Chaos, Solitons and Fractals 153 (2021) 111494

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

FKMAWCW: Categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning

Amin Golzari Oskouei[a], Mohammad Ali Balafar[a,*], Cina Motamed[b]

[a] Department of Computer Engineering, University of Tabriz, Tabriz, Iran
[b] Department of Computer Science, University of Orleans, Orléans, France

ARTICLE INFO

Article history:
Received 26 June 2021
Revised 23 September 2021
Accepted 2 October 2021

Keywords:
Fuzzy k-modes
Attribute weighting
Cluster weighting
Clustering

ABSTRACT

The fuzzy k-modes (FKM) is a popular method for clustering categorical data. However, the main problem of this algorithm is that it is very sensitive to the initialization of primary clusters, so inappropriate initial cluster centers lead to poor local optima. Another problem with the FKM is the equal importance of the attributes used during the clustering process, which in real applications, the importance of the attributes are different, and some attributes are more important than others. Some versions of FKM have been presented in the literature, each of which has somehow solved one of the above problems. In this paper, we propose a new clustering method (FKMAWCW) to solve mentioned problems at the same time. In the proposed clustering process, a local attribute weighting mechanism is used to weight the attributes of each cluster properly. Also, a cluster weighting mechanism is proposed to solve the initialization sensitivity. Attribute weight and cluster weight are learned simultaneously and automatically during the clustering process. In addition, to reduce the noise sensitivity, a new distance function is proposed. So, the proposed algorithm can tolerate noisy environment. Extensive experiments on 11 benchmark datasets and an artificially generated dataset show that the proposed algorithm performs better than the state-of-the-art algorithms. This paper presents mathematical analyses to obtain updating functions, providing the convergence proof of the algorithm. The implementation source code of FKMAWCW is made publicly available at https://github.com/Amin-Golzari-Oskouei/FKMAWCW.

1. Introduction

Clustering is an important tool in data mining, machine learning, pattern recognition, and computer vision [1–4]. Clustering aims to divide a series of samples into clusters so that the similarity within the clusters increases and the similarity between them decreases [5–8]. Clustering algorithms are divided into two groups according to the data type: numerical and categorical clustering algorithms. Each attribute in the categorical data contains at least two distinct values with no precedence or delay. In other words, no particular order can be considered between these values. There are some challenges in clustering categorical data [9,10]: 1) for numerical data, the representative of each cluster (centers) often consists of the mean of the samples in each attribute domain of the cluster. Calculating the mean for categorical data is impractical, and 2) standard distance functions such as Manhattan and Euclidean for categorical data are unusable because there is no order between categorical values [11,12].

The fuzzy k-modes (FKM) clustering algorithm [13] is one of the most popular clustering algorithms that is applied for clustering categorical data [14]. FKM has shown successful results in various applications such as [15–18]. In this method, a sample can be assigned to several clusters with different degrees of membership. Fuzzy clustering methods (such as FKM) compared to hard clustering methods (such as k-modes (KM) [19]) can retain more information and achieve better results [20,21]. Also, the fuzzy membership function in this algorithm helps us to discover the complex relationships between a sample and all clusters more accurately [22].

However, the main problem of the FKM algorithm is its sensitivity to the initial cluster centers, which can drop its performance in clustering [23,24]. Another problem with this algorithm is that it considers the same importance for all attributes, while in many real applications, some attributes are more important than others, and giving more importance to these attributes in the clustering process improves the quality of clustering [17,25]. As the number of attributes increases, some attributes may be less important in some clusters while more important in others. Hence, considering the same weight for all of the attributes makes the result of clustering unsatisfactory [17,25]. Also, these methods are extremely sensitive to noise, due to Hamming distance.

[*] Corresponding author.
E-mail addresses: a.golzari@tabrizu.ac.ir (A. Golzari Oskouei), balafarila@tabrizu.ac.ir (M.A. Balafar), motamed@free.fr (C. Motamed).

https://doi.org/10.1016/j.chaos.2021.111494
0960-0779/© 2021 Elsevier Ltd. All rights reserved.

Regarding the FKM's sensitivity to the initialization problem, various methods have been proposed so far. Among them, some methods such as [23], try to eliminate dependence on random initial conditions by spreading the initial cluster representatives in the data space at the initialization step. Some methods apply a scheme to prevent the formation of low-quality clusters during the algorithm's iteration. For example, the methods presented in [26] start from random centers, and the desired centers are calculated automatically during algorithm iterations.

To solve the second problem of FKM (i.e., the problem of giving equal importance to various attributes), different attribute weighting techniques have been proposed, as well. One effective solution for identifying important attributes is to apply a weighting scheme on the attributes [27]. Attribute weighting can be classified into two general groups as follows: 1) global attribute weighting (same weights for attributes in all clusters), and 2) local attribute weighting (different weights for attributes in each cluster). The local weighting mechanism has shown a better performance than the global weighting scheme.

As mentioned before, k-partitioning clustering methods (such as KM and FKM) methods are sensitive to initialization. This sensitivity exists in both algorithms with and without attribute weighting. In fact, attribute weighting is necessary but not enough. Recently in [28], to overcome the two mentioned problems at the same time, a new fuzzy c-means (FCM) clustering algorithm based on the local attribute weighting and cluster weighting schemes was proposed. In this method, cluster weighting was used to reduce the FCM's sensitivity to the selection of initial centers, and attribute weighting was used to increase the accuracy of the clustering. However, this method, such as other FCM-based methods, is not suitable for categorical data clustering due to the distance function used. Our method is a modified version of this method that is adapted for categorical data clustering by introducing a new distance function.

In this paper, inspired by [28], we present a new clustering method to solve mentioned problems simultaneously. In the proposed FKMAWCW algorithm, each attribute is weighted locally (assign different weights for attributes in each cluster). We also assign weight to clusters to handle the initialization problem. Calculating the clusters' weight is performed based on the sum of intra-cluster weighted-feature distances (SIWD). Cluster weights are calculated automatically based on the samples assigned to the clusters during iteration. These weights prevent the creation of clusters with large SIWDs, and systematically, higher-quality clusters are obtained regardless of the initial centers. Also, we define a new distance function based on a combination of frequency probability-based distance [29] and non-Euclidean distance [30]. Using the proposed distance function, the proposed algorithm is robust to noise. In this way, the proposed algorithm can tolerate noisy environment.

The performance of the proposed approach is evaluated on 11 benchmark datasets and compared with the results of other successful clustering algorithms. The obtained results show the high efficiency of the FKMAWCW against the competitors. Extensive experiments are performed to evaluate the effectiveness of each solution applied in the proposed approach.

The remainder of the paper is organized as follows: Section 2 provides an overview of existing clustering methods. In Section 3, the proposed FKMAWCW method is described in detail. In Section 4, the experimental results are presented. In Section 5, the conclusions and possible future works are discussed.

2. Related work

Here, we review existing solutions to each of mentioned challenges in Section 1. Subsection 2.1 reviews the related works for reducing the initialization sensitivity, and Subsection 2.2 reviews related works based on the attribute weight mechanism.

2.1. Methods for reducing initialization sensitivity

Initialization is important in k-partitioning clustering methods. Selecting appropriate initial centers is crucial for achieving a good local minimum [31–34]. Various methods have been introduced to solve the problem of sensitivity to initialization, and here some of these works are reviewed.

Wu et al. [35] proposed an initialization method based on density function. The problem with this algorithm is that the complexity of this algorithm is exponential. To reduce the complexity, random sampling is suggested. Due to the randomness of this step, the same results may not be obtained in all restarts [36].

In 2009, Cao et al. [24] introduced the initialization method that works by the distance between samples and the density of samples. Their method selects a sample with the highest mean density as the initial center for a cluster. To calculate the other centers, the distance between the samples, the previously known clusters, and the mean density of the sample are used. The problem with this method is that a boundary sample may be selected as the first center, which may influence the selection of the initial centers of the other clusters [37].

In [38], Nguyen et al. introduced an extension of the k-means algorithm for clustering categorical data. They proposed a new dissimilarity measure based on an information theoretic definition of similarity that considers the amount of information of two values in the domain set. The definition of cluster centers is generalized using kernel density estimation approach. Then, the new algorithm is proposed by incorporating an attribute weighting scheme that automatically measures the contribution of individual attributes for the clusters. Recently, an improved version of this method was presented in [39].

Khan and Ahmad [37] introduced an initialization algorithm for KM. In this algorithm, multiple data clustering is performed based on attribute values in different clusters. The outputs of this algorithm are used as the initial cluster centers. Also, they proposed a new mechanism for selecting the most relevant attributes, namely prominent attributes.

Jiang et al. [23] proposed two methods for KM algorithm initialization, in which the samples are selected as the initial centers that are not outliers. The first method is the traditional distance-based outlier detection technique and the second method is the partition entropy-based outlier detection technique. The weighted distance function was also used to calculate the distance between two points. In this method, additional parameters are needed, such as the outlier-ness degree of the candidate points for the initial centers and the distance between the initial centers of the candidate and all currently existing initial centers, which can be considered a limitation for this method [40].

Peng et al. [27] introduced an initialization algorithm for KM. This method consists of two steps. At first, the attributes are assigned a global weight, and the distance between two points is calculated by the new criterion based on the attribute weighting. In the second step, the initial centers of the clusters are selected based on the weight of attributes and the combination of distance and density criteria.

2.2. Attribute weighting methods

Solutions that have been proposed for attribute weighting may be classified into two general groups. The first one is global attribute weighting, and the second one is local attribute weighting. These groups are introduced, the methods are reviewed, advantages, and disadvantages of each method are expressed.

1) Global attribute weighting

This group includes algorithms that assign weights to the attributes globally. That is, weights assigned to attributes are identical in all clusters.

For the first time, global attribute weighting was introduced in [18]. Subsequently, the papers [41–43] presented various versions of global attribute weighting clustering methods. However, most of these methods are the weighted versions of k-means or fuzzy c-means.

In 2015, an algorithm based on FKM was introduced [17]. The attribute weighting mechanism in this algorithm improves the clustering quality. The proposed algorithm is not sensitive to noise, but it is sensitive to the initialization of cluster centers.

Huang [43] proposed a weighted KM clustering algorithm to increase the efficiency of the KM algorithm for high-dimensional categorical datasets. This algorithm can automatically calculate weights in the k-mode clustering process. The attributes' weight is calculated by the inverse ratio with the sum of variance within the cluster for each attribute. In this way, the noise attributes are recognized, and their effect on clustering is significantly reduced. Although this algorithm has high clustering accuracy, it is extremely sensitive to initialization [44].

Bai and Liang [44] presented an extended version of the Huang method [43]. In this method, intra-cluster variance and inter-cluster information were used to calculate the attribute weight. They proposed new objective functions for several clustering algorithms (including hard and soft clustering). The balance of intra-cluster and inter-cluster information is made possible by new parameters. Since there is no prior knowledge about these parameters, the exact values are adjusted by a trial and error process, so it is difficult to find the appropriate value [16].

In 2019, a new clustering method called genetic intuitionistic weighted FKM (GIWFKM) [15] was introduced. It is based on FKM and genetic algorithms. This paper first presents the intuitionistic weighted FKM (IWFKM) algorithm that uses intuitionistic sets, new distance function, and weighting attributes. Then, the GIWFKM algorithm is introduced, which combines the IWFKM and the genetic algorithms. The GIWFKM algorithm also uses the unsupervised attribute selection method. The attribute selection is based on the correlation coefficient to eliminate some redundant attributes that can both improve clustering performance and reduce computational time.

2) Local attribute weighting

In contrast to the first group, this group includes algorithms that assign weights to the attributes locally, i.e., the weights assigned to the attributes are different in each cluster. Recently, local attribute weighting methods have gained more attention for clustering categorical data. In this section, we review some of these works.

Chan introduced an attribute-weighted clustering algorithm. The proposed algorithm is highly efficient for mixed datasets. However, for categorical data, the Chan clustering algorithm faces some problems in calculating weights. If there are the same attribute values in some dimension, weight one is given to those attributes and zero for the rest. This means that other attributes are ignored [25].

Cao [25] introduced a new weighting mechanism to solve the problem of Chan's algorithm [45]. In this method, the weight of each attribute in each cluster is calculated based on complement entropy.

Bai et al. [46] introduced the MWKM method, a local attribute weighted algorithm for high dimensional categorical data, an improved version for the KM algorithm. MWKM calculates two weights for each attribute and uses these weights to identify subsets of each cluster's attributes. In addition to common parameters in weighted k-partition clustering algorithm (such as the number of clusters and attribute weight control parameter), their proposed algorithm requires two extra parameters T_s and T_v. These parameters are used in the MWKM objective function to help identify optimal clusters. Experiments in MWKM show that the proposed algorithm improves the clustering accuracy, but it suffers from parameter dependence [47].

In 2016, Chen et al. [48] proposed a soft subspace clustering algorithm for clustering categorical data using fuzzy attribute selection. The distance between samples is calculated using a probabilistic distance function.

In 2018, Jia and Cheung [49] presented a soft clustering method with local attribute weighting for mixed data (combining numerical and categorical data). In this method, the weight of attributes in each cluster is calculated by combining the intra-cluster similarity and the dissimilarity between the clusters. In the proposed method, the appropriate number of clusters is automatically found. In this method, the numerical attributes are converted to categorical attributes by a discretization method, which eliminates some important information [50].

In [51], the authors developed a novel and robust FCM clustering algorithm. The proposed method combined FCM and nonnegative spectral clustering into a unified model, which could further exploit the prior knowledge of data pairs such that both the quality of affinity graph and the clustering performance could be improved. Also, in [52], a novel unsupervised feature selection method was proposed via exploiting the fuzzy membership. By embedding the FCM problem, the fuzzy information and cluster structure were well exploited. Due to the linearity of fuzzy membership, the FKM problem will result in the trivial solution without the support of regularization. To tackle this problem, the FCM problem was embedded with the adaptive loss regularized regression problem concerning the fuzzy membership, such that a sparse and nontrivial solution could be achieved. In other words, the embedded problem performed fuzzy clustering and subspace regression simultaneously. Consequently, the closed-form solutions regarding sparse fuzzy membership and projection matrix could be obtained to evaluate the contribution of each feature simultaneously.

