A Multivariate Fuzzy Weighted *K*-Modes Algorithm with Probabilistic Distance for Categorical Data

Ren-Jieh Kuo[1], Maya Cendana[1,*], Thi Phuong Quyen Nguyen[2] & Ferani E. Zulvia[1]

[1]Department of Industrial Management, National Taiwan University of Science and Technology, No. 43, Section 4, Kee-Lung Road, Taipei, Taiwan 106
[2]Faculty of Project Management, The University of Danang–University of Science and Technology, No. 54, Nguyen Luong Bang, Danang, Vietnam
*E-mail: d10901809@mail.ntust.edu.tw

Abstract. Data clustering is a data mining approach that assigns similar data to the same group. Traditionally, cluster similarity considers all attributes equally, but in real-world applications, some attributes may be more important than others. Therefore, this study proposes an algorithm that utilizes multivariate fuzzy weighting to demonstrate the varying importance of each attribute, using a Gini impurity measure for weight assignment. Additionally, the proposed algorithm implements probabilistic distance to reduce sensitivity to noise. Probabilistic distance offers more detailed information and better interpretation than Hamming distance, which ignores attribute positions. Probabilistic distance utilizes information about the attribute's position within and between clusters. This enhances clustering performance by creating clusters with more similar attributes. Therefore, the proposed Multivariate Fuzzy Weighted K-Modes with Probabilistic Distance for Categorical Data (MFWKM-PD) algorithm, based on the multivariate fuzzy K-modes algorithm, not only considers detailed membership calculations but also considers the varying contributions of attributes and their positions in distance calculation. This study evaluated the proposed MFWKM-PD using several benchmark datasets. The experiments validated that the proposed MFWKM-PD shows promising results compared to other algorithms in terms of accuracy, NMI, and ARI.

Keywords: *categorical data; fuzzy clustering; Gini impurity; MFWKM-PD; probabilistic distance.*

1 Introduction

Nowadays, the amount of digital data is growing exponentially and becoming more complex both on variety and sources. Therefore, the question of "how to effectively process this abundance of data to derive meaningful insight that can support decision makers" has become more crucial than in previous times when traditional data analysis methods were still effective. Data mining is an approach to explore and extract important information from data. It uses a variety of techniques, such as clustering and classification. Classification is a supervised learning method where the data is given predetermined labels. In contrast,

Received March 28, 2024, Revised July 13th, 2024, Accepted for publication July 26th, 2024

clustering groups similar data points together without using labels and is called an unsupervised learning technique. Clustering works without prior knowledge of specific labels to help make better decisions in many areas related to data segmentation and grouping. For example, customer segmentation on purchasing behavior, in which customers are grouped together according to similarities in their buying patterns. In molecular biology, cluster analysis can help scientists find the relationships between genes from large amounts of genomic data. Examples can also be found in other areas, such as medicine, biology, e-commerce [1-3], and so on.

Data clustering generates clusters according to the similarity of the data. Therefore, distance measurement is a critical issue in data clustering. The distance measurement used for calculating data similarity (or dissimilarity) should be defined based on the data type. The existing distance measurement, such as Hamming distance, has difficulty capturing the relation between the attributes since this metric only considers the dissimilarities between objects or objects with their centroids. Other distance measurements can be used, such as probabilistic distance based on kernel density estimation [4]. As opposed to the distance from the object to the centroid, probabilistic distance is specified as the distance from the object to the cluster.

Moreover, clustering methods can be divided into two categories based on how the clusters are produced, i.e., hierarchical and partitional clustering. Hierarchical clustering allows clusters to have sub-clusters, while partitional clustering only assigns each unlabeled object to one cluster [5]. Furthermore, in partitional clustering, clustering methods can be divided into hard and soft clustering, which differ in membership weight. Hard clustering, such as *K*-means [6] or *K*-modes [7], assigns each instance to a single cluster. In contrast, soft or fuzzy clustering, like in the fuzzy *c*-means algorithm (FCM) [8] and the fuzzy *K*-modes algorithm (FKM) [9], assigns each instance to multiple clusters with different membership values. Unlike FCM, FKM was developed to cluster the categorical data. However, there are some drawbacks when the number of values of each attribute increases. For instance, they treat all attributes as equally important, while in a real application, the contribution of the attributes can be different. The multivariate fuzzy *K*-modes algorithm (MFKM) is an algorithm that can handle differences between the values of each attribute of different clusters [10].

