Fuzzy Clustering Optimal k Selection Method Based on Multi-objective Optimization

Wang Lisong

Nanjing University of Aeronautics and Astronautics

Cui Guonan (✉ cgn843174009@qq.com)

Nanjing University of Aeronautics and Astronautics <https://orcid.org/0000-0003-4334-2273>

Cai Xinye

Nanjing University of Aeronautics and Astronautics

Research Article

Keywords: Clustering Validity index, Fuzzy Clustering, Multi-Objective Optimization Algorithm, The Number of Clustering

Posted Date: July 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-429631/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Fuzzy Clustering Optimal k Selection Method Based on Multi-objective Optimization

Lisong Wang · Guonan Cui · Xinye Cai

Received: date / Accepted: date

Abstract Because of the complexity of data sets from the real world, it is difficult to classify the data sets clearly and effectively, thus we prefer to adopt fuzzy clustering approaches to analyze the data sets. However, due to the variety of fuzzy clustering algorithms, and the different number of clusters will lead to different clustering results. The number of clusters is closely related to the clustering division, so how to determine the number of fuzzy clustering (k) has become a problem. Until now, many researchers have proposed utilizing fuzzy clustering validity indexes to deal with this kind of problem. However, the effectiveness index of fuzzy clustering can only be evaluated on the basis of the fuzzy clustering algorithm FCM to divide the clusters. When the range of k value is too large, FCM's clustering for different k values is quite time-consuming. From this perspective, this paper proposes a fuzzy clustering optimal k selection method based on multi-objective optimization (FMOEA-K). Different from the traditional methods, this method combines the fuzzy clustering effectiveness index with multi-objective optimization algorithm (MOEA), and uses multi-objective optimization algorithm to search the appropriate cluster center concurrently. Because of the concurrency of the multi-objective optimization algorithm, the calculation time is shortened. The experimental results show that compared with the traditional method, the FMOEA-K can shorten the calculation time and improve the accuracy of calculating the optimal k value.

Keywords Clustering Validity index · Fuzzy Clustering · Multi-Objective Optimization Algorithm · The Number of Clustering

Lisong Wang
Nanjing University of Aeronautics and Astronautics
Tel.: +139-51-947365
Fax: +139-51-947365
E-mail: wangls@nuaa.edu.cn

Guonan Cui
Nanjing University of Aeronautics and Astronautics
Xinye Cai
Nanjing University of Aeronautics and Astronautics

1 Introduction

As the development of the Internet technology, clustering analysis has been applied to clustering engineering including machine learning and pattern recognition[34,14], life science and medicine including genetics, biology, microbiology[22, 23,10], computer science including network mining, spatial database analysis, image segmentation[15,18,21] widely. Clustering is one kind of unsupervised learning method, which can actively group data points, such that the data points of the same cluster have extremely high similarity and with data points which belongs to different clusters have great difference. Now cluster algorithm can be divided into two types[37]. The first category is hard clustering algorithm, such as k-means, k-means++ [2], intelligent K-means[1], etc, hard clustering algorithm divides a data set into multiple clusters, and each object belongs to only one cluster. The second one we denote fuzzy cluster algorithm allowing each object to belong to multiple clusters according to the degree of membership including Fuzzy C-Means[6], MAFC[30], etc.

Determining the value of k plays an crucial role in our work, no matter hard clustering algorithm or fuzzy clustering algorithm. Because of the complexity of data sets from the real world, it is difficult to determine k without prior knowledge and understanding of data structure. Some researchers propose to utilize the clustering validity index to solve to address this contradiction[13,17]. There are two categories of clustering validity index: external validity index and internal validity index. The external validity index evaluates the partitions by comparing them with the previously assumed results[9]; the internal validity index evaluates the partition by checking the results, and generally measures the rationality of the results by the compactness (comp) and separation degree (sep). The smaller of the value $comp + sep$, the more rational of k . Through the clustering validity index, the structure of the data set can be better analyzed from the data sets, so as to obtain the optimal classification number of the data sets[31]. Compared with external validity index, the internal validity index pay more attention to the degree of compactness and separation between clusters. In the past few decades, the internal validity index of fuzzy clustering has been proposed to verify the fuzzy clustering, for instances, partition coefficient(PC)[5], partition entropy(PE)[4], XB[36], FS[16], SC[39], PBM index[27,28], VM index[40]. Besides of this, PCAES index[35], SSD[33], PBM indicators[3] and etc. However, the effectiveness index of fuzzy clustering can only be evaluated on the basis of the fuzzy clustering algorithm FCM to divide the clusters. When the range of k values is too large, it is quite time-consuming to divide clusters by FCM corresponding to different k values. On the other hand, some clustering effectiveness indicators only show a monotonically decreasing state as the number of clusters k increases only when the cluster center is close to the correct cluster center. When the cluster center is not close to the correct cluster center, they cannot determine the optimal number of clusters k .