Recently, new clustering methods, namely (deep clustering), have been introduced that combine deep neural networks and clustering methods. These methods use deep learning models to map raw data into embedded space. Then, clustering algorithms (such as k-means) are applied in this new embedded space. Some of these methods are introduced in [53,54]. These methods are able to cluster any type of data such as image, text, and so on. However, in this research field, to the best of our knowledge, a method specifically for clustering categorical data has not yet been introduced.

3. Proposed approach

3.1. Formulation of the FKMAWCW

Concerning the FKM problems in Section 1, we aim to design a clustering algorithm that effectively and efficiently reduces initialization sensitivity by weighting the clusters, and a local attribute weighting mechanism is used to improve clustering accuracy as well. Also, a new distance function was introduced, which is a combination of frequency probability-based distance [29] and non-Euclidean distance [30]. This distance function is not sensitive to noise.

Eq. (1) shows the objective function in the proposed method:

$$F(\mathbf{Z},\mathbf{W},\mathbf{C},\boldsymbol{U})=\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}{u_{nk}}^{\alpha}w_{km}^{q}z_{k}^{p}d^{2}(x_{nm},c_{km}). \tag{1}$$

Subject to

$$u_{nk} \in [0,\ 1],\ \sum_{k=1}^{K} u_{nk} = 1 \text{ where } 0 \le u_{nk} \le 1;$$

$$w_{km} \in [0,\ 1],\ \sum_{k=1}^{M} w_{km} = 1 \text{ where } 0 \le w_{km} \le 1; \tag{2}$$

$$z_k \in [0,\ 1],\ \sum_{k=1}^{K} z_k = 1 \text{ where } 0 \le z_k \le 1\ .$$

Here, $\mathbf{U} = [u_{nk}]$ is a membership matrix, u_{nk} represents the membership of n-th data point to the k-th cluster, $\mathbf{C} = [c_{km}]$ is a matrix of cluster centers, c_{km} represent the m-th attribute in the k-th cluster, $\mathbf{W} = [w_{km}]$ is an attribute weight matrix, w_{km} represents the weight of m-th attribute in the k-th cluster, $\boldsymbol{z} = [z_k]$ represents cluster weight vector having the length of K, $\mathrm{z_k}$ represents the weight of k-th cluster, N is the number of data points, M refers to the number of attributes, and K refers to the number of clusters, $\mathbf{X} = [x_{nm}]$ is a dataset matrix, x_{nm} represents the m-th attribute in the n-th data point, and α is the fuzzification coefficient ($\alpha > 1$). The parameter q is in the range $q < 0$ and $q > 1$. The parameter p is within the range $0 \le p\ < 1$. p is a prior that controls the sensitivity of the weight change at each iteration. More details about parameter p are provided in Subsection 3.3. $d^2(x_{nm}, c_{km})$ is a proposed distance function. More details about this distance function are provided in Subsection 3.2.

Minimizing $F(\mathbf{Z}, \mathbf{W}, \mathbf{C}, \boldsymbol{U})$ with respect to normalization constraints Eq. (2)) is a constrained nonlinear optimization problem. We first fix $\boldsymbol{C}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ and find necessary conditions on $\boldsymbol{U}$ to minimize $\boldsymbol{F}(\boldsymbol{U})$. Then we fix $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{Z}$ and minimize $\boldsymbol{F}(\boldsymbol{C})$ with respect to $\boldsymbol{C}$. then, we fix $\boldsymbol{U}$, $\boldsymbol{Z}$, and $\boldsymbol{C}$ and minimize $\boldsymbol{F}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$. finally, we fix $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{C}$ and minimize $\boldsymbol{F}(\boldsymbol{Z})$ with respect to $\boldsymbol{C}$. then, we fix $\boldsymbol{U}$, $\boldsymbol{Z}$ and $\boldsymbol{C}$ and minimize $\boldsymbol{F}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$. The matrices $\boldsymbol{U}$, $\boldsymbol{C}$, $\boldsymbol{W}$, and Z, are updated according to the Eqs. (3) to (8) respectively. The proof is shown in the Appendix.

$$u_{nk} = \begin{cases} 1,\ if\ x_{n,1\ldots m} = c_{k,1\ldots m} \\ 0,\ if\ x_{n,1\ldots m} = c_{l,1\ldots m} \\ \dfrac{1}{\sum_{l=1}^{K}\left[\frac{z_k^p \sum_{m=1}^{M} w_{km}^q d^2(x_{nm}, c_{km})}{z_l^p \sum_{m=1}^{M} w_{lm}^q d^2(x_{nm}, c_{lm})}\right]^{\frac{1}{(\alpha-1)}}},\ else \end{cases} \tag{3}$$

where $1 \le l \le k$ and $l \ne k$.

$$c_{km} = a_{mr}\ \in DOM\ (A_m), \tag{4}$$

where (see Eqs. (5) and (6)):

$$r = \arg max_{1<t<n_m} \sum_{n=1}^{N} {u_{nk}}^{\alpha},\ constraints\ to\ x_{nm} = a_{mt} \tag{5}$$

$$\sum_{n=1}^{N} {u_{nk}}^{\alpha}, constraints\ to\ x_{nm} = a_{mr} \ge \sum_{n=1}^{N} {u_{nk}}^{\alpha}, \\ constraints\ to\ x_{nm} = a_{mt} \tag{6}$$

n_m indicates the number of the m-th attribute domains. For m-th attribute, a_{mr} and a_{mt} indicate the r-th and t-th domains, respectively.

$$w_{km} = \begin{cases} \frac{1}{h_m}, & if\ Dw_{km} = 0 \text{ and } h_m = |\{s : Dw_{ks} = 0\}| \\ 0, & if\ Dw_{km} \ne 0 \text{ and } \exists s \text{ where } Dw_{ks} = 0 \\ \dfrac{1}{\sum_{s=1}^{M}\left[\frac{Dw_{km}}{Dw_{ks}}\right]^{\frac{1}{q-1}}}, & if\ Dw_{ks} \ne 0 \text{ where } \forall\ 1 \le s \le M \end{cases} \tag{7}$$

where, $Dw_{km} = \sum_{n=1}^{N} {u_{nk}}^{\alpha} d^2(x_{nm}, c_{km})$,

$$z_k = \begin{cases} \frac{1}{g_k}, & if\ Dz_k = 0 \text{ and } g_k = |\{l : Dz_l = 0\}| \\ 0, & if\ Dz_k \ne 0 \text{ and } \exists l \text{ where } Dz_l = 0 \\ \dfrac{1}{\sum_{l=1}^{K}\left[\frac{Dz_k}{Dz_l}\right]^{\frac{1}{p-1}}}, & if\ Dz_l \ne 0 \text{ where } \forall\ 1 \le l \le K \end{cases} \tag{8}$$

where, $Dz_k = \sum_{n=1}^{N}\sum_{m=1}^{M} {u_{nk}}^{\alpha}\ w_{km}^q d^2(x_{nm}, c_{km})$.

3.2. Proposed distance function

We define a new distance function based on a combination of frequency probability-based distance [29] and non-Euclidean distance [30]. Using this distance function, noise attributes have less effect on the results during the clustering process. In this way, the FKMAWCW can tolerate noisy environment. Eq. (9) shows the proposed distance function.

$$d^2(x_{nm}, c_{km}) = 1 - exp\left(-\gamma_m(\delta(x_{nm}, c_{km}).p(x_{nm} = c_{km}))^2\right) \tag{9}$$

where γ_m shows the inverse standard deviation from the mode (SDM) [55] of the m-th attribute of the x dataset. $\delta(x_{nm}, c_{km})$ is defined as Eq. (10).

$$\delta(x_{nm}, c_{km}) = \begin{cases} 1 - \beta,\ if\ x_{nm} =\ c_{km} \\ \beta,\ if\ x_{nm} \ne\ c_{km} \end{cases} \tag{10}$$

The parameter β is within the range of $0.5 < \beta \le 1$. If $\beta =\ 1$, $\delta(x_{nm}, c_{km})$ will become the conventional Hamming distance.

Considering the attribute A_m for x_{nm} and c_{km}, $p(x_{nm} = c_{km})$ is a probability where the values x_{nm} and c_{km} are equal. This probability is calculated based on the frequency of the state of $x_{nm} =\ c_{km}$ in the whole dataset [29] (Eq. (11)).

$$p(x_{nm} = c_{km}) = p(A_m = x_{nm}\boldsymbol{X}).\ p^-(A_m = x_{nm}\boldsymbol{X}) + p(A_m = c_{km}\boldsymbol{X}) \\ .\ p^-(A_m = c_{km}\boldsymbol{X}) \tag{11}$$

Where

$$p(A_m = x_{nm}\boldsymbol{X}) = \frac{\sigma_{A_m = x_{nm}}(\boldsymbol{X})}{\sigma_{A_m}(\boldsymbol{X}\ )\ except\ to\ empty\ values} \tag{12}$$

$$p^-(A_m = x_{nm}\boldsymbol{X}) = \frac{\sigma_{A_m = x_{nm}}(\boldsymbol{X}) - 1}{\sigma_{A_m}(\boldsymbol{X})\ except\ to\ empty\ values - 1} \tag{13}$$

In Eqs. (12) and (13), the operation $\sigma_{A_m = x_{nm}}(\boldsymbol{X})$ counts the number of samples in the data set $\boldsymbol{X}$ that have the value x_{nm} for attribute A_m.

3.3. Discussion

As stated in the introduction, k-mode based algorithms are sensitive to initialization. After a bad initialization, some clusters with large SIWDs[1] (sum of the intra-cluster weighted-feature distance) may be merged, and those with low SIWDs are broken into smaller clusters. Therefore, even if there are some clusters with balanced SIWDs in the dataset, after running the algorithm, some clusters with unbalanced SIWDs may be formed. This problem often happens in k-mode clustering-based algorithms. Preventing the formation of large SIWD clusters helps evade poor solutions after a bad initialization and balances the clusters [56,57]. Weighting the clusters solves this problem to a large extent by balancing the SIWD of the clusters.

[1] SIWD can also be interpreted as the intra-cluster variance. In this case, the cluster variance is defined as the sum, and not the average, of the squared distances from the instances belonging to the cluster to its center. In fact, SIWD is exactly the same Dz_k used in Eq. (8).

To investigate the cluster weighting mechanism, inspired by the research carried out in [56], we rewrite the proposed objective function as Eq. (14).

$$F(\mathbf{Z},\mathbf{W},\mathbf{C},\boldsymbol{U}) = \sum_{k=1}^{K} z_k^p \mathrm{SIWD}_k = \sum_{k=1}^{K} z_k^p \sum_{n=1}^{K}\sum_{m=1}^{M} u_{nk}{}^{\alpha} w_{km}^{q} d^2(x_{nm}, c_{km}) \tag{14}$$

Considering Eq. (12), the weight of the clusters is obtained by Eq. (15).

$$z_k = \frac{\mathrm{SIWD}_k{}^{\frac{1}{1-p}}}{\sum_{l=1}^{K} \mathrm{SIWD}_l{}^{\frac{1}{1-p}}} \tag{15}$$

where SIWD_k indicates k-th cluster SIWD.

According to Eq. (15), for a given partitioning of the data, the weights are set proportionally to the cluster SIWDs. The weight of attributes and clusters are involved in the assignment of samples to clusters (Eq. (3)). Apparently, for higher weighted clusters, the weighted distance of their representatives from the samples increases. Consequently, a cluster with large SIWD may lose some of its current samples that are away from its center, and its SIWD is expected to decrease. At the same time, low SIWD clusters, due to the small weights, may also acquire samples that are not close to their centers, and their SIWD will increase. The used weighting scheme limits the emergence of large SIWD clusters and allows high-quality solutions to be systematically uncovered, irrespective of the initialization.

As mentioned earlier, similar to the method presented in [56], the value of p is chosen in the range $0 \le p < 1$. For $p = 1$, the estimation of the weights simplifies to Eq. (16):

$$z_k = \begin{cases} \approx 1. \ k = \mathrm{argmax}_{1 \le \dot{k} \le K} \mathrm{SIWD}_{\dot{k}} \\ \approx 0. \ otherwise. \end{cases} \tag{16}$$

In each restart of the algorithm, only the cluster with higher SIWD_k gains a weight close to 1, and as a result, all the samples of that cluster are randomly transferred to one of the clusters having a weight close to zero. This leads to the formation of an empty cluster, and the algorithm will not continue properly. If $p > 1$, the objective function mainly focuses on the weighting of the clusters, and as a consequence, weights become more influential than the other parameters. Hence, it does not converge to a minimum point. Therefore, only $0 \le p < 1$ can be permitted.

Furthermore, p has an inverse relationship with the similarity measure between the weights of the clusters (see Eq. 15). It can be shown that, the greater (smaller) p value, the less (more) similar the weight values become, as the relative differences of the SIWDs among the clusters are enhanced (suppressed). This remark also holds for the $z_k{}^p$ values, which are the actual coefficients used in the objective function (see Eq. 1). To clarify this, we define $\frac{z_k}{z_{\dot{k}}}$ as the similarity ratio between the weights of the clusters (see Eqs. (17) and (18)). If the value of this ratio is close to 1, it would indicate greater similarity between the weights of the clusters.

$$\frac{z_k}{z_{\dot{k}}} = \left(\frac{\mathrm{SIWD}_k}{\mathrm{SIWD}_{\dot{k}}}\right)^{\frac{1}{1-p}} \tag{17}$$

$$\frac{z_k{}^p}{z_{\dot{k}}{}^p} = \left(\frac{\mathrm{SIWD}_k}{\mathrm{SIWD}_{\dot{k}}}\right)^{\frac{p}{1-p}} \tag{18}$$

Considering $0 \le p < 1$, as p increases, the value of the $\frac{1}{1-p}$ and $\frac{p}{1-p}$ exponents grow, thus the relative differences of the cluster SIWDs are enhanced, and both ratios deviate more from 1, i.e., the weights and coefficients $z_k{}^p$ attain less similar values (the exact opposite holds when p is decreased). In other words, p adjusts how intensely the differences of the cluster SIWDs are reflected on the weights.