MFKM is an algorithm constructed by using the multivariate fuzzy c-means algorithm (MFCM) [11]. Both take a multivariate approach to determining the degree of membership. Certain methodologies employ a weighting technique. As an illustration, the weighted multivariate membership fuzzy c-means (WMFCM-M) algorithm and the multivariate membership integrated with weighted distances by cluster and variable (WMFCM-D) algorithm use the distance

parameter's weight to measure the variability dispersion within clusters, thereby facilitating the identification of cluster shapes [12]. Additional research has explored the utilization of impurity metrics, including entropy and Gini impurity, to enhance clustering performance [13-15]. Moreover, Kim [14] proposed a method for assigning weights that considers the attribute distributions within and between clusters. As a result, each attribute has a different weight and contribution that impacts the separation of objects into different clusters.

This study intended to develop a clustering algorithm that addresses the limitations of the above-mentioned algorithms by combining a weighting method and probabilistic distance, namely the multivariate fuzzy weighted K-modes algorithm with probabilistic distance for categorical data (MFWKM-PD). Three validity indexes—accuracy, normalized mutual information, and modified Rand index—were used to measure the performance of MFWKM-PD. This study implemented the proposed algorithm on several benchmark datasets. The results were compared with some clustering algorithms.

The remaining of this paper is organized as follows. Section 2 presents a brief review of the MFKM algorithm. Section 3 discusses the proposed MFWKM-PD algorithm. Section 4 presents the experimental results. Finally, the concluding remarks are given in Section 5.

2 Multivariate Fuzzy *K*-Modes Algorithm (MFKM)

The most popular fuzzy clustering algorithm to cluster categorical data is called the FKM algorithm; nevertheless, it reflects the same membership degrees for every attribute [9]. On the other hand, assigning different values for each attribute in different clusters is possible. Therefore, the FKM algorithm may be combined with the multivariate approach called multivariate fuzzy *K*-modes (MFKM) [10]. This algorithm was adopted from the multivariate fuzzy *c*-means algorithm for numerical data (MFCM) [12]. MFKM and MFCM are based on the same idea: to find a multivariate fuzzy partition and form multiple membership matrices.

MFKM has three important characteristics: (1) being able to determine each object's significance for a certain group based on each attribute; (2) being able to extract more information from data, which provides higher clustering quality; and (3) being a multivariate data analysis tool. Consequently, it produces different membership degrees, allowing it to handle the ambiguity of the data precisely [10].

In general, the MFKM algorithm has to represent the different memberships for each attribute in the different clusters and calculate the distance between clusters and centroids. The illustration of the membership distribution can be seen in

Table 1, where there are five records (*i*) with two attributes (*l*) and three clusters (*j*).

Table 1 Membership distribution example of MFKM algorithm.

	X1	**X2**	***j* = 1**		***j* = 2**		***j* = 3**		**Total**
			***l* = 1**	***l* = 2**	***l* = 1**	***l* = 2**	***l* = 1**	***l* = 2**	
***i* = 1**	1	2	0.08	0.22	0.20	0.23	0.03	0.24	1.00
***i* = 2**	1	1	0.37	0.30	0.03	0.11	0.15	0.04	1.00
***i* = 3**	4	5	0.01	0.02	0.10	0.11	0.17	0.59	1.00
***i* = 4**	9	8	0.32	0.09	0.09	0.32	0.02	0.16	1.00
***i* = 5**	8	7	0.07	0.13	0.59	0.16	0.01	0.04	1.00
			1.61		1.94		1.45		5.00

Given *X*, a set of *n* categorical objects. Each object x_i is defined as a set of *m* categorical attributes so that $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$. MFKM partitions *X* into *k* clusters denoted as $C_1, C_2, \dots, C_k$ by minimizing the objective function in Eq. (1).

$$J^M = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{l}^{m} u_{jli}^{\alpha} d(x_{il}, z_{jl}) \tag{1}$$

with subject to constraints in Eqs. (2) to (4):

$$0 \le u_{jli} \le 1, \quad \forall j = 1, \dots, k; i = 1, \dots, n; l = 1, \dots, m, \tag{2}$$

$$\sum_{j=1}^{k} \sum_{l=1}^{m} u_{jli} = 1, \quad \forall i = 1, \dots, n, \tag{3}$$

$$0 < \sum_{i=1}^{n} \sum_{l=1}^{m} u_{jli} < n, \quad \forall j = 1, \dots, k, \tag{4}$$

where α is a fuzziness component, and *k* is a predetermined number of clusters, while $U = [u_i]$ with (*i*=1, 2, …, *n*) is a multivariate fuzzy partition. *U* contains *n* multivariate membership matrix $u_i = [u_{jli}]$ where u_{jli} is the membership degree of the object *i* to cluster *j* on attribute *l*. z_{jl} is the centroid of cluster *j* on attribute *l*. $d(x_{il}, z_{jl})$ is the distance between x_{il} and its responding centroid z_{jl}. The matching distance measure is represented in Eq. (5) as follows:

$$d(x_{il}, z_{jl}) = \begin{cases} 0, & if\ x_{il} = z_{jl} \\ 1, & if\ x_{il} \neq z_{jl} \end{cases}, \tag{5}$$

where z_{jl} is the j^{th} element of Z_j and x_i is the i^{th} point of X_i.