This paper presents an algorithm for clustering optimal k selection based on Multi-Objective Optimization. Different from the traditional methods, this method combines the fuzzy clustering effectiveness index with multi-objective optimization algorithm, and uses multi-objective optimization algorithm to search the appropriate cluster center concurrently. Because of the concurrency of the multi-objective optimization algorithm, the calculation time is shortened, and FMOEA-K converts

the clustering effectiveness index, combined with MOEAs to provide the correct clustering center, reducing the error of calculating the optimal k value.

In section 2, we introduce some basic concepts of Multi-Objective Optimization, morphological similarity distance and some Fuzzy Clustering validity indexes. Then we proceed to describe the algorithm in detail in section 3. In section 4, FMOEA-K is used in the experiment and compared with other clustering validity indexes. Finally, we conclude our work and make a discussion of the related work.

2 Related work

2.1 Multi-Objective Optimization

The definition of Multi-Objective Optimization Problem(MOP):

$$\begin{aligned} \text{minimize } F(x) &= \{f_1(x), f_2(x), \dots, f_m(x)\} \\ \text{subject to } &x \in R^n \end{aligned} \quad (1)$$

where $x = \{x_1, x_2, \dots, x_n\}$ is solution with n variables. The objective function $f(x)$ is an objective vector of x . $F : R^n \rightarrow R^m$ represents the mapping from n -dimensional decision variable space to m -dimensional objective function space. Let a and b be two solutions, each of which has n decision variables, if and only if $\forall l = 1, 2, \dots, m$ and $\exists k = 1, 2, \dots, m$, $f(a)_l > f(b)_l$, a is considered to Pareto dominate b , denoted by $a \prec b$. A solution $x \in R^n$ is not Pareto dominated by any one solution, then x is called a Pareto optimal solution[8]. The set of all the Pareto optimal solutions is called the Pareto set and its projection in the objective space is called the Pareto front(PF)[26]. The ideal point (u^{ideal}) is the best objective values in PF: $u^{ideal} = \min_{x \in R^n} f(x) = \{f_1(x), f_2(x), \dots, f_m(x)\}$, and the nadir point (u^{nadir}) is the worst values in PF: $u^{nadir} = \max_{x \in PF} f(x) = \{f_1(x), f_2(x), \dots, f_m(x)\}$.

2.2 Morphological Similarity Distance

Morphological similarity distance :

Zhong Li et al. [25] show that most clustering algorithms are based on Euclidean distance, but the traditional distance can't accurately measure the similarity as shown in Fig.1. Shape similarity distance (MSD) studies the relationship between vector similarity and difference, so that reduces the error of similarity. the definition is as follows:

$$\begin{aligned} MSD(i, j) &= ED(i, j) * (2 - ASD(i, j) / SAD(i, j)) \\ ASD(i, j) &= \left| \sum_{v=1}^n (x_{jv} - x_{iv}) \right| \end{aligned} \quad (2)$$

Where $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ is vector, ASD is the absolute sum of the two vector differences, ED is Euclidean distance, SAD is Manhattan distance.

Fig. 1 Similarity of vector(a, b) $\angle$ vector(a, c) = vector(a, b) by Euclidean distance; Similarity of vector(a, b) $\angle$ vector(a, c) $\angle$ vector(a, b) by MSD

3 The Fuzzy Clustering validity index and its existing problems

3.1 The indicators of PC and PE

Researchers propose an allocating coefficient algorithm(PC)[5] and an allocating entropy algorithm(PE)[4]. The definitions of PC and PE are as follows.

$$V_{PC} = \frac{1}{n} \sum_{j=1}^n \left(\sum_{i=1}^k u_{ij}^2 \right) \quad (3)$$

$$V_{PE} = -\frac{1}{n} \sum_{j=1}^n \sum_{i=1}^k u_{ij} \log_a u_{ij} \quad (4)$$

The core idea of algorithm is that when a data object x_j has been clearly divided into the i th cluster, it would get a large value in u_{ij} . When getting the minimum value of PC or PE, the large value in u_{ij} is the optimized number of Clustering(k). However, the indicators of PE and PC are depending on the membership matrix to ensure the data sets, ignoring the geometric features of data sets. Meanwhile, as the increase of k , the indicators of PC and PE tend to decline.