Therefore, for a high p value ($p \cong 1,\ p \ne 1$), large SIWD clusters accumulate considerably higher z_k and $z_k{}^p$ values compared to low SIWD clusters, resulting in an objective that severely penalizes clusters with high SIWD. Note that an extremely high p may force clusters with large SIWD to lose most, or even all their samples, as their enormous weights will excessively distance the samples from their centers (Eq. (4)), something not desired of course.

3.4. pseudo-code of the FKMAWCW

The pseudo-code of the FKMAWCW clustering method shows in Fig. 1. Similar to the method presented in [56], to find the appropriate value of p, we apply an iteration-based approach using three parameters p_{init}, p_{step}, and p_{max}. We start the algorithm using a small value (p_{init}) for p. In each iteration, we increase p as much as p_{step} until the maximum value (p_{max}) is obtained. If an empty cluster or a cluster with one sample appears, we decrease p by p_{step} regardless of whether p equals p_{max} or not. At this point, we choose the values of u_{nk}, w_{km}, and z_{km} corresponding to the previous p. The algorithm continues until in two successive iterations, the difference between the two objective function values is less than the threshold value ε, or the number of iterations reaches the maximum (t_{max}). The implementation source code of FKMAWCW is made publicly available at https://github.com/Amin-Golzari-Oskouei/FKMAWCW.

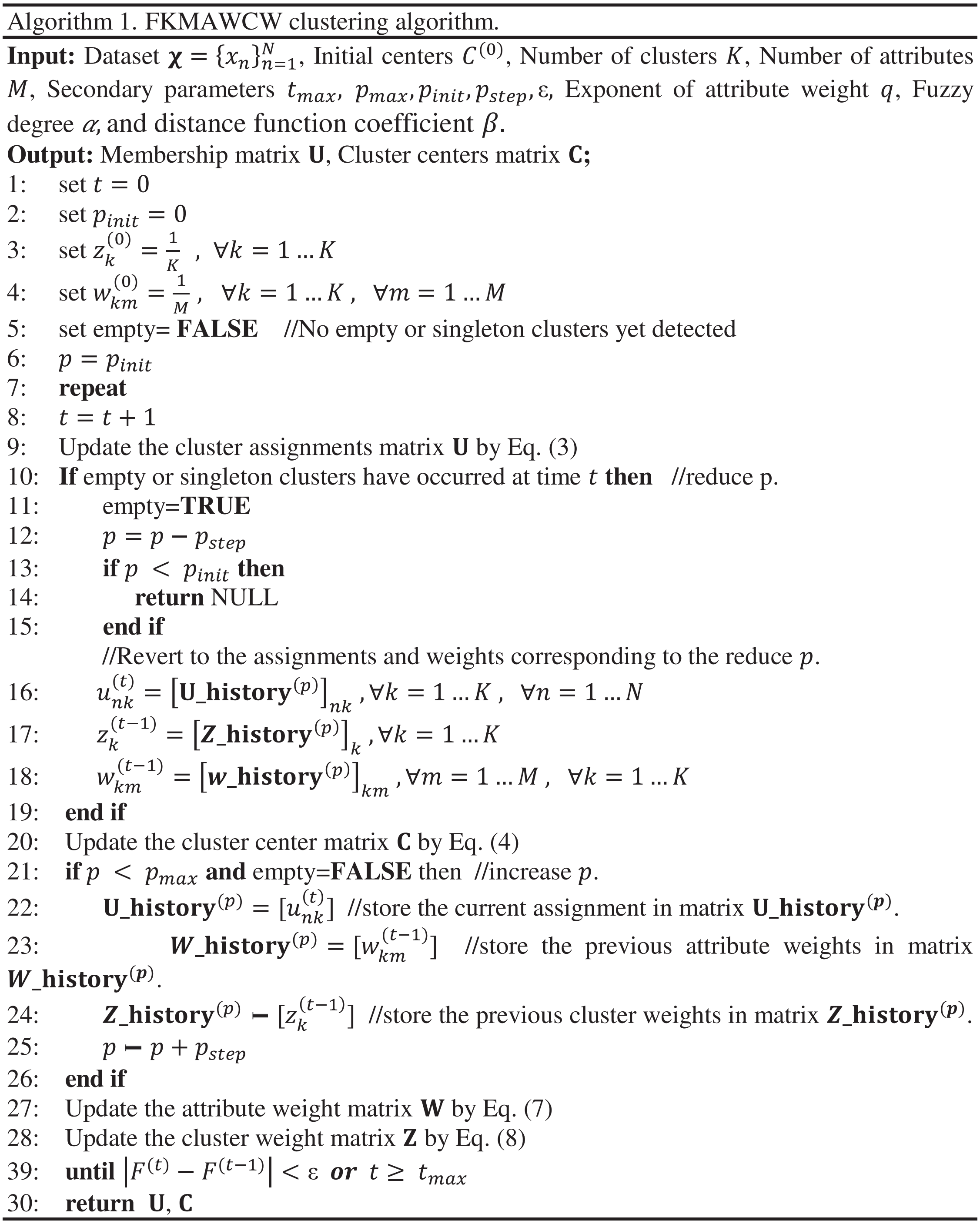
3.5. Computational complexity

According to the FKMAWCW, it can be found that the complexity of computation for the FKMAWCW depends on four steps. The steps for updating $\boldsymbol{U}$, $\boldsymbol{C}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$. The computation complexity for the second step is $O(\hat{M}KN)$, and for the rest of the steps is $O(MKN)$, where $\hat{M} = \sum_{m=1}^{M} T_m$ and T_m indicate the number of m-th attribute domains. Thus, the computational complexity for an iteration of the algorithm is $O(KN(3M + \hat{M}))$. Although due to the extra step of calculating the weight of attributes and clusters, our algorithm has more computational complexity than the FKM algorithm. When $\hat{M} \gg M$, the asymptotic difference is negligible.

4. Experiments

In this section, the performance of the proposed approach is evaluated. The results are compared with the following groups of methods:

1) Initialization sensitivity reduction methods:
   - Wu [58]: A New Initialization Method for Clustering Categorical Data
   - Khan [37]: Cluster center initialization algorithm for KM clustering;
   - Cao [24]: A new initialization method for categorical data clustering;
   - Jiang [23]: Initialization of KM clustering using outlier detection techniques
   - Peng [27]: Attribute weight-based clustering centers algorithm for initializing KM clustering.
   - Mod-2 [38]: A k-Means-Like algorithm for Clustering categorical data using an information theoretic-based dissimilarity measure;
   - Mod-3 [39]: A method for k-means-like clustering of categorical data;
2) Attributes-weighted methods:
   - IWFKM [15]: intuitionistic weighted fuzzy k -modes algorithm for categorical data;



```
Algorithm 1. FKMAWCW clustering algorithm.
Input: Dataset $\boldsymbol{\chi} = \{x_n\}_{n=1}^{N}$, Initial centers $C^{(0)}$, Number of clusters $K$, Number of attributes $M$, Secondary parameters $t_{max}, p_{max}, p_{init}, p_{step}, \varepsilon$, Exponent of attribute weight $q$, Fuzzy degree $\alpha$, and distance function coefficient $\beta$.
Output: Membership matrix U, Cluster centers matrix C;
    set $t = 0$
    set $p_{init} = 0$
    set $z_k^{(0)} = \frac{1}{K}$ , $\forall k = 1 \dots K$
    set $w_{km}^{(0)} = \frac{1}{M}$ , $\forall k = 1 \dots K$ , $\forall m = 1 \dots M$
    set empty= FALSE    //No empty or singleton clusters yet detected
    $p = p_{init}$
    repeat
    $t = t + 1$
    Update the cluster assignments matrix U by Eq. (3)
    If empty or singleton clusters have occurred at time $t$ then   //reduce p.
        empty=TRUE
        $p = p - p_{step}$
        if $p < p_{init}$ then
            return NULL
        end if
        //Revert to the assignments and weights corresponding to the reduce $p$.
        $u_{nk}^{(t)} = \left[\mathbf{U\_history}^{(p)}\right]_{nk}$ ,$\forall k = 1 \dots K$ , $\forall n = 1 \dots N$
        $z_k^{(t-1)} = \left[\boldsymbol{Z\_history}^{(p)}\right]_k$ ,$\forall k = 1 \dots K$
        $w_{km}^{(t-1)} = \left[\boldsymbol{w\_history}^{(p)}\right]_{km}$ ,$\forall m = 1 \dots M$ , $\forall k = 1 \dots K$
    end if
    Update the cluster center matrix C by Eq. (4)
    if $p < p_{max}$ and empty=FALSE then  //increase $p$.
        $\mathbf{U\_history}^{(p)} = [u_{nk}^{(t)}]$ //store the current assignment in matrix $\mathbf{U\_history}^{(\boldsymbol{p})}$.
        $\boldsymbol{W\_history}^{(p)} = [w_{km}^{(t-1)}]$   //store the previous attribute weights in matrix $\boldsymbol{W\_history}^{(\boldsymbol{p})}$.
        $\boldsymbol{Z\_history}^{(p)} \leftarrow [z_k^{(t-1)}]$ //store the previous cluster weights in matrix $\boldsymbol{Z\_history}^{(\boldsymbol{p})}$.
        $p \leftarrow p + p_{step}$
    end if
    Update the attribute weight matrix W by Eq. (7)
    Update the cluster weight matrix Z by Eq. (8)
    until $\left|F^{(t)} - F^{(t-1)}\right| < \varepsilon$ or $t \geq t_{max}$
    return U, C
```

Fig. 1. Proposed clustering algorithm.

- EWKM [25]: A weighting KM algorithm for subspace clustering of categorical data;
- Saha [17]: Categorical FKM Clustering with Automated Attribute Weight Learning;
- SBC[59]: space structure and clustering of categorical data;
- Chan [45]: An optimization algorithm for clustering through weighted dissimilarity measures
- Jia [49]: Subspace Clustering of Categorical data

Parameter ε and the maximum number of iterations are common in the implemented methods, which are set to 10^{-5} and 100, respectively. The parameter α in fuzzy algorithms is set to 2. In the FKMAWCW, the required parameters are set as follows: $p_{step} = 0.01$, $p_{init} = 0$, $p_{max} = 0.5$, $q = 2$, and β is chosen from the range $\beta \in \{1,\ 0.9,\ 0.99\}$. In the experiments, the best results for compared algorithms have been quoted directly from the relevant articles.

4.1. Dataset

To evaluate the proposed method efficiency, the synthetic and real datasets were used as follows:

1) Synthetic dataset

To easily examine the efficiency of the solutions proposed in our method to solve each of the existing challenges, we use a syn-

Table 1
Real-world dataset.

Dataset	Number of data points	Number of dimensions	Number of Classes
Balance scale	625	4	3
Car evaluation	1728	6	4
Chess	3196	36	2
Dermatology	366	34	6
Lung	32	56	3
Lymphography	148	18	4
Mushroom	8124	22	2
Nursery	12960	8	5
Soybean	47	35	2
Voting	435	16	2
Zoo	101	17	7

thetic dataset. In this dataset, there are 300 four-dimensional samples in 3 classes, 100 samples per class. The domain of each attribute is a to j, i.e. $DOM(Feature_m) = \{a, b, c, d, e, f, g, h, i, j\}, \forall 1 \leq m \leq M,$ where M represents the total number of attributes. Samples 1 to 100 are in class 1, and the first attribute values in them are a, samples 101 to 200 are in class 2, and the second attribute values in them are e, samples 201 to 300 are in class 3, and the third attribute values in them are j. The rest of the values in the entire dataset are randomly selected from a to j.

2) Real-world dataset

To evaluate the performance of the FKMAWCW and compare its results with other state-of-the-art methods, we use 11 standard and real-world datasets from UCI [60] data repository. The details of this dataset are summarized in Table 1.

4.2. Performance criteria

The four criteria of the *Adjusted Rank Index (ARI)* [61], *Accuracy (ACC), Precision (PR),* and *Recall (RE)* are used in the conducted experiments to measure the efficiency of the algorithms.

Adjusted Rand Index:

$$ARI\ (T, C) = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{19}$$

In Eq. (19), T is the target class label, C is the result of the clustering algorithm, a, b, c, d are the number of samples that are both in class similar to C and T, in one class in T but different in class C, in a class in C but different in T and in different classes in both T and C.

Accuracy:

$$ACC\ (T, C) = \frac{\sum_{k=1}^{K} e_k}{N} \tag{20}$$

Precision:

$$PR\ (T, C) = \frac{\sum_{k=1}^{K} \left(\frac{e_k}{e_k + f_k}\right)}{N} \tag{21}$$

Recall:

$$RE\ (T, C) = \frac{\sum_{k=1}^{K} \left(\frac{e_k}{e_k + g_k}\right)}{N} \tag{22}$$

In Eqs. (20 to 22), e_k is the number of samples that are correctly assigned to the class T_k $(e_k = |T_k \cap^{C_k}|)$. f_k is the number of samples that are incorrectly assigned to the class $T_k(f_k = |C_k| - e_k)$. g_k is the number of samples that are incorrectly rejected from T_k $(g_k = |T_k| - e_k)$. In this paper, the mean of these criteria is used per 100 restarts.

