Arrington, J.A.: Validity-guided (re) clustering with applications to image segmentation. *IEEE Trans. Fuzzy Syst.* **4**(2), 112–123 (1996)
41. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. *Pattern Recogn.* **37**(3), 487–501 (2004)
42. Wu, C.H.: A new fuzzy clustering validity index with a median factor for centroid-based clustering. *IEEE Trans. Fuzzy Syst.* **23**(3), 701–718 (2014)
43. Feng, Z., Fan, J.C.: A novel validity index in fuzzy clustering algorithm. *Int. J. Wirel. Mob. Comput.* **10**(2), 183–190 (2016)
44. Haouas, F.: A new efficient fuzzy cluster validity index: Application to images clustering. In: 2017 IEEE International Conference on Fuzzy Systems, pp. 1–6. IEEE (2017)
45. Geng, J.Y., Qian, X.Z., Zhou, S.B.: New fuzzy clustering validity index. *Appl. Res. Comput.* **36**(4), 1001–1005 (2019)
46. Ouchicha, C., Ammor, O., Meknassi, M.: A new validity index in overlapping clusters for medical images. *Autom. Control. Comput. Sci.* **54**(3), 238–248 (2020)
47. Liu, Y.: A new robust fuzzy clustering validity index for imbalanced datasets. *Inf. Sci.* **547**, 579–591 (2021)
48. Wang, L.S., Binet, D.: TRUST: a trigger-based automatic subjective weighting method for network selection. In: 2009 Fifth Advanced International Conference on Telecommunications, pp: 362–368. IEEE, (2009)
49. Su, X.: An new fuzzy clustering algorithm based on entropy weighting. *J. Comput. Inf. Syst.* **6**(10), 3319–3326 (2010)
50. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: *Annals of Mathematical Statistics*, Vol. 38 (1967)
51. Yager, R.R.: On the dempster-shafer framework and new combination rules. *Inf. Sci.* **41**(2), 93–137 (1987)
52. Wang, H.: A new method of cognitive signal recognition based on hybrid information entropy and DS evidence theory. *Mob. Netw. Appl.* **23**(4), 677–685 (2018)
53. Li, P., Wei, C.P.: An emergency decision-making method based on DS evidence theory for probabilistic linguistic term sets. *Int. J. Disaster Risk Reduct.* **37**, 101–178 (2019)
54. Lu, S., Li, P., Li, M.: An improved multi-modal data decision fusion method based on DS evidence theory. In: 2020 IEEE 4th

- Information Technology, Networking, Electronic and Automation Control Conference, Vol. 1, pp. 1684–1690 (2020)
55. Ye, F., Chen, J., Li, Y.B.: Improvement of DS evidence theory for multi-sensor conflicting information. *Symmetry* **9**(5), 69 (2017)
 56. Li, Y.B.: Based on DS evidence theory of information fusion improved method. In: 2010 International Conference on Computer Application and System Modeling, Vol. 1, pp. 416–419 (2020)
 57. Li, F.J.: Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method. *Inf. Sci.* **378**, 389–409 (2017)



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.



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.