3.2 The indicator of FS

Fukuyama et al, [16] propose Compactness-separation model, which acquires the indicators of FS from the perspective of Clustering compactness and separation degree, yet due to the magnitude of compactness and separation degree is not consistent, it generates data randomness that is defined as follows:

$$V_{FS} = \sum_{i=1}^k \sum_{j=1}^n u_m^{ij} \|x_j, c_i\| - \sum_{i=1}^k \sum_{j=1}^n u_m^{ij} \|c_i, \bar{c}\| \quad (5)$$

3.3 The indicator of XB

[36] not only considers the membership matrix and geometric structure, but also combine the membership matrix by defining the clustering relationship.

$$V_{XB} = \frac{\sum_{i=1}^k \sum_{j=1}^n u_{ij} \|x_j, c_i\|}{n * \min_{i \neq j} \|c_i, c_j\|} \quad (6)$$

3.4 The indicator of PCAES

[35] propose a partition coefficient and exponential separation index(PCAES). It assigns a normalized distribution coefficient and a separation index to each identified cluster, which better judges the selected approach whether has an outstanding ability of clustering. However, PCAES lacks consideration for increments, especially when handling stream data, we cannot supervise the evolution of the stream data well. The definition of the PCAES is shown in equation(7), where u_M is the degree of clustering.

$$V_{PCAES} = \sum_{i=1}^k \sum_{j=1}^n \frac{u_{ij}^2}{u_M} - \sum_{i=1}^k \exp\left(-\frac{k * \min_S \neq i \|c_i - c_s\|^2}{\sum_l = 1^k} \|c_l - \bar{c}\|^2\right) \quad (7)$$

3.5 The indicator of SC

Zahid et.al[39] propose the indicator of SC that not only considers the position of measured separation ratio and compactness, but also the data structure and the geometric properties of membership matrix. Yet SC always defines the compactness and separation of measurement of data structure, ignoring the definition of each clustering, which is shown in equation(8).

$$V_{SC} = \frac{\sum_{i=1}^k \|c_i - \bar{c}\|^2 / k}{\sum_{i=1}^k (\sum_{j=1}^n u_{ij} \|x_j - c_j\|^2 / \sum_{j=1}^n u_{ij})} + \frac{\sum_{i=1}^{k-1} \sum_{l=i+1}^k (\sum_{n=1}^j \min(u_{ij}, u_{li})^2) / \sum_{n=1}^j \min(u_{ij}, u_{lj})}{\sum_{j=1}^n (\max_{1 \leq i \leq k} u_{ij})^2 / \sum_{j=1}^n \max_{1 \leq i \leq k} u_{ij}} \quad (8)$$

4 FMOEA-K Algorithm

In this section, in order to solve some problems of clustering validity index, we propose a fuzzy clustering optimal k selection algorithm based on Multi-Objective Optimization, which combines the fuzzy clustering validity index with Multi-Objective Optimization algorithm to solve the optimal k -value problem of fuzzy clustering.

4.1 The compactness and separation of fuzzy clustering indexes

In this paper, the existing fuzzy clustering effectiveness index is improved and used as the conflict function of the bi-objective model. The FS index [16] combines the membership degree u and the Euclidean distance between the data points and the data center, which can well measure the compactness within the cluster, but ignores the geometric characteristics of the data and produces certain randomness [38]. Therefore, in order to consider the geometric characteristics of the data, the Euclidean distance in FS index is replaced by morphological similarity distance (MSD) [25,29]. Compared with Euclidean distance, MSD takes into account the shape difference between vectors and reduces the randomness caused by geometric features.

Let $X = x_1, x_2, x_3, \dots, x_n$ is the data set for cluster analysis, and $x_i = x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}$ represents the characteristic of N of x_i . Therefore, the compactness function based on MSD is:

$$Comp(k, U) = \sum_{i=1}^k \sum_{j=1}^n u_{ij} MSD(x_j, c_i) \quad (9)$$

Where k is the number of clustering centers, $C = \{c_1, c_2, \dots, c_k\}$ is the set of clustering centers, U membership matrix, $u_{ij} \in U$ is the membership coefficient of the j -th data in the i -th cluster, and $MSD(x_j, c_i)$ represents the deviation of the j -th data in the i -th cluster.