4.3. Experiment 1: the effect of attribute weighting

In this section, we investigate the effect of the proposed attribute weighting schema on the final clustering results. We aim to examine whether the proposed algorithm, properly assigns weights to attributes or not. To this end, we use the synthetic dataset. To demonstrate the importance of each attribute in this dataset, we illustrate the synthetic dataset (see Fig. 2). As shown in this figure, it is understood that Cluster1 (samples 1 to 100), Cluster2 (samples 101 to 200), and Cluster3 (samples 201 to 300) are mainly formed based on Attribute1, Attribute2, and Attribute3, respectively. In other words, some attributes in some clusters are noisy. These noise attributes include the fourth attribute in all three clusters, the second and third attributes in Cluster1, the first and third attributes in Cluster2, and the first and second attributes in Cluster3. These attributes do not provide useful information for optimal clustering. This means that the first, second, and third attributes are more important in Cluster1, Cluster2, and Cluster3, respectively.

The FKMAWCW is run on the synthetic dataset. We examine the weights obtained for each cluster. Fig. 3 shows the final weights assigned to each of the attributes in the different clusters. According to this figure, it is observed that the FKMAWCW gives more weight to the first, second, and third attributes in Cluster1, Cluster2, and Cluster3, respectively. This level of difference in the weights obtained is in line with the prediction made on the data illustrated in Fig. 2, indicating the proper performance of the local weighting method adopted in the proposed algorithm. In other words, the proposed algorithm can properly distinguish the noisy attributes from non-noisy attributes and give more weight to the non-noisy attributes in each cluster.

4.4. Experiment 2: the effect of cluster weighting

In this section, we investigate the effect of the proposed cluster weighting schema on the final clustering results. We run the proposed algorithm once by assigning weights to the clusters employing our algorithm, and once without assigning weights. In the case of the latter experiment, we give the same weight for all the clusters so that $z_k = \frac{1}{\boldsymbol{K}}$. In both experiments, the algorithm is restarted 1000 times from the same randomly chosen initial centers, and average results are reported. Considering the number of restarts 1000 times, both appropriate and inappropriate initial centers are covered. The clustering results from both experiments are presented in Fig. 4. As shown in this figure, the proposed algorithm has a better performance through cluster weighting than the case without cluster weighting. When the weight is not assigned to the clusters, the efficiency of the proposed algorithm is decreased. By using cluster weighting, the *ACC, ARI, PR, and RE* rates of the proposed method are improved 2.4%, 2.24%, 2.1%, and 1.51%, respectively.

4.5. Experiment 3: robustness to noise

As discussed in Subsection 3.2, the proposed distance function is robust to noise. To evaluate the robustness of the proposed distance function and compare it with other distance functions, in this experiment, we add noise to the synthetic dataset. Noise applied on synthetic datasets is a combination of empty and random (in range $\{a, b, c, d, e, f, g, h, i, j\}$) values. This noise is applied on the synthetic dataset with different percentages (see Fig. 5).

We run the FKMAWCW with different well-known distance functions on the synthetic dataset with different percentages of noise. The FKMAWCW, with all tested distance functions (the proposed, Hamming, and frequency probability-based distance func-

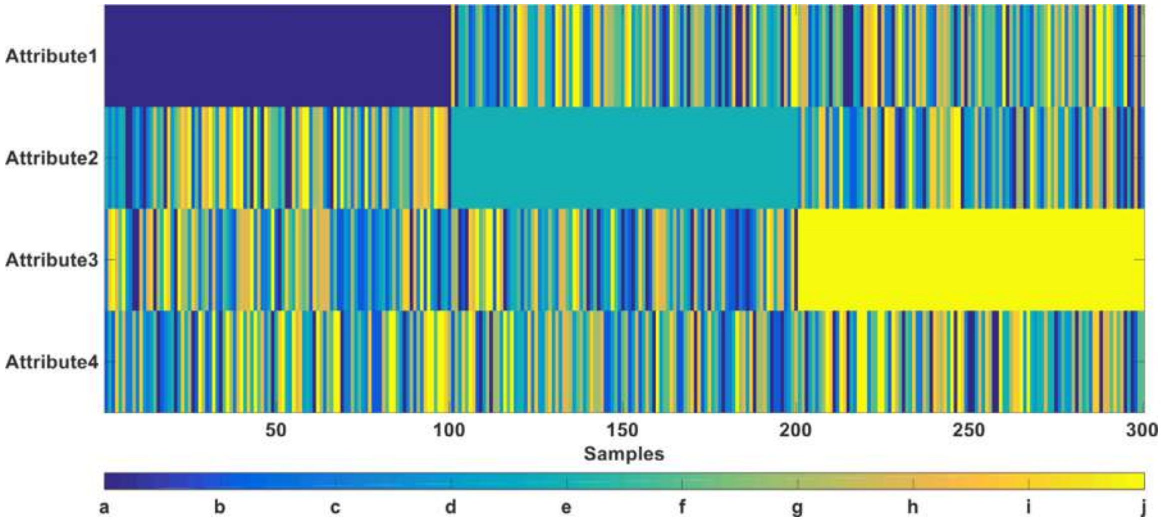


Fig. 2. Synthetic dataset.

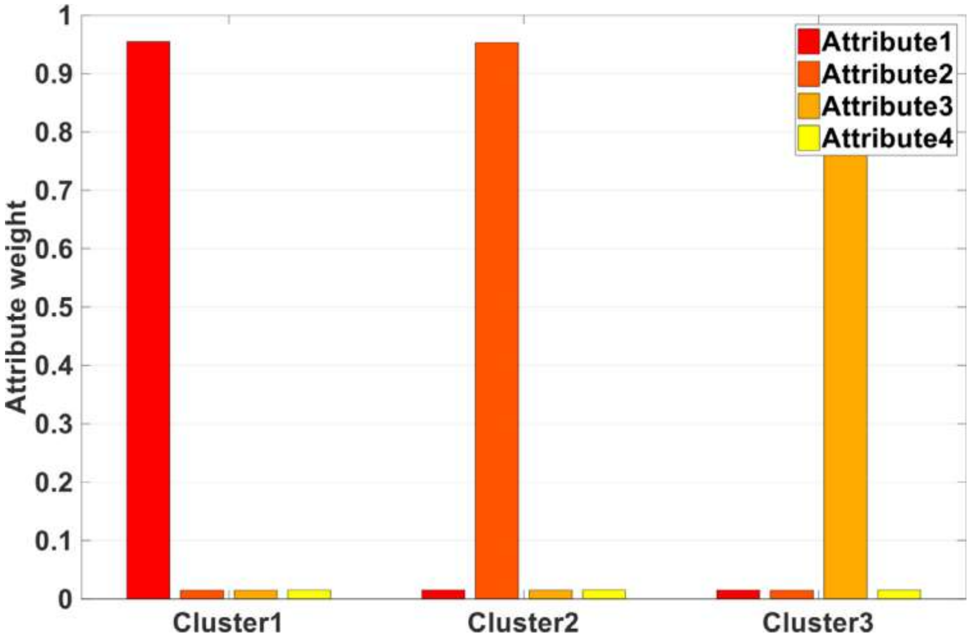


Fig. 3. Weights assigned to the attributes of Cluster1, Cluster2, and Cluster3 in the synthetic dataset.

tions), is restarted 10 times from the same randomly chosen initial centers. Table 2 shows the obtained results for all tested distance functions. As shown in Table 2, the proposed method performs better than Hamming distance for all noise levels and has a lower performance than frequency probability for 25%, 30%, and 45% noise values. Although the frequency probability method works better for some noise levels, the result is extremely low for noise below 20%. Also, for noise below 20%, it has even lower overall performance than the Hamming distance. In general, the proposed algorithm with the new distance function has a better overall performance than the Hamming distance and frequency probability functions and is more robust to noise.

To investigate the behavior of the objective function on a synthetic dataset with different noise levels, we illustrate the value of the objective function during the algorithm iterations (see Fig. 6) to show how the proposed algorithm alternates between the $\boldsymbol{U}$, $\boldsymbol{C}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ optimization steps to get a local optimum of $\boldsymbol{F}$. As shown in Fig. 6, the value obtained for the objective function with the proposed distance function is reduced in the initial iterations markedly. Finding the suitable cluster centers in the initial iterations is the reason for such behavior. The value of the objective function has been significantly reduced from first to about 15th iterations and slight decreases from iterations 15th to the end. This shows that the proposed algorithm achieves a nearly optimal solution in the initial iterations. Also, the overall convergence time (number of iterations) of the proposed distance criterion is less than the others. This indicates that the proposed method has good convergence.

4.6. Clustering results on real-world datasets

In this section, we evaluate the FKMAWCW and other algorithms on real-world datasets.

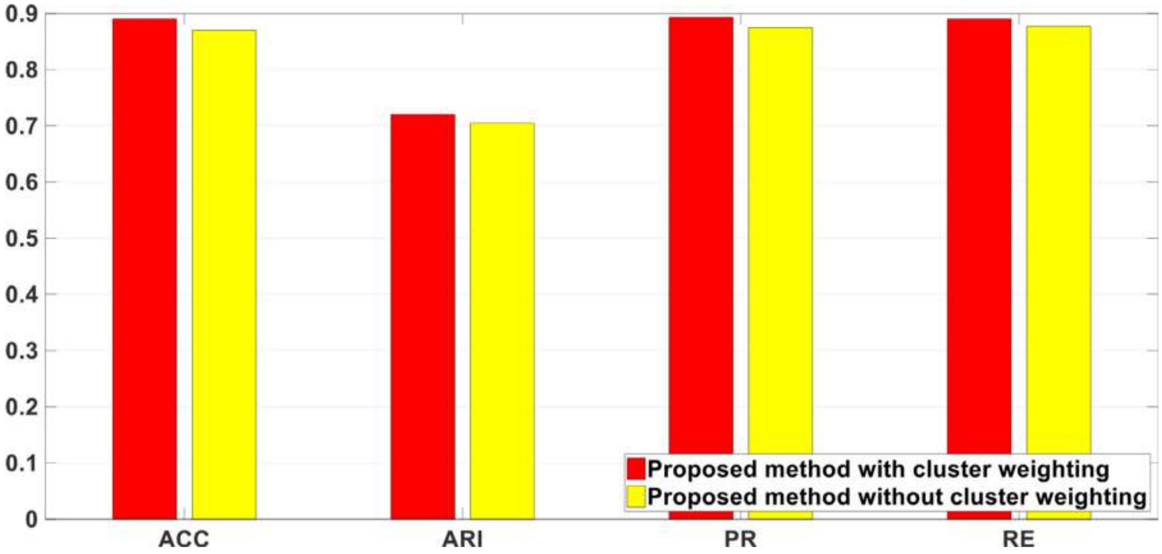


Fig. 4. Effect of cluster weighting on the quality of clustering on the synthetic dataset.

Table 2
Comparison of different distance functions with different percentages of noise.

Distance	Metric	Noise								
		5%	10%	15%	20%	25%	30%	35%	45%	50%
Hamming	**ACC**	0.8467	0.7637	0.6324	0.5870	0.5810	0.5593	0.5130	0.5222	0.5047
	ARI	0.5898	0.4759	0.2624	0.2190	0.1976	0.1895	0.1338	0.1137	0.0971
	PR	0.8528	0.7840	0.6240	0.5959	0.5858	0.5455	0.4742	0.5374	0.5372
	RE	0.8467	0.7637	0.6324	0.5870	0.5810	0.5593	0.5130	0.5222	0.5047
Frequency probability	**ACC**	0.7217	0.7314	0.6940	0.6678	**0.6546**	0.6387	0.5733	**0.6121**	0.5396
	ARI	0.4749	0.4540	0.3977	0.3554	**0.2988**	**0.2634**	0.1911	**0.1893**	**0.1283**
	PR	0.7843	0.7347	0.7056	0.6523	0.6505	**0.6574**	0.5777	**0.6472**	**0.5650**
	RE	0.7217	0.7314	0.6940	0.6678	0.6546	0.6387	0.5733	**0.6121**	0.5396
Proposed	**ACC**	**0.8583**	**0.7973**	**0.7367**	**0.7200**	0.6457	**0.6473**	**0.6003**	0.5500	**0.5430**
	ARI	**0.6176**	**0.5107**	**0.4080**	**0.3875**	0.2742	0.2588	**0.1942**	0.1335	0.1191
	PR	**0.8632**	**0.8034**	**0.7390**	**0.7240**	**0.6515**	0.6527	**0.6114**	0.5599	0.5544
	RE	**0.8583**	**0.7973**	**0.7367**	**0.7200**	**0.6457**	**0.6473**	**0.6003**	0.5500	**0.5430**

Table 3
Comparison of the proposed method with attributes-weighted methods.

Datasets	Saha [17]	SBC [59]	IWFKM [15]	FKMAWCW
Lung	0.58	0.63	0.597	**0.6406**
	0.14	0.216	0.232	**0.2682**
Dermatology	0.635	0.793	0.695	**0.7936**
	0.305	0.545	0.391	**0.7466**
Mushroom	0.644	0.798	**0.825**	0.8182
	0.238	0.387	0.376	**0.4053**
Zoo	0.72	0.579	**0.821**	**0.8218**
	0.719	0.404	0.711	**0.7806**
Voting	0.82	0.876	**0.899**	0.8922
	0.577	0.564	**0.644**	0.6137
Soybean	0.893	0.936	0.894	**1**
	0.788	0.85	0.719	**1**

4.6.1. Experiment 4: FKMAWCW algorithm vs. attributes-weighted methods

This section considers evaluating the proposed FKMAWCW algorithm compared to the benchmark attributes-weighted algorithms in terms of the *ACC* and *ARI*. Note that 11 datasets are used to evaluate in the proposed FKMAWCW in Subsections 4.6.2 and 4.6.3. However, only 6 datasets are selected to compare with the benchmark algorithms because these are the mutual datasets that were used to conduct the experiment on both the proposed and the benchmark algorithms. The performance of the different methods is shown in Table 3. This table shows the *ACC*, and *ARI* rates from top to bottom for each dataset, respectively. The results for compered methods have been quoted directly from the relevant publications. In Table 3, the best rates are bold-faced.