The MFKM clustering algorithm is described as follows:

1. Initialization

   $(Fix\ c, 2 \le k \le n; fix\ m, 1 \le \alpha \le \infty; fix\ T; and\ fix\ \varepsilon > 0\ )$

   Randomly initialize $u_{jli} (j = 1, \dots, k; l = 1, \dots, m; i = 1, \dots, n)$ of pattern x belonging to cluster C_i on attributes *l* such that $u_{jli} \in [0,1]$ and $\sum_{j=1}^{k} \sum_{l=1}^{m} u_{jli} = 1$. Do t = 1.

2. Representation
   (Fix membership u_{jli} of pattern $x(x = 1, \dots, n)$ belonging to class C_i on the attributes $l(l = 1, \dots, m)$). Compute centroid z_j of cluster $C_j(i = 1, \dots, k)$ using Eq. (6):

$$r^* = \underset{1 \le r \le p_j}{arg\ max} \left\{ \sum_{l:x_{jl}=a_l^r} u_{jli}^{\alpha} \right\}. \tag{6}$$

3. Feature Membership
   (Fix centroid z_j of cluster $C_j(i = 1, \dots, k)$). Update the fuzzy membership degree, u_{jli}, using Eq. (7):

$$u_{jli} = \left[\sum_{p=1}^{k} \sum_{q=1}^{m} \left(\frac{d(x_{il}, z_{jl})}{d(x_{iq}, z_{pq})} \right)^{\frac{1}{\alpha - 1}} \right]^{-1}. \tag{7}$$

4. Stopping Criterion
   If $\left\| J_M^{t+1} - J_M^t \right\| \le \varepsilon \ or \ t > T$, go to Step (5); otherwise, update t = t + 1 and go Step 2.

5. Class Membership
   (Fix centroid z_j and membership $u_{jli} = (j = 1, \dots, k), (l = 1, \dots, m), (i = 1, \dots, n)$). Compute the fuzzy membership degree of object *x* belonging to cluster C_j, (γ_{ji}) using Eq. (8):

$$\gamma_{ji} = \sum_{l=1}^{m} u_{jli}. \tag{8}$$

3 Proposed Algorithm

In real-world applications, the attributes have different contributions; therefore, the importance of the attributes can be different. Using impurity metrics such as entropy and Gini impurity, each attribute can be weighed during the clustering process by considering the within-cluster and between-cluster relationship [14]. Moreover, the proposed algorithm uses probabilistic distance, which differs from possibilistic distance, instead of Hamming distance [4] to reduce the noise sensitivity.

Thus, the MFKM algorithm, which integrates the weighting method and probabilistic distance, becomes the multivariate fuzzy weighted *K*-modes algorithm with probabilistic distance (MFWKM-PD).

In Eq. (9), the objective function of the proposed MFWKM-PD is defined to find *U*, *Z*, and *W* to minimize *F(U, Z, W)* as follows:

$$F(U, Z, W) = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{l=1}^{m} u_{jli}^{\alpha} w_l d(x_{il}, z_{jl}) \tag{9}$$

Give X, a set of n categorical objects. Each object x_i is described by a set of m categorical attributes so that $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$, k is a predefined number of clusters denoted as $C_1, C_2, \ldots, C_k$, α is a fuzziness component, U as a multivariate fuzzy partition contains n multivariate membership matrix $u_i = [u_{jli}]$ where u_{jli} is the membership degree of the object i to cluster j on attribute l, Z represents cluster centers, and w_l is the weight of attribute l.

As w_l is a weight of attribute l, consider $w_{l,w}$ is a weight of attribute l based on within-cluster information as well as $w_{l,b}$ as a weight of attribute l based on between-cluster information. Thus, in Eq. (10), the weight of attribute l (w_l) is defined as follows:

$$w_l = aw_{l,w} + (1-a)w_{l,b}. \tag{10}$$

The adjusted parameter between two weight components $w_{l,w}$ and $w_{l,b}$ is defined as a, where $a(0 \le a \le 1)$. Since the probabilistic distance on the proposed algorithm uses Gini Impurity, the final weights of $w_{l,w}$ and $w_{l,b}$ are calculated based on Eqs. (11) and (12):

$$w_{l,w} = \frac{\sum_{j=1}^{k} \delta_j \left(1 - I_w(f_{lj})\right)}{\sum_{l=1}^{m} \sum_{j=1}^{k} \delta_j \left(1 - I_w(f_{lj})\right)}, \tag{11}$$