The separability function tests the relationship between different clusters, and the degree of separation can be obtained by using the distance measure between fuzzy clusters. Because some fuzzy clusters may have overlapping points (as shown in Fig. 2), the average value of overlapping degree between all fuzzy clusters can be calculated by using the separation function proposed in [29], which can reduce the separation error between clusters. Let F_p and F_q is two fuzzy clusters belonging to a fuzzy partition (k, U) . The separation function is shown in formula (10) (11) (12) (13):

Fig. 2 Clustering with overlapping points

$$Sep(k, U) = \frac{2}{k(k-1)} \sum_{p \neq q}^k S(F_p, F_q) \quad (10)$$

$$S(F_p, F_q) = \sum_{j=1}^n S(x_j : F_p, F_q) h(x_j) \quad (11)$$

$$S(x_j : F_p, F_q) = \min(F_p(x_j), F_q(x_j)) \quad (12)$$

$$F_p(x_j) = \frac{\sum_{j=1}^n u_{pj}^m (x_j - c_p)(x_j - c_p)^T}{\sum_{j=1}^n u_{pj}^m} \quad (13)$$

Where $S(F_p, F_q)$ and $p, q \in C$ are the similarity of fuzzy clustering F_p, F_q in dataset X . $F_p(x_j)$ is the covariance matrix of x_j in the p -th cluster. $h(x_j)$ represents a kind of weight. It can adjust the emphasis on overlapping data points according to the sharing degree of overlapping data points between fuzzy clusters, as shown in formula (14):

$$h(x_j) = - \sum_{i=1}^k u_{pj}(x_j) \log_a u_{qj}(x_j) \quad (14)$$

FDCS is defined as the sum of compactness Comp and separation Sep. The smaller the value of FDCS, the smaller the degree of overlap of fuzzy clusters, and the more compact the cluster inside. As shown in formula (15):

$$FDCS(k, U) = Comp(k, U) + Sep(k, U) \quad (15)$$

4.2 Bi-Objective model establishment

As the number of clusters k increase, the value of FDCS will decrease. But FDCS can't be used as the objective function of bi-objective directly, because the value of FDCS will decrease with the increase of k only when the cluster center is right as show in Fig. 3. If FDCS is used as the objective function under the incorrect clustering center, the original correct k non dominant solution may be controlled and the result will fall into the local optimum [19].

Fig. 3 (a) The value of FDCS under real cluster center (b) The value of FDCS under random cluster centers

Theorem1:

$$\min F(x) = \{f_1(x) = 1 - \exp^{-FDCS(k,U)} - k, f_2(x) = k\} \quad (16)$$

Where theorem 1 is monotone decreasing.

Proof: If we assume $f(x)$ decrease monotonically as k increases, then it satisfies $s_1 - s_2 > 0, f_1(s_1) - f_1(s_2) < 0$:

$$\begin{aligned} f_2(s_1) - f_2(s_2) &= (k_1 - k_2) \geq 1 \\ f_1(s_1) - f_1(s_2) &= (k_2 - k_1) - (\exp^{-FDCS(k_2,U)} - \exp^{-FDCS(k_1,U)}) \end{aligned} \quad (17)$$

$$\text{Where } k_2 - k_1 \leq -1 \text{ and } -1 < \exp^{-FDCS(k_2,U)} - \exp^{-FDCS(k_1,U)} < 1$$

We can conclude from (17):

As k increases, $f_1(x)$ keeps monotonically decreasing.

Theorem 2: Each optimized pareto is corresponding to the optimized value of $\min F(x)$

Proof: Let x' be the decision variable corresponding to an optimal (s, k) on PF, i.e in the current solution, $\forall s' \in X, s' \succ s$, so we can get the minimized value $f_1(x')$. And because the second part of $f_1(x')$ is constant. As a result, when $f_2(x) = k, 1 - \exp^{-FDCS(k,U)}$ is at its minimum. Therefore, the optimized value of $\min F(x)$ is x' .

According to the Theorem2, we can transform FDCS to equation(17) form Theorem 1, FDCS can be guaranteed that $f_1(x)$ decreases monotonically with the increase of objective function $f_2(x)$ under the incorrect clustering center.

4.3 Algorithm design

After the bi-objective model is established, the bi-objective model needs to be optimized by MOEA, and finally the PF composed of the optimal solution corresponding to each k value is obtained [32]. If the PF distribution is not uniform, the subsequent PF analysis may fall into a local optimum, and the resulting k value is not accurate. The CDG proposed in the literature [7] uses grid positioning to ensure the diversity of PF. Therefore, CDG is used as the optimization algorithm to optimize the dual objective model.