It is not difficult to see that the proposed FKMAWCW algorithm outperforms its rivals since it achieves better results on 5 datasets (i.e., *Lung, Dermatology, Mushroom, Zoo*, and *Soybean*) in a total of 6 tested datasets. For the *Voting* dataset, the best *ACC* and *ARI* are obtained by IWFKM [15] algorithm. After the proposed method, the IWFKM [15], SBC [59], and Saha [17] methods have relatively good performance, respectively.

Table 4 shows the average results for FKMAWCW and other compared algorithms. As shown in this table, FKMAWCW has the best results. In terms of *ACC* and *ARI* metrics, after FKMAWCW, the methods IWFKM [15], SBC [59], and Saha [17] have better outcomes, respectively.

In some methods, such as Jia [49], EWKM [25], and Chan [45], only some datasets have been tested. In Jia [49], only the *ACC* criterion is reported. In this method, the *ACC* criteria for *Soybean, Voting,* and *Zoo* datasets are 1, 0.8786, and 0.7624, respectively, which is not an improvement over the proposed method. In Chan [45], for the *Mushroom, Soybean*, and *Voting* datasets, the *ACC* cri-

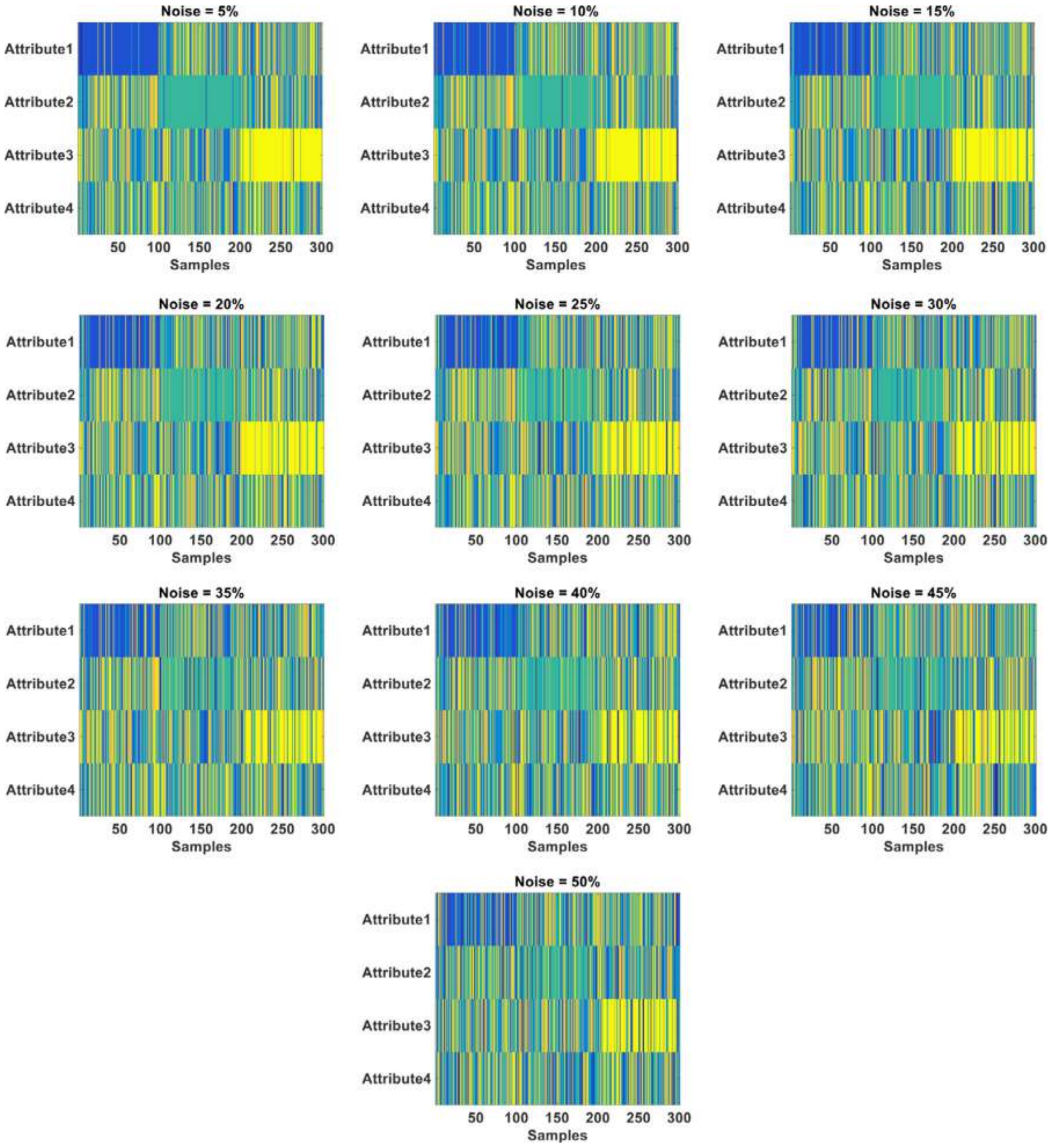


Fig. 5. Synthetic dataset with different percentages of noise.

Table 4
The average performance of FKMAWCW and other attributes-weighted methods.

Metrics	Saha [17]	SBC [59]	IWFKM [15]	FKMAWCW
ACC	0.7153	0.7687	0.7885	**0.8277**
ARI	0.4612	0.4943	0.5122	**0.6357**

teria are 0.6195, 0.7117, and 0.7894, respectively, and the *ARI* criteria are 0.0020, 0.5328, and 0.3638, respectively. The results of this method are also worse than the proposed method. In EWKM [25], for the *Mushroom, Soybean*, and *Voting* datasets, the *ACC* criteria are 0.7905, 0.8972, and 0.8651, respectively, and the *ARI* criteria are 0.3586, 0.8054, and 0.5345, respectively. The results of this method, like the previous two methods, are also worse than the proposed method.

4.6.2. Experiment 5: FKMAWCW algorithm vs. initialization sensitivity reduction methods

In this section, the performance of FKMAWCW is compared with initialization sensitivity reduction methods. The performance of the different methods is shown in Table 5. This table shows the *ACC*, *PR*, and *RE* rates from top to bottom for each dataset, respectively. The results of other methods have been quoted directly from the relevant publications.

As shown in Table 5, FKMAWCW has the best overall performance for all datasets except *Mushroom* and *Zoo*. In these two datasets, the method Khan [37] performs better than the FKMAWCW method. Generally, the results show that FKMAWCW has a higher performance than other successful methods in this field.

To compare the proposed FKMAWCW algorithm with other benchmark algorithms, Table 6 shows the result of the hypothesis

Table 5
Comparison of Proposed Method with initialization sensitivity reduction methods.

Datasets	Khan [37]	Cao [24]	Mod-2 [38]	Mod-3 [39]	FKMAWCW
Lung	0.5	0.5	0.5022	0.4922	**0.6406**
	0.6444	0.5584	0.5726	0.5515	**0.7423**
	0.5168	0.5014	0.4595	0.4512	**0.6614**
Mushroom	**0.8815**	0.8754	0.7474	0.7682	0.8182
	0.8975	0. 9019	0.7586	0.7718	0.8566
	0.878	0.8709	0.7472	0.7686	0.8483
Soybean	0.9574	**1**	0.907	0.8666	**1**
	0.9583	**1**	0.906	0.8552	**1**
	0.9705	**1**	0.9006	0.8567	**1**
Voting	0.8506	0.8621	0.8764	0.8764	**0.8922**
	0.8484	0. 8571	0.8724	0.8724	**0.9541**
	0.8672	0.8755	0.8921	**0.892**	0.8387
Zoo	**0.8911**	0.8812	0.7601	0.7524	0.8218
	0.7224	**0.8702**	0.6518	0.6193	0.7806
	0.7716	0.6714	0.6503	0.6494	0.7289
Balance scale	0.4129	0.376	0.431	0.4323	**0.5054**
	0.3609	0.3282	0.4177	0.4238	**0.4798**
	0.3541	0.3228	0.3957	0.4009	**0.4588**
Car evaluation	0.3576	0.4936	0.3725	0.3831	**0.5795**
	0.2415	**0.3826**	0.3407	0.344	0.2855
	0.2499	**0.4875**	0.365	0.3639	0.2609
Chess	**0.704**	0.5663	0.5279	0.5385	0.5583
	0.5312	0.5796	0.5326	0.5425	**0.6035**
	0.554	0.5537	0.5296	0.5394	0.2465
Dermatology	0.6175	0.5874	0.728	0.7404	**0.7926**
	0.6841	0.5604	0.6696	0.6589	**0.7544**
	0.6165	0.526	0.696	0.6943	**0.7821**
Lymphography	0.5068	0.3514	0.5433	0.5341	**0.6139**
	0.4226	0.2698	**0.475**	0.4599	0.4221
	0.4451	0.2955	**0.5438**	0.5328	0.4408
Nursery	0.2804	0.3651	0.3165	0.3128	**0.3662**
	0.2304	0.2978	0.2954	0.293	**0.3078**
	0.2044	0.2273	0.2501	0.2426	**0.2501**

Table 6
The result of the statistical test for the FKMAWCW algorithm and other state-of-the-art methods.

Metrics	Datasets	FKMAWCW vs. Khan	FKMAWCW vs. Cao	FKMAWCW vs. Mod-2	FKMAWCW vs. Mod-3
ACC	*Lung*	+	+	+	+
	Mushroom	-	-	+	+
	Soybean	+	=	+	+
	Voting	+	+	+	+
	Zoo	-	-	+	+
	Balance scale	+	+	+	+
	Car evaluation	+	+	+	+
	Chess	-	-	+	+
	Dermatology	+	+	+	+
	Lymphography	+	+	+	+
	Nursery	+	+	+	+
PR	*Lung*	+	+	+	+
	Mushroom	-	-	+	+
	Soybean	+	=	+	+
	Voting	+	+	+	+
	Zoo	+	-	+	+
	Balance scale	+	+	+	+
	Car evaluation	+	-	-	-
	Chess	+	+	+	+
	Dermatology	+	+	+	+
	Lymphography	-	+	-	-
	Nursery	+	+	+	+
RE	*Lung*	+	+	+	+
	Mushroom	-	-	+	+
	Soybean	+	=	+	+
	Voting	-	-	-	-
	Zoo	-	+	+	+
	Balance scale	+	+	+	+
	Car evaluation	+	-	-	-
	Chess	-	-	-	-
	Dermatology	+	+	+	+
	Lymphography	-	+	-	-
	Nursery	+	+	+	+

test on each dataset. Symbol "+" indicates that the proposed FKMAWCW algorithm performs a better result. Similarly, the symbol "=" indicates the equal result or no difference between the two algorithms, while "−" means the worse result of the FKMAWCW algorithm. According to the statistical result in Table 6, the FKMAWCW algorithm performs the worse result than those of all benchmark algorithms on the *Voting* and *Chess* datasets in term of the *RE*. Compared with the Khan [37] algorithm, the FKMAWCW algorithm yields the better results on 8 datasets in term of the *ACC* (i.e., *Lung, Soybean, Voting, Balance scale, Car evaluation, Dermatology, Lymphography,* and *Nursery*), 9 datasets in term of the *PR* (i.e., *Lung, Soybean, Voting, Zoo, Balance scale, Car evaluation, Chess,*

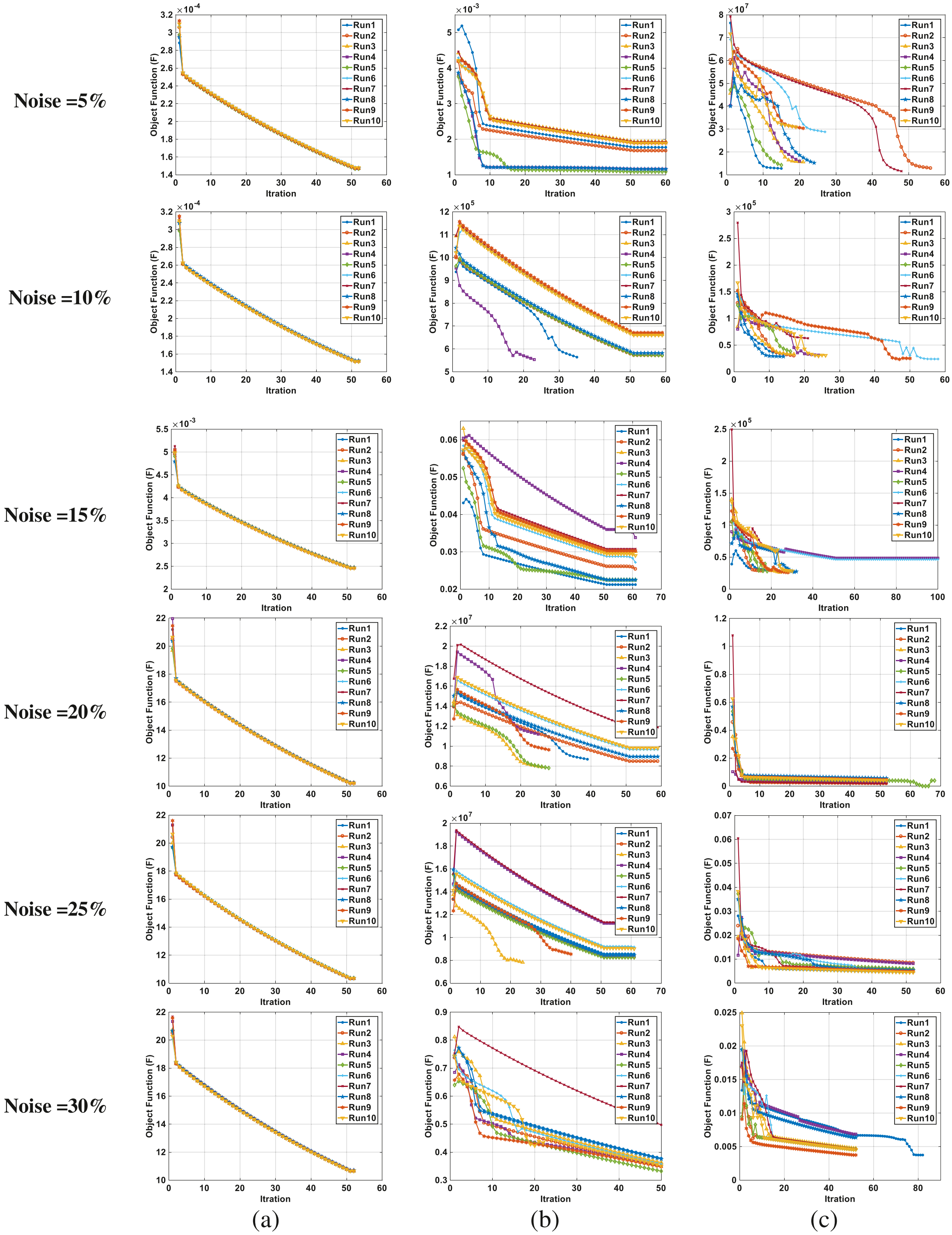


Fig. 6. Convergence speed of the objective function for different distance distances. (a) Hamming distance, (b) frequency probability-based distance, and (c) proposed distance function.