The initialization of algorithm related parameters is shown in Algorithm 1. Membership U of the data set is obtained by FCM algorithm. The number of grid neighbors, grid distance, ideal point and nadir point of GS are initialized.

After the GS is established, a bi-objective model based on CDG Multi-Objective Optimization algorithm is shown in Algorithm 2. In the follow section, we will explain the steps of the algorithm in detail:

In steps 1-2, the population P is randomly initialized, and an ideal and nadir point is determined based on P [7].

In steps 4-8, N children are produced from P , and an empty set Q is defined for sorting the children. For each solution x their matching solution is obtained by neighbor selection (NS) [7]. In step 7, randomly select two solutions x_i, x_j from NS , and generate their children y from solutions x, x_i, x_j through de operation [24], and then add y to set Q .

Algorithm 1 Initialization**Require:**

X : data set
 k_{min} : maximum number of clusters
 k_{max} : minimum number of clusters
 P : the current population
 u^{ideal} : the approximate ideal point
 u^{nadir} : the approximate nadir point
 n : the number of grid

Ensure: Membership matrix U

- 1: normalize the data;
- 2: Random initialization the cluster center c_i for $k = k_{min}, \dots, k_{max}$;
- 3: **repeat**
- 4: Iteratively call function $FCM(U, V)$ for value of k ;
- 5: Use FCM to cluster and get membership matrix $U(k, X)$;
- 6: $k++$;
- 7: **until** $k > k_{max}$;
- 8: Initializing the $GS(P)$ of CDG: initializing the grid distance and grid neighbors in GS;

Algorithm 2 k -value optimization of fuzzy clustering**Require:**

$mGen$: Maximum generation ;
 N : the population size;
 T : the maximum grid distance for neighborhood;
 k : a range of k values;
 P : the current population;

- 1: Update ideal point and nadir point;
- 2: Initialize $GS(P)$
- 3: **repeat**
- 4: Initialize an empty set $Q = \emptyset$;
- 5: **for** each $x \in P$ **do**
- 6: Obtain the neighboring solution as the mating pool of x by neighbor selection (NS);
- 7: Selection two solution randomly from NS , generate an offspring solution y . from two solution, the two solution by DE operators, and y is added to Q ;
- 8: **end for**
- 9: Update ideal and nadir point, $P = P \cup Q$;
- 10: Update the $GS(P)$;
- 11: Selection parent if the number of $P < N$, randomly select $N - P$ from $\bar{P}$ and add to P , otherwise use Rank-Base Selection(RBS) to select N solutions;
- 12: $gen++$;
- 13: **until** $gen > maxGen$
- 14: Return PF of the current population P ;
- 15: Return the optimal k from PF by DB index;

In step 9, the ideal point and nadir point are updated by using the combination population $P = PUQ$ to make it closer to the real value. In step 10, the GS will be updated with the new population P .

In step 11, a solution with the ideal(nadir) value of all subproblems is selected.

Finally, in step 14-15, after obtaining PF about k value through MOEA, use Davies-Bouldin (DB) index [11] (in this paper, we will still use MSD to replace the Euclidean distance in DB index) as shown in formula (18) (19). The DB index is used to transform the current PF, and optimal k value is obtained by the ratio of the sum of the scattering within the cluster and the separation between the clusters. The smaller the DB index, the better the number of clusters k divided. When the DB is within the specified range, the minimum value is the optimal

k value.

$$DB = \frac{1}{k} \sum_{i=1}^k R_i \quad (18)$$

$$R_i = \max_{j, j \neq i} \left\{ \frac{S_i + S_j}{d_{ij}} \right\}$$

Where $S_i = \frac{1}{V_i} \sum_{x \in V_i} MSD(x, c)$ is the scattering value in the i -th cluster, V_i is the number of data points in the cluster, and c is the cluster center of the cluster. $d_{ij} = MSD(c_i, c_j)$ is the morphological similarity distance between the two clustering centers.

5 Experiment and result analysis

5.1 Experiment Preparation

In order to prove that the PF obtained by this algorithm is uniform, we compare it with the PF obtained by other Multi-Objective Optimization algorithms EMO-KC [33], NSGA-III [12, 20]) when the range of k is large.

In this paper, four UCI real data sets (iris, Bupa Lever Disorder, Wisconsin Diagnostic Burst Cancer, Wine) are used for experiments in the range of $k \in [2, 21]$. The parameter in this paper settings in [7] are referred for the optimization algorithm experiments. The parameter configuration of DE operation in FMOEA-K is the same as in [32]: $CR = 1.0, F = 0.5, \eta = 20$ and $p_m = \frac{1}{n}$. Fig. 4 shows the PF obtained by four optimization algorithms in the interval $k \in [2, 18]$.

It can be observed that the PF of all algorithms is evenly distributed in Wine, WDBC and BLD. However, the PF of EMO-KC[33] is obviously unevenly distributed in Iris data set, while the PF of FMOEA-K and NSGA-III is still evenly distributed. This shows that the PF generated by FMOEA-K is still evenly distributed in the range of k .

The above experiments prove that the PF obtained by FMOEA-K through Multi-Objective Optimization algorithm is uniform. Next, we need to prove whether the optimal clustering number k obtained by FMOEA-K is accurate. We add 6 new artificial datasets and 6 real datasets in the optimal k -value comparison experiment, and compare the results of FMOEA-K and clustering validity index (PC, PE, XB, pcaes, SC, VW). These 6 artificial data sets are: Data-3, Data-3Noise, Data-4, Data-5, Data-6 and Data-4X. The artificial dataset is shown in Fig. 5, where the number in the name represents the number of clusters in the dataset. Table 1 gives a brief description of the datasets described above, where Data-4X takes into account a strong overlap, there are noise points in Data-3Noise. The artificial data set mentioned above is similar to the data set used in [9, 40]. In most literatures, iris has only two clusters. They think that the number of clusters reaches the optimal when $k = 2$, but there are still some clustering algorithms that can generate three clusters, so generally speaking, $k = 2$ or 3 are considered reasonable.

Fig. 4 PF of FMOEA-K, NSGA-III, EMO-KC in interval $k \in [2, 18]$

Fig. 5 All the artificial datasets

Table 1 Brief information about the dataset

Datasets	Data size	Data dimension	The number of clusters
Data-3	100	2	3
Data-4	200	2	4
Data-3Noise	130	2	3
Data-5	2000	2	4 or 5
Data-6	300	2	6
Data-4X	1000	2	4
Wine	178	13	3
WBCD	569	30	2
BLD	345	7	2
Iris	150	4	2 or 3
Breast Cancer	699	9	2
Pima	768	8	2

5.2 Experiment results analysis

Fig. 6 shows the optimal k obtained by FMOEA-K on different data sets. The horizontal axis shows the value range of k , and the vertical axis shows the calculated results (marked with triangle filled points) after the data is normalized. In order to reduce the randomness, when testing the optimal k value obtained by FMOEA-K, we first normalize the values of all data sets to the interval $[0,1]$, and use FMOEA-K to test each target data set 20 times. Finally, according to the corresponding results, we calculate the average value as the final value of the test results. Table 2 shows the optimal number of clusters calculated by different algorithms from the artificial data sets and the real data sets. k_{opt} represents the expected optimal number of clusters. Combined with the results shown in Table 2 and Fig. 6, most of the algorithms in Data-3, Data-4, and Data-6 with obvious clustering and separation effect can get the expected results. The results of PCAES, VW and FMOEA-K on the Data-3Noise show that they are robust to noise. For the data set of Data-5, the points in some areas of the data set overlap. Some algorithms such as PC, XB and PCAES suggest that the optimal k is 4, while FMOEA-K, VW and SC suggest that the optimal k is 5. VW and FMOEA-K can still find the optimal k which is consistent with the expectation in the Data-4X with connection points and point overlap. For BLD, WBCD, Wine in real datasets, almost all algorithms calculate the same optimal k as expected. Although the results of different algorithms in Iris dataset are $k = 2$ or $k = 3$, the optimal k of FMOEA-K is also in this rea reasonable range. According to the comparison of the above results with six different clustering validity indices, we believe that the optimal k obtained by FMOEA-K is effective.

We compare the calculation time of FMOEA-K and the traditional method (using the combination of XB indicator and FCM). Under the same data set Iris, the maximum value of the k value range is set to 10, 50, 100, 150. , The result is shown in Fig. 7. According to the results, as the range of k values continues to expand, it takes less time to obtain the optimal k value due to the parallel characteristics of MOEAs in FMOEA-K than traditional methods.