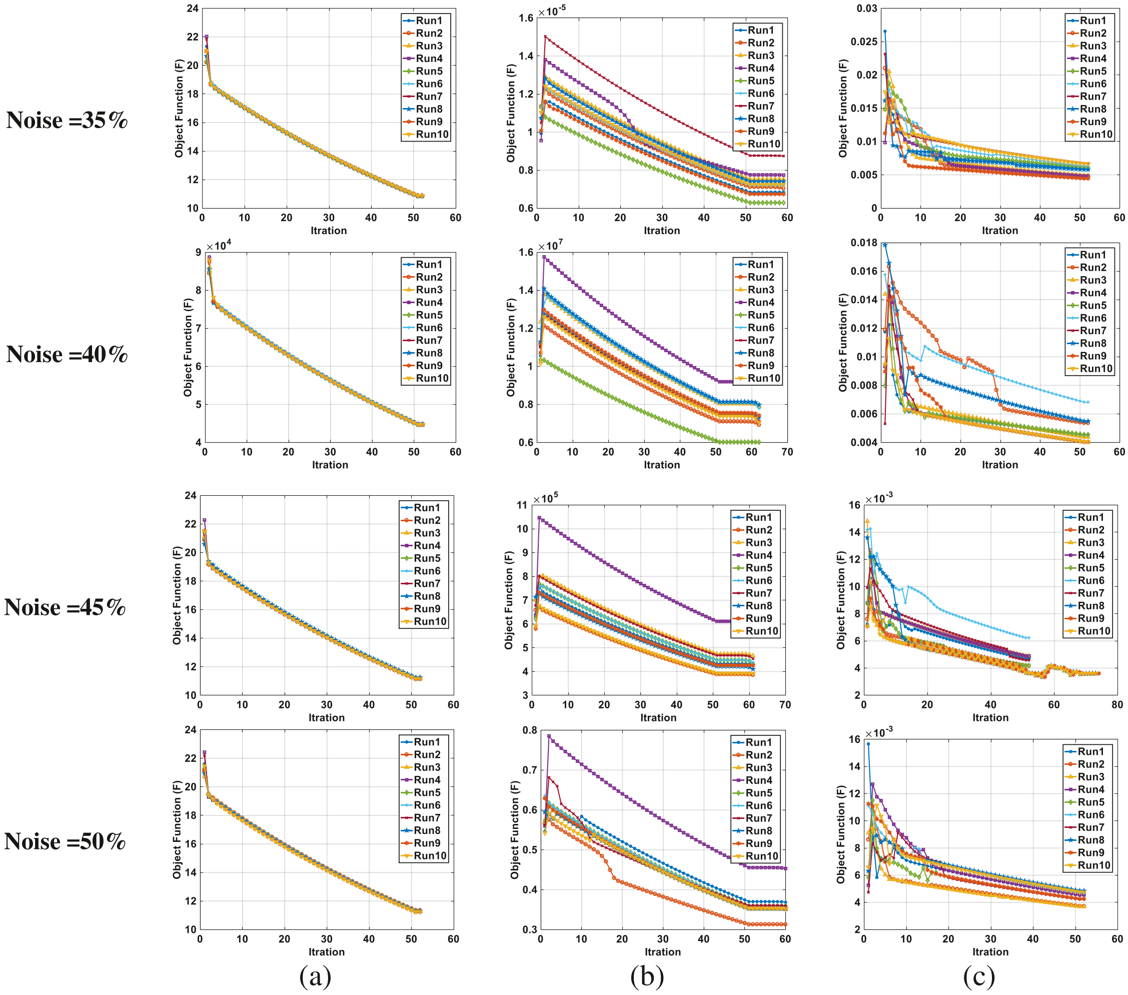


Fig. 6. Continued

Dermatology, and *Nursery*), and 6 datasets in term of the *RE* (i.e., *Lung, Soybean, Balance scale, Car evaluation, Dermatology*, and *Nursery*). Compared with the Cao [24] algorithm, the FKMAWCW algorithm performs better on 7 datasets in term of the *ACC* (i.e., *Lung, Voting, Balance scale, Car evaluation, Dermatology, Lymphography*, and *Nursery*), 7 datasets in term of the *PR* (i.e., *Lung, Voting, Balance scale, Chess, Dermatology, Lymphography*, and *Nursery*), and 6 datasets in term of the *RE* (i.e., *Lung, Zoo, Balance scale, Dermatology, Lymphography*, and *Nursery*). On the *Soybean* dataset, there is no difference in the performance of two algorithms. Regarding the Mod-2 [38] and Mod-3 [39] algorithms, the performance of the FKMAWCW algorithm is significantly better since there are: better results on all datasets in term of *ACC*, only 2 worse results on the *Car evaluation* and *Lymphography* in term of *PR*; and 3 worse results on the *Voting, Car evaluation*, and *Lymphography* datasets in term of *RE*. In summary, the proposed FKMAWCW, which takes advantage of [28] and the new distance metric for categorical data, can obtain better results than some existing initialization sensitivity reduction clustering methods.

In Table 7, only 5 datasets are selected to compare with the benchmark algorithms because these are the mutual datasets that were used to conduct the experiment on both the proposed and the benchmark algorithms. As shown in Table 7, the FKMAWCW has the best performance for *Lung* and *Voting* datasets. In the *Mushroom* and *Zoo* datasets, respectively, the Peng [27] and Jiang [23] algorithms have a higher performance than the proposed algorithm.

Table 7
Comparison of Proposed Method with Wu, Jiang, and Peng methods.

Datasets	Wu [58]	Jiang [23]	Peng [27]	FKMAWCW
Lung	0.5	0.625	0.5262	**0.6406**
	0.5584	0.6833	0.6017	**0.7423**
	0.5014	0.5932	0.5938	**0.6614**
Mushroom	0.8754	0.8941	**0.9185**	0.8182
	0. 9019	0.9138	**0.9108**	0.8566
	0.8709	0.8903	**0.9078**	0.8483
Soybean	**1**	**1**	**1**	**1**
	1	**1**	**1**	**1**
	1	**1**	**1**	**1**
Voting	0.8621	0.869	0.8671	**0.8922**
	0. 8571	0.863	0.8759	**0.9541**
	0.8755	**0.8811**	0.879	0.8387
Zoo	0.8812	**0.901**	0.8933	0.8218
	0.8702	**0.8906**	0.8911	0.7806
	0.6714	**0.8432**	0.8378	0.7289

4.6.3. Experiment 6: effect of attribute and cluster weighting on real-world datasets

To investigate the effect of the attribute and cluster weighting used in our approach on the final clustering quality, we com-

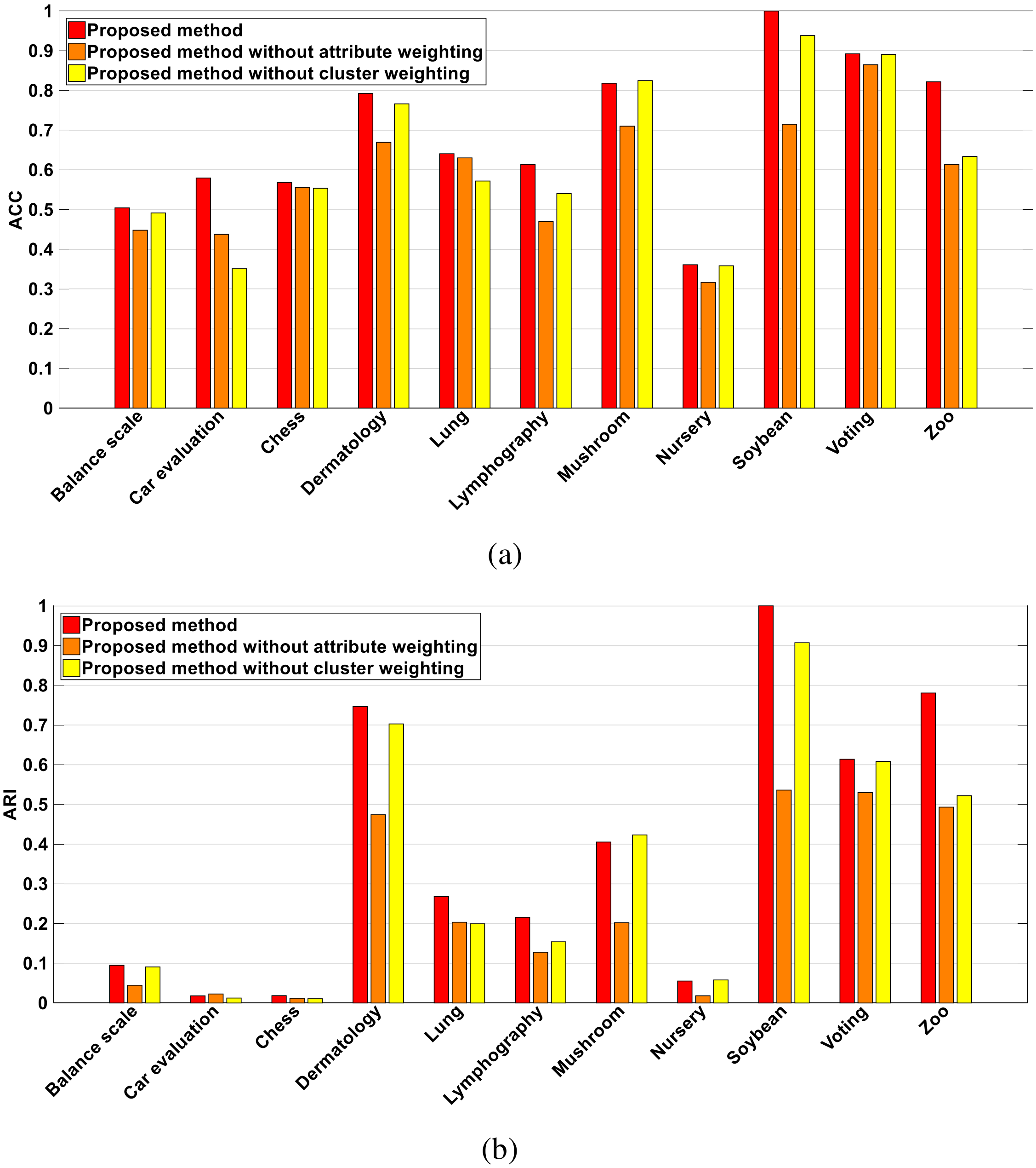
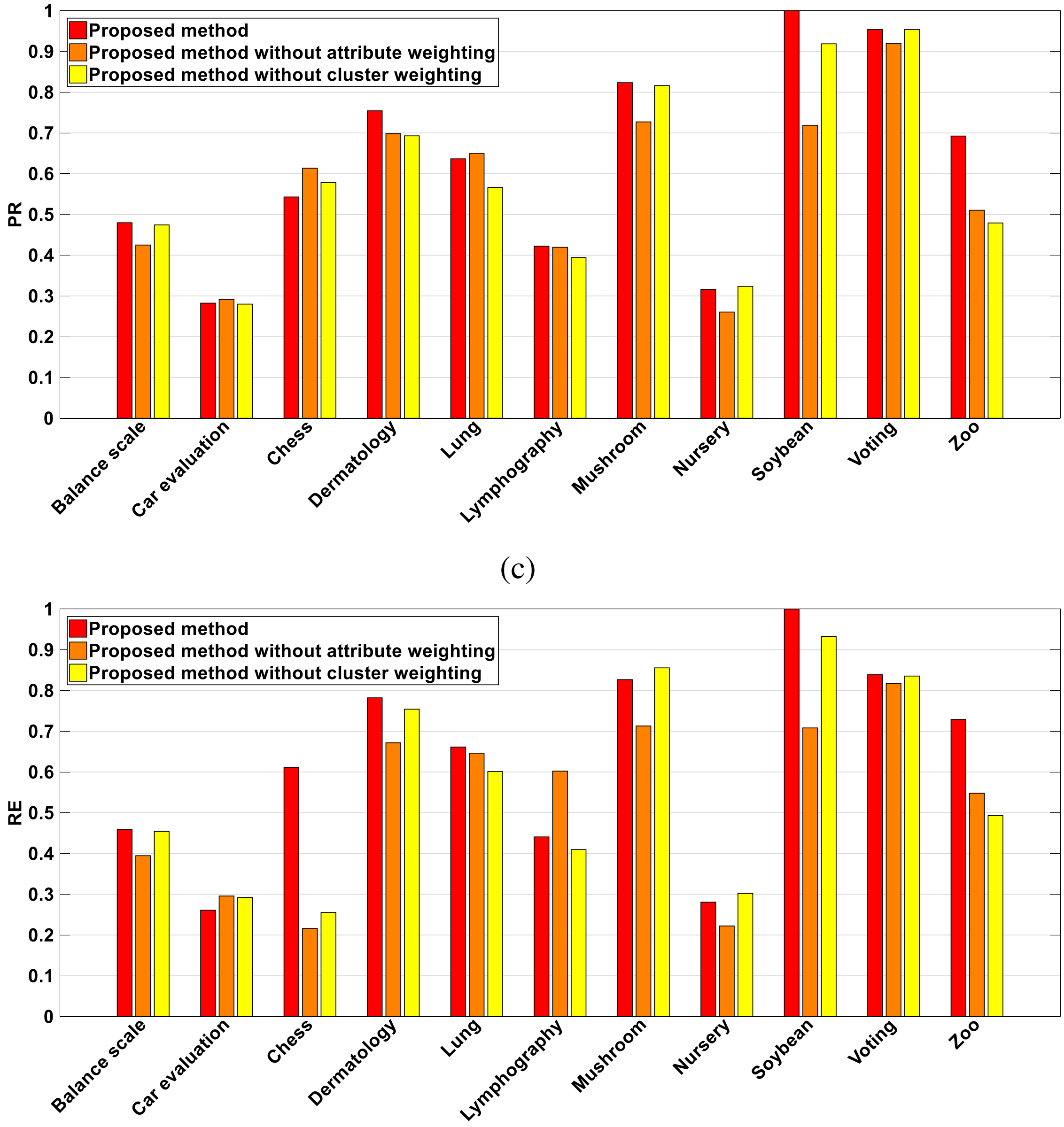