In order to further study the accuracy of FMOEA-K results, we will use FMOA-K and 6 validity index combined with FCM algorithm to get k_{opt} on Data-4X obtains the corresponding clustering results, as shown in Fig. 8. It can be seen

Fig. 6 The optimal k value obtained by FMOEA-K in the difference datasets

that the clustering results of FMOEA-K and VW on Data-4X according to the corresponding k_{opt} are in line with the experimental expectations, while the results of other indicators based on their corresponding k_{opt} are not in line with the expectations and are disturbed. According to the clustering results, FMOEA-K can well identify the k_{opt} with interference point data in the process of calculating k_{opt} . To sum up, the optimal k value of FMOEA-K in different data sets is effective.

Table 2 Brief information about the dataset

Dataset	k_{opt}	PC	PE	XB	PCAES	SC	VW	FMOEA-K
Data-3	3	2	2	4	3	3	3	3
Data-3Noise	3	2	2	4	3	4	3	3
Data-4	4	4	2	4	4	4	4	4
Data-5	4 or 5	4	6	4	4	5	5	5
Data-6	6	6	6	4	6	4	6	6
Data-4X	4	2	2	5	2	3	4	4
WBCD	2	2	2	2	2	2	2	2
BLD	2	2	2	2	2	4	4	2
Wine	3	2	2	3	3	3	3	3
Iris	2 or 3	2	2	2	2	3	2	2
Breast Cancer	2	2	2	2	10	2	2	2
Pima	2	2	2	10	9	2	2	2

Fig. 7 Time-consuming comparison between FMOEA-K and traditional methods

6 Conclusion

This paper proposes FMOEA-K algorithm which combines Multi-Objective optimization algorithm to solve the problem that the effectiveness index of Fuzzy Clustering with the increase of clustering k . Different from general effectiveness index, we convert clustering task to dual-mode model, which includes cluster indexes and cluster number k . Meanwhile, we get the PF of dual-mode model by

(a) FMOEA-K

(b) PC

Fig. 8 Clustering results of FMOEA-K and different clustering indexes combined with FCM algorithm on Data-4X

utilizing Multi-Objective optimization. Finally, the optimal value k is analyzed by DB index, so that the number of clustering k can be accurately under different types of data sets.

The comparison results with other 6 different cluster validity indexes show that FMOEA-K is effective for most data sets in this study. We could use more real

data sets to test FMOEA-K in the further study and apply it into the analysis of flight test data.

7 Compliance with Ethical Standards

This paper complies with the ethical standards and the specific content is as follows:

Funding: This study was funded by National Natural Science Foundation of China (62072234).

Ethical approval : All applicable international, national, and/or institutional guidelines for the care and use of animals were followed.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

1. de Amorim, R.C.: Constrained intelligent k-means: Improving results with limited previous knowledge. In: 2008 The Second International Conference on Advanced Engineering Computing and Applications in Sciences, pp. 176–180. IEEE (2008)
2. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Tech. rep., Stanford (2006)
3. Begum, S.A.: On limiting behavior of the pbm and the fuzzified pbm cluster validity index. Assam University Journal of Science and Technology **5**(2), 93–98 (2010)
4. Bezdek, J.C.: Cluster validity with fuzzy sets (1973)
5. Bezdek, J.C.: Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology **1**(1), 57–71 (1974)
6. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences **10**(2-3), 191–203 (1984)
7. Cai, X., Mei, Z., Fan, Z., Zhang, Q.: A constrained decomposition approach with grids for evolutionary multiobjective optimization. IEEE Transactions on Evolutionary Computation **22**(4), 564–577 (2017)
8. Coello, C.A.C., Lamont, G.B., Van Veldhuizen, D.A., et al.: Evolutionary algorithms for solving multi-objective problems, vol. 5. Springer (2007)
9. Cui, H., Zhang, K., Fang, Y., Sobolevsky, S., Ratti, C., Horn, B.K.: A clustering validity index based on pairing frequency. IEEE Access **5**, 24884–24894 (2017)
10. Cui, L., Li, G., Lin, Q., Chen, J., Lu, N.: Adaptive differential evolution algorithm with novel mutation strategies in multiple sub-populations. Computers & Operations Research **67**, 155–173 (2016)
11. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
12. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. IEEE transactions on evolutionary computation **18**(4), 577–601 (2013)
13. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters (1973)
14. Fatehi, M., Asadi, H.H.: Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data, a case study from the dalli cu-au porphyry deposit in central iran. Ore Geology Reviews **81**, 245–255 (2017)
15. Fu, Z., Ren, K., Shu, J., Sun, X., Huang, F.: Enabling personalized search over encrypted outsourced data with efficiency improvement. IEEE transactions on parallel and distributed systems **27**(9), 2546–2559 (2015)
16. Fukuyama, Y.: A new method of choosing the number of clusters for the fuzzy c-mean method. In: Proc. 5th Fuzzy Syst. Symp., 1989, pp. 247–250 (1989)
17. Gath, I., Geva, A.B.: Unsupervised optimal fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence **11**(7), 773–780 (1989)