Fig. 7. Effect of weighting on the quality of real-world datasets clustering. (a) *ACC*, (b) *ARI*, (c) *PR* and (d) *RE.*

pare the proposed method once without attribute weighting and once without cluster weighting on 11 real-world datasets. In the case of without attribute weighting, we give the same weights to all attributes in each cluster. Also, in the case of without cluster weighting, we assign the same weights to all clusters. These weights are fixed during the algorithm run and are not updated. The compared results based on the *ACC*, *ARI*, *PR*, and *RE* are shown in Fig. 7.

As shown in Fig. 7, by using both weighting techniques (proposed method), the *ACC*, *ARI*, *PR*, and *RE* rates of the proposed approach, compared to the without attribute weighting mode, are improved by an average of 12%, 70%, 10%, and 27% on all testing datasets, respectively. For some datasets, such as *Soybean* and *Zoo*, the effect of the attribute weighting technique is more remarkable than other datasets. Also, compared to the without cluster weighting mode, the *ACC*, *ARI*, *PR*, and *RE* rates are improved by an average of 12%, 23%, 7%, and 17% on all testing datasets, respectively. These results show that in the proposed method, the weighting of attributes has a more significant effect on the formation of optimal clusters than the weighting of clusters. Although the overall results are better, for some datasets such as *Voting* and *Nursery*, the performance of the proposed method is worse than the without cluster weighting mode (for *PR* and *RE* criteria). This is because the used cluster weighting technique forms balance clusters in terms of SIWD. Therefore, this technique forms better clusters when there are natural groups with balanced SIWD in the dataset. So, this tactic may prove problematic when natural groups with different amounts of SIWD existing the dataset, a common scenario in practice, as it will hinder the clustering process from unveiling the true structure of the data.



(c)

(d)

Fig. 7. Continued

5. Conclusion

Much research has been carried out to design a clustering algorithm with weighted attributes. Extensive research is also carried out to make the FKM algorithm less sensitive to initialization. Most of these investigations have provided a solution to each of the existing challenges alone. Also, most of them use the Hamming similarity criterion, which is sensitive to noise.

In this study, we presented a new method for clustering categorical data based on the FKM clustering algorithm. During the clustering process, we used an attribute weighting scheme and a cluster weighting strategy to have better results. Used weights were calculated automatically and simultaneously during the learning process. Also, a new distance function based on the combination of the non-Euclidean distance and the frequency probability-based distance is used. Experimental results on large real-world datasets and a synthetic dataset showed that the proposed algorithm correctly assigns weight to each attribute due to its importance in each cluster, is not sensitive to the initialization and noise.

As future direction, it is of interest to investigate the application of the FKMAWCW in the clustering mixed data. We are also interested in the automatic determination of the number of clusters during the clustering process.

Declaration of Competing Interest

None declared under financial, general, and institutional competing interests.

CRediT authorship contribution statement

Amin Golzari Oskouei: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Mohammad Ali Balafar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing. **Cina Motamed:** Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – review & editing.

Acknowledgements

All persons who have made substantial contributions to the work reported in the manuscript, but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

Appendix

Theorem 1. *Let* $\boldsymbol{C}$, $\boldsymbol{W}$, *and* $\boldsymbol{Z}$ *be fixed,* $\boldsymbol{U}$ *is a strict local minimum of the* $\boldsymbol{F}(\boldsymbol{U})$ *if and only if* $\boldsymbol{U}$ *is calculated via Eq. (3).*

Proof. We use the Lagrangian multiplier technique to solve the following unconstrained minimization problem (see Eq. (A.1)).

$$G(\boldsymbol{U},\ \boldsymbol{\Lambda}) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}{}^{\alpha} \sum_{m=1}^{M} w_{km}^{q} z_k^p d^2(x_{nm}, c_{km}) - \sum_{n=1}^{N} \delta_n \left(\sum_{k=1}^{K} u_{nk} - 1 \right) \tag{A.1}$$

where $\boldsymbol{\Lambda} = [\delta_1, \delta_2,\ \ldots,\ \delta_N\ ]^{\boldsymbol{T}}$ is a vector containing the Lagrange multipliers corresponding to the constraints.

The optimization problem in Eq. (A.1) can be decomposed into N independent sub-minimization problems (see Eq. (A.2)).

$$G_n(\boldsymbol{U},\ \delta_n)\ = \sum_{k=1}^{K} u_{nk}{}^{\alpha} \sum_{m=1}^{M} w_{km}^{q} z_k^p d^2(x_{nm}, c_{km}) - \delta_n \left(\sum_{k=1}^{K} u_{nk} - 1 \right) \tag{A.2}$$

for $1 \le n \le N$

By setting the gradient of $G_n(\boldsymbol{U},\ \delta_n)$ to zero with respect to δ_n and u_{nk}, we obtain Eqs. (A.3 and A.4)

$$\frac{\partial\ G_n(\boldsymbol{U},\ \delta_n)}{\partial\ \delta_n}\ = -\left(\sum_{k=1}^{K} u_{nk} - 1 \right) = 0 \tag{A.3}$$

$$\frac{\partial\ G_n(\boldsymbol{U},\ \delta_n)}{\partial u_{nk}} = \alpha u_{nk}{}^{(\alpha-1)}\ z_k^p \sum_{m=1}^{M} w_{km}^{q} d^2(x_{nm}, c_{km}) - \delta_n = 0 \tag{A.4}$$

from (A.4), we obtain Eq. (A.5)

$$u_{nk} = \left[\frac{\delta_n}{\alpha\ z_k^p \sum_{m=1}^{M}\ w_{km}^{q}\ d^2(x_{nm}, c_{km})} \right]^{\frac{1}{\alpha-1}} \tag{A.5}$$

Substituting (A.5) into (A.3), we have Eq. (A.6)

$$\sum_{l=1}^{K} u_{nl} = \sum_{l=1}^{K} \left[\frac{\delta_n}{\alpha\ z_l^p \sum_{m=1}^{M}\ w_{lm}^{q}\ d^2(x_{nm}, c_{lm})} \right]^{\frac{1}{\alpha-1}} = 1 \tag{A.6}$$

It follows that Eq. (A.7)

$$\delta_n = \frac{\alpha}{\left[\sum_{l=1}^{K} \left[\frac{1}{z_l^p \sum_{m=1}^{M}\ w_{lm}^{q}\ d^2(x_{nm}, c_{lm})} \right]^{\frac{1}{\alpha-1}} \right]^{\alpha-1}} \tag{A.7}$$

Substituting Eq. (A.7) into Eq. (A.5), we obtain (3). This completes the proof.

Secondly, we can prove that Eq. (3) is the sufficient condition for the minimum of $\boldsymbol{F}(\boldsymbol{U})$.

Proof. If we show that the second partial derivative of Eq. (A.2) is positive, it can be proved that u_{nk} defined by Eq. (3) is a local minimum of Eq. (A.2); the derivative of Eq. (A.2) with respect to u_{nk} is as follows Eq. (A.8):

$$\frac{\partial}{\partial u_{nk}} \left(\frac{\partial\ G_n(\boldsymbol{U},\ \delta_n)}{\partial u_{nk}} \right) = \alpha(\alpha-1)\ u_{nk}{}^{\alpha-2}\ z_k^p \sum_{m=1}^{M} w_{km}^{q} d^2(x_{nm}, c_{km}) \tag{A.8}$$

Since we know $d^2(x_{nm}, c_{km}) \ge 0$, $w_{km} \ge 0$, $z_k \ge 0$, and $\alpha > 1$ are positive. So the Eq. (A.8) is positive definite. $\boldsymbol{F}(\boldsymbol{U})$ must have the minimum point, and Eq. (3) is sufficient for $\boldsymbol{U}$ to be a local minimum of $\boldsymbol{F}(\boldsymbol{U})$. Then Theorem 1 can be validated. This completes the proof. □

Theorem 2. *Let* $\boldsymbol{U}$, $\boldsymbol{C}$, *and* $\boldsymbol{Z}$ *be fixed,* $\boldsymbol{W}$ *is a strict local minimum of the* $\boldsymbol{F}(\boldsymbol{W})$ *if and only if* $\boldsymbol{W}$ *is calculated via Eq. (7).*

Proof. We use the Lagrangian multiplier technique to solve the following unconstrained minimization problem (see Eq. (A.9)).

$$G(\boldsymbol{W},\ \boldsymbol{\Lambda}) = \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}{}^{\alpha} \sum_{m=1}^{M} w_{km}^{q} z_k^p d^2(x_{nm}, c_{km}) - \sum_{k=1}^{K} \psi_k \left(\sum_{m=1}^{M} w_{km} - 1 \right) \tag{A.9}$$

where $\boldsymbol{\Lambda} = [\psi_1, \psi_2,\ \ldots,\ \psi_K\ ]^{\boldsymbol{T}}$ is a vector containing the Lagrange multipliers corresponding to the constraints.

The optimization problem in Eq. (A.9) can be decomposed into K independent sub-minimization problems (see Eq. (A.10)).

$$G_k(\boldsymbol{W},\ \psi_k)\ = \sum_{n=1}^{N} u_{nk}{}^{\alpha} \sum_{m=1}^{M} w_{km}^{q} z_k^p d^2(x_{nm}, c_{km}) - \psi_k \left(\sum_{m=1}^{M} w_{km} - 1 \right) \tag{A.10}$$

for $1 \le k \le K$

By setting the gradient of $G_k(\boldsymbol{W},\ \psi_k)$ to zero with respect to ψ_k and w_{km}, we obtain Eqs. (A.11 and A.12)

$$\frac{\partial\ G_k(\boldsymbol{W},\ \psi_k)}{\partial\ \psi_k}\ = -\left(\sum_{m=1}^{M} w_{km} - 1 \right) = 0 \tag{A.11}$$

$$\frac{\partial\ G_k(\boldsymbol{W},\ \psi_k)}{\partial\ w_{km}} = q\ w_{ks}^{q-1} z_k^p \sum_{n=1}^{N} u_{nk}{}^{\alpha}\ d^2(x_{ns}, c_{ks}) - \psi_k = 0 \tag{A.12}$$

from (A.12), we obtain Eq. (A.13)

$$w_{km} = \left[\frac{\psi_k}{q z_k^p \sum_{n=1}^{N}\ u_{nk}{}^{\alpha}\ d^2(x_{nm}, c_{km})} \right]^{\frac{1}{q-1}} \tag{A.13}$$

Substituting (A.13) into (A.11), we have Eq. (A.14)

$$\sum_{s=1}^{M} w_{ks} = \sum_{s=1}^{M} \left[\frac{\psi_k}{q z_k^p \sum_{n=1}^{N}\ u_{nk}{}^{\alpha}\ d^2(x_{ns}, c_{ks})} \right]^{\frac{1}{q-1}} = 1 \tag{A.14}$$

It follows that Eq. (A.15)

$$\psi_k = \frac{q}{\left[\sum_{s=1}^{M} \left[\frac{1}{z_k^p \sum_{n=1}^{N}\ u_{nk}{}^{\alpha}\ d^2(x_{ns}, c_{ks})} \right]^{\frac{1}{q-1}} \right]^{q-1}} \tag{A.15}$$

Substituting Eq. (A.15) into Eq. (A.13), we obtain Eq. (7). This completes the proof.

Secondly, we can prove that Eq. (7) is the sufficient condition for the minimum of $\boldsymbol{F}(\boldsymbol{W})$.

Proof. If we show that the second partial derivative of Eq. (A.10) is positive, it can be proved that w_{km} defined by Eq. (7) is a local minimum of Eq. (A.10); the derivative of Eq. (A.10) with respect to u_{nk} is as follows Eq. (A.16):

$$\frac{\partial}{\partial w_{km}}\left(\frac{\partial\ G_k(\boldsymbol{W},\ \psi_k)}{\partial w_{km}}\right) = \sum_{n=1}^{N} u_{nk}{}^{\alpha} q(q-1)\ w_{km}^{q-2}\ z_k^p d^2(x_{nm}, c_{km}). \tag{A.16}$$

Since we know $d^2(x_{nm} - c_{km}) \geq 0$, $w_{km} \geq 0$, $z_k \geq 0$, $q > 1$, and $q < 0$ are positive. So the Eq. (A.16) is positive definite. $\boldsymbol{F}(\boldsymbol{w})$ must have the minimum point, and Eq. (7) is sufficient for $\boldsymbol{W}$ to be a local minimum of $\boldsymbol{F}(\boldsymbol{W})$. Then Theorem 2 can be validated. This completes the proof. □

Theorem 3. *Let $\boldsymbol{U}$, $\boldsymbol{C}$, and $\boldsymbol{W}$ be fixed, $\boldsymbol{Z}$ is a strict local minimum of the $\boldsymbol{F}(\boldsymbol{Z})$ if and only if $\boldsymbol{Z}$ is calculated via Eq. (8).*

Proof. The proof of Theorem 3 is the same as the proof of Theorem 1 and 2. □

Theorem 4. *Let $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ be fixed, $\boldsymbol{C}$ is a strict local minimum of the $\boldsymbol{F}(\boldsymbol{C})$ if and only if $\boldsymbol{C}$ is calculated via Eq. (4).*

To obtain the c_{km} update formula, the objective function can be written as follows Eq. (A.17):

$$\begin{aligned} F(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W},\boldsymbol{Z}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}{}^{\alpha} \sum_{m=1}^{M} w_{km}^{q}\ z_k^p d^2(x_{nm}, c_{km}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk}{}^{\alpha} \hat{d}^2(x_{nm}, c_{km}), \end{aligned} \tag{A.17}$$

By incorporating w_{km}^q and z_k^p into the equation of computing distance $\hat{d}^2(x_{nm}, c_{km})$, the proposed objective function becomes the objective function of the FKM algorithm [13]. In the proposed algorithm and FKM, the $\hat{d}^2(x_{nm}, c_{km})$ and $u_{nk}{}^{\alpha}$ are in the range [0, 1]. Therefore, the function for updating c_{km} and proof of optimization in FKM can also be adopted for the proposed method.

References

[1] Niño-Adan I, Manjarres D, Landa-Torres I, Portillo E. Feature weighting methods: A review. Expert Syst Appl 2021;184:115424. doi:10.1016/j.eswa.2021.115424.

[2] Baradarani A, Wu QJ. Wavelet-based moving object segmentation. In: Pattern recognition: recent advances; 2010. p. 151.

[3] Hashemzadeh M, Zademehdi A. Fire detection for video surveillance applications using ICA K-medoids-based color model and efficient spatio-temporal visual features. Expert Syst Appl 2019;130:60–78. doi:10.1016/j.eswa.2019.04.019.

[4] Bouyer A, Hatamlou A. An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms. Appl Soft Comput 2018;67:172–82. doi:10.1016/j.asoc.2018.03.011.

[5] Kuwil FH, Atila Ü, Abu-Issa R, Murtagh F. A novel data clustering algorithm based on gravity center methodology. Expert Syst Appl 2020;156:113435. doi:10.1016/j.eswa.2020.113435.

[6] Asgari-Chenaghlu M, Feizi-Derakhshi M-R, farzinvash L, Balafar M-A, Motamed C. TopicBERT: a cognitive approach for topic detection from multimodal post stream using BERT and memory–graph. Chaos, Solitons Fractals 2021;151:111274. doi:10.1016/j.chaos.2021.111274.

[7] Bouyer A, Farajzadeh N. An optimized K-harmonic means algorithm combined with modified particle swarm optimization and cuckoo search algorithm. J Intell Syst 2020;29(1):1–18.

[8] Bouyer A. An optimized k-harmonic means algorithm combined with modified particle swarm optimization and Cuckoo Search algorithm. Found Comput Decis Sci 2016;41(2):99–121.

[9] Li Y, Li D, Wang S, Zhai Y. Incremental entropy-based clustering on categorical data streams with concept drift. Knowledge-Based Syst 2014;59:33–47. doi:10.1016/j.knosys.2014.02.004.

[10] Chen H, Chuang K, Chen M. On data labeling for clustering categorical data. IEEE Trans Knowl Data Eng 2008;20(11):1458–72. doi:10.1109/TKDE.2008.81.

[11] Kim M, Ramakrishna RS. Projected clustering for categorical datasets. Pattern Recognit Lett 2006;27(12):1405–17. doi:10.1016/j.patrec.2006.01.011.

[12] Nikzad-Khasmakhi N, Balafar M, Feizi-Derakhshi MR, Motamed C. ExEm: expert embedding using dominating set theory with deep learning approaches. Expert Syst Appl 2021;177:114913. doi:10.1016/j.eswa.2021.114913.

[13] Zhexue H, Ng MK. A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans Fuzzy Syst 1999;7(4):446–52. doi:10.1109/91.784206.

[14] Yuan F, Yang Y, Yuan T. A dissimilarity measure for mixed nominal and ordinal attribute data in k-Modes algorithm. Appl Intell 2020;50(5):1498–509. doi:10.1007/s10489-019-01583-5.

[15] Kuo RJ, Nguyen TPQ. Genetic intuitionistic weighted fuzzy k-modes algorithm for categorical data. Neurocomputing 2019;330:116–26. doi:10.1016/j.neucom.2018.11.016.

[16] Zhu S, Xu L. Many-objective fuzzy centroids clustering algorithm for categorical data. Expert Syst Appl 2018;96:230–48. doi:10.1016/j.eswa.2017.12.013.

[17] Saha A, Das S. Categorical fuzzy k-modes clustering with automated feature weight learning. Neurocomputing 2015;166:422–35. doi:10.1016/j.neucom.2015.03.037.

[18] DeSarbo WS, Carroll JD, Clark LA, Green PE. Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. Psychometrika 1984;49(1):57–78. doi:10.1007/bf02294206.

[19] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery 1998;2(3):283–304. doi:10.1023/A:1009769707641.

[20] Sivarathri S, Govardhan A. "Experiments on hypothesis" fuzzy K-means is better than K-means for clustering,. Int J Data Mining Knowledge Manage Process 2014(5):21 vol. 4.

[21] Stetco A, Zeng X-J, Keane J. Fuzzy C-means++: Fuzzy C-means with effective seeding initialization. Expert Syst Appl 2015;42(21):7541–8. doi:10.1016/j.eswa.2015.05.014.

[22] Rui X, Wunsch D. Survey of clustering algorithms. IEEE Trans Neural Netw 2005;16(3):645–78. doi:10.1109/TNN.2005.845141.

[23] Jiang F, Liu G, Du J, Sui Y. Initialization of K-modes clustering using outlier detection techniques. Inform Sci 2016;332:167–83. doi:10.1016/j.ins.2015.11.005.

[24] Cao F, Liang J, Bai L. A new initialization method for categorical data clustering. Expert Syst Appl 2009;36(7):10223–8. doi:10.1016/j.eswa.2009.01.060.

[25] Cao F, Liang J, Li D, Zhao X. A weighting k-modes algorithm for subspace clustering of categorical data. Neurocomputing 2013;108:23–30. doi:10.1016/j.neucom.2012.11.009.

[26] Bai L, Liang J, Dang C. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowledge-Based Syst 2011;24(6):785–95. doi:10.1016/j.knosys.2011.02.015.

[27] Peng L, Liu Y. Attribute weights-based clustering centres algorithm for initialising K-modes clustering. Cluster Comput 2019;22(3):6171–9. doi:10.1007/s10586-018-1889-5.

[28] Hashemzadeh M, Oskouei AGolzari, Farajzadeh N. New fuzzy C-means clustering method based on feature-weight and cluster-weight learning. Appl Soft Comput 2019;78:324–45. doi:10.1016/j.asoc.2019.02.038.

[29] Jia H, Cheung Y, Liu J. A new distance metric for unsupervised learning of categorical data. IEEE Trans Neural Netw Learn Syst 2016;27(5):1065–79. doi:10.1109/TNNLS.2015.2436432.

[30] Zhi X-b, Fan J-l, Zhao F. Robust local feature weighting hard c-means clustering algorithm. Neurocomputing 2014;134:20–9. doi:10.1016/j.neucom.2012.12.074.

[31] Bhopale AP, Tiwari A. Swarm optimized cluster based framework for information retrieval. Expert Syst Appl 2020;154:113441. doi:10.1016/j.eswa.2020.113441.

[32] Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl 2013;40(1):200–10. doi:10.1016/j.eswa.2012.07.021.

[33] Dinh D-T, Huynh V-N. k-PbC: an improved cluster center initialization for categorical data clustering. Appl Intell 2020;50(8):2610–32. doi:10.1007/s10489-020-01677-5.

[34] Supratid S, Kim H. Modified fuzzy ants clustering approach. Appl Intell 2009;31(2):122–34. doi:10.1007/s10489-008-0117-z.

[35] Wu S, Jiang Q, Huang JZ. A new initialization method for clustering categorical data. In: Pacific-Asia conference on knowledge discovery and data mining. Springer; 2007. p. 972–80.

[36] Ahmad A, Hashmi S. K-Harmonic means type clustering algorithm for mixed datasets. Appl Soft Comput 2016;48:39–49. doi:10.1016/j.asoc.2016.06.019.

[37] Khan SS, Ahmad A. Cluster center initialization algorithm for K-modes clustering. Expert Syst Appl 2013;40(18):7444–56. doi:10.1016/j.eswa.2013.07.002.

[38] Nguyen T-HT, Huynh V-N. A k-means-like algorithm for clustering categorical data using an information theoretic-based dissimilarity measure. In: FoIKS. Springer; 2016. p. 115–30.

[39] Nguyen T-HT, Dinh D-T, Sriboonchitta S, Huynh V-N. A method for k-means-like clustering of categorical data. J Ambient Intell Human Comput 2019. doi:10.1007/s12652-019-01445-5.

[40] Naouali S, Salem SBen, Chtourou Z. Clustering categorical data: A survey. Int J Inform Technol Decision Making 2020;19(01):49–96.

[41] Xing H-J, Ha M-H. Further improvements in Feature-Weighted Fuzzy C-Means. Information Sciences 2014;267:1–15. doi:10.1016/j.ins.2014.01.033.

[42] Hung W-L, Yang M-S, Chen D-H. Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation. Pattern Recognit Lett 2008;29(9):1317–25. doi:10.1016/j.patrec.2008.02.003.

[43] Jian Y. General C-means clustering model. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1197–211. doi:10.1109/TPAMI.2005.160.
[44] Bai L, Liang J. The k-modes type clustering plus between-cluster information for categorical data. Neurocomputing 2014;133:111–21. doi:10.1016/j.neucom.2013.11.024.
[45] Chan EY, Ching WK, Ng MK, Huang JZ. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognit 2004;37(5):943–52. doi:10.1016/j.patcog.2003.11.003.
[46] Bai L, Liang J, Dang C, Cao F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognit 2011;44(12):2843–61. doi:10.1016/j.patcog.2011.04.024.
[47] Bouguessa M. Clustering categorical data in projected spaces. Data Mining Knowledge Discovery 2015;29(1):3–38. doi:10.1007/s10618-013-0336-8.
[48] Chen L, Wang S, Wang K, Zhu J. Soft subspace clustering of categorical data with probabilistic distance. Pattern Recognit 2016;51:322–32. doi:10.1016/j.patcog.2015.09.027.
[49] Jia H, Cheung Y. Subspace clustering of categorical and numerical data with an unknown number of clusters. IEEE Trans Neural Netw Learning Syst 2018;29(8):3308–25. doi:10.1109/TNNLS.2017.2728138.
[50] Du H, Fang W, Huang H, Zeng S. MMDBC: density-based clustering algorithm for mixed attributes and multi-dimension data. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), 2018; 2018. p. 549–52. doi:10.1109/BigComp.2018.00093.
[51] Zhang R, Nie F, Guo M, Wei X, Li X. Joint learning of fuzzy K-means and nonnegative spectral clustering with side information. IEEE Trans Image Process 2019;28(5):2152–62. doi:10.1109/TIP.2018.2882925.
[52] Zhang R, Li X. Regularized regression with fuzzy membership embedding for unsupervised feature selection. IEEE Trans Fuzzy Syst 2020. doi:10.1109/TFUZZ.2020.3026834.
[53] Zhang R, Li X, Zhang H, Nie F. Deep fuzzy K-means with adaptive loss and entropy regularization. IEEE Trans Fuzzy Syst 2020;28(11):2814–24. doi:10.1109/TFUZZ.2019.2945232.
[54] Zhang R, Tong H, Xia Y, Zhu Y. Robust embedded deep K-means clustering. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 2019. p. 1181–90.
[55] Kvålseth TO. Measuring variation for nominal data. Bull Psycho Soc 1988;26(5):433–6. doi:10.3758/BF03334906.
[56] Tzortzis G, Likas A. The MinMax k-Means clustering algorithm. Pattern Recognit 2014;47(7):2505–16. doi:10.1016/j.patcog.2014.01.015.
[57] Liu J, Guo Y, Li D, Wang Z, Xu Y. Kernel-based MinMax clustering methods with kernelization of the metric and auto-tuning hyper-parameters. Neurocomputing 2019;359:173–84. doi:10.1016/j.neucom.2019.05.056.
[58] Wu S, Jiang Q, Huang JZ. A new initialization method for clustering categorical data. In: Zhou Z-H, Li H, Yang Q, editors. Advances in knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 972–80.
[59] Qian Y, Li F, Liang J, Liu B, Dang C. Space structure and clustering of categorical data. IEEE Trans Neural Netw Learn Syst 2016;27(10):2047–59. doi:10.1109/TNNLS.2015.2451151.
[60] [Online]. Available: http://archive.ics.uci.edu/ml/index.php.
[61] Hoffman M, Steinley D, Brusco MJ. A note on using the adjusted Rand index for link prediction in networks. Social Netw 2015;42:72–9. doi:10.1016/j.socnet.2015.03.002.