18. Gu, B., Sun, X., Sheng, V.S.: Structural minimax probability machine. *IEEE Transactions on Neural Networks and Learning Systems* **28**(7), 1646–1656 (2016)
19. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. *IEEE transactions on Evolutionary Computation* **11**(1), 56–76 (2007)
20. Jain, H., Deb, K.: An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, part ii: handling constraints and extending to an adaptive approach. *IEEE Transactions on evolutionary computation* **18**(4), 602–622 (2013)
21. Khan, I., Huang, J.Z., Ivanov, K.: Incremental density-based ensemble clustering over evolving data streams. *Neurocomputing* **191**, 34–43 (2016)
22. Li, B., Wang, M., Li, X., Tan, S., Huang, J.: A strategy of clustering modification directions in spatial image steganography. *IEEE Transactions on Information Forensics and Security* **10**(9), 1905–1917 (2015)
23. Li, H., Li, C., Hu, J., Fan, X.: A resampling based clustering algorithm for replicated gene expression data. *IEEE/ACM transactions on computational biology and bioinformatics* **12**(6), 1295–1303 (2015)
24. Li, H., Zhang, Q.: Multiobjective optimization problems with complicated pareto sets, moea/d and nsga-ii. *IEEE transactions on evolutionary computation* **13**(2), 284–302 (2008)
25. Li, Z., Yuan, J., Zhang, W.: Fuzzy c-mean algorithm with morphology similarity distance. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, pp. 90–94. *IEEE* (2009)
26. Miettinen, K.: Nonlinear multiobjective optimization, vol. 12. Springer Science & Business Media (2012)
27. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. *Pattern recognition* **37**(3), 487–501 (2004)
28. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy sets and systems* **155**(2), 191–214 (2005)
29. Rezaee, B.: A cluster validity index for fuzzy clustering. *Fuzzy Sets and Systems* **161**(23), 3014–3025 (2010)
30. Shang, R., Zhang, W., Li, F., Jiao, L., Stolkin, R.: Multi-objective artificial immune algorithm for fuzzy clustering based on multiple kernels. *Swarm and Evolutionary Computation* **50**, 100485 (2019)
31. Vaidya, J., Shafiq, B., Basu, A., Hong, Y.: Differentially private naive bayes classification. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 571–576. *IEEE* (2013)
32. Wang, L., Cui, G., Zhou, Q., Li, K.: A multi-clustering method based on evolutionary multiobjective optimization with grid decomposition. *Swarm and Evolutionary Computation* **55**, 100691 (2020)
33. Wang, R., Lai, S., Wu, G., Xing, L., Wang, L., Ishibuchi, H.: Multi-clustering via evolutionary multi-objective optimization. *Information Sciences* **450**, 128–140 (2018)
34. Webb, A.R.: Statistical pattern recognition. John Wiley & Sons (2003)
35. Wu, K.L., Yang, M.S.: A cluster validity index for fuzzy clustering. *Pattern Recognition Letters* **26**(9), 1275–1291 (2005)
36. Xie, X., Beni, G.: A validity measure for fuzzy clustering: *Ieee—transactions on pattern analysis and machine intelligence* (1991)
37. Xu, R., Wunsch, D.: Survey of clustering algorithms. *IEEE Transactions on neural networks* **16**(3), 645–678 (2005)
38. Yang, S., Li, K., Liang, Z., Li, W., Xue, Y.: A novel cluster validity index for fuzzy c-means algorithm. *Soft Computing* **22**(6), 1921–1931 (2018)
39. Zahid, N., Limouri, M., Essaid, A.: A new cluster-validity for fuzzy clustering. *Pattern recognition* **32**(7), 1089–1097 (1999)
40. Zhang, Y., Wang, W., Zhang, X., Li, Y.: A cluster validity index for fuzzy clustering. *Information Sciences* **178**(4), 1205–1218 (2008)