$$l_{l,b} = \frac{1 - I_b(g_l^r)}{\sum_{l=1}^{m} 1 - I_b(g_l^r)} \tag{12}$$

where $\sum_{l=1}^{m} w_{l,w} = 1$, For categorical attributes, δ_j is a weight that is proportional to the number of data samples that are part of the j-th cluster. It can be defined as $\delta_j = \frac{n_j}{n}$, where n_j is the number of objects in cluster C_j. Furthermore, the Gini impurities of each category attribute, which are based on the distribution of categories within the clusters, are presented in Eq. (13) as $I_w(f_{lj})$:

$$I_w(f_{lj}) = 1 - \sum_{r=1}^{h_l} \left(f_{lj}^r\right)^2 \tag{13}$$

where h_l is the set of categories in attribute l. The Gini impurity, $I_b(g_l^r)$, is a distribution of categories across clusters. It is defined in Eq. (14):

$$I_b(g_l^r) = 1 - \sum_{j=1}^{k} \left(g_{lj}^r\right)^2, \tag{14}$$

Regarding the distribution of categories, the distribution of categories of attribute l in cluster j can be considered as f_{lj} and f_{lj}^r as the frequency ratio of category r of attribute l in cluster j. Moreover, the distribution of attribute l across clusters is considered as g_l^r and g_{lj}^r as the frequency ratio of objects whose category value of the attribute l is a_l^r.

The fuzzy weighting method is used to compute f_{lj}^r and g_{lj}^r with membership degree u. Suppose the probability $x_{il} = a_l^r$ in the j-th cluster, and the probability

of the object with $x_{il} = a_l^r$ also belong to the *j*-th cluster, then f_{lj}^r and g_{lj}^r are defined in Eqs. (15) and (16):

$$f_{lj}^r = \frac{\sum_{x_{il}=a_l^r} u_{ji}}{\sum_{i=1}^n u_{ji}} \tag{15}$$

and

$$g_{lj}^r = \frac{\sum_{x_{il}=a_l^r} u_{ji}}{\sum_{x_{il}=a_l^r} \sum_{j=1}^k u_{ji}} \tag{16}$$

The proposed MFWKM-PD algorithm employs probabilistic distance instead of Hamming distance [4]. Rather than calculating the object-to-centroid distance, this method computes the object-to-cluster distance. Therefore, distance $d(x_{il}, z_{jl})$ in Eq. (9) becomes $d(x_{il}, C_j)$ which represents the distance of x_i to cluster *j* on attribute *l* as defined in Eq. (17):

$$d(x_{il}, Z_j) = \sum_{t \in h_l} \left[p(x_l = t|x_{il}) - p(x_l = t|C_j)\right]^2 \tag{17}$$

where the h_l is the set of categories. For instance, the attribute *d* takes $|h_l|$ discrete values. In the set, an arbitrary category is indicated by $t \in h_l$, where $l \in [1, |h_l|]$. Since all the categories in h_l are assumed to be independent of one another, Eq. (18) is used to estimate the probability $p(x_l = t|x_{il})$:

$$p(x_l = t|x_{il}) = I(t = x_{il}) \tag{18}$$

where $I(t = x_{il}) = 0 \; if \; x_{il} \neq t$, and otherwise, $I(t = x_{il}) = 0$

The kernel functions $p(x_l = t|C_j)$ determine the probability density of x_l, where $t \in \mathrm{h_l}$. It is defined in Eq. (19):

$$p(x_l = t|C_j) = \beta_j \frac{1}{|h_l|} + (1 - \beta_j) f_j(t) \tag{19}$$

The smoothing parameter, β_j, is called the bandwidth, and the frequency estimator of *t* in cluster $\mathrm{C_j}$ is formulated in Eq. (20):

$$f_j(t) = \frac{1}{n_j} \sum_{x_i \in j} I(t = x_{il}) \tag{20}$$

For categorical data, the optimal bandwidth of β_j lies within [0, 1] and can be calculated using Eq. (21):

$$\beta_j = \frac{\sum_{l=1}^m s_{jl}^2}{(n_j - 1) \sum_{l=1}^m \left(\frac{|h_l|-1}{|h_l|} - s_{jl}^2\right)} \tag{21}$$

The number of objects in cluster *j* is represented as n_j. The sample dispersion of categorical attributes is measured by the Gini diversity index S_{jl}^2, which is provided in Eq. (22):

$$S_{jl}^2 = 1 - \sum_{t \in h_l} [f_j(t)]^2 \tag{22}$$

Finally, the probabilistic distance is defined in Eq. (23):

$$d(x_{il}, C_j) = \sum_{t \in h_l} \left[I(t = x_{il}) - \beta_j \frac{1}{|h_l|} - (1 - \beta_j) f_j(t) \right]^2 \tag{23}$$

Therefore, to minimize *F(U, Z, W)*, the objective function of the proposed MFWKM-PD algorithm is presented in Eq. (24)"

$$F(U, Z, W) = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{l=1}^{m} u_{jli}^{\alpha} w_l d(x_{il}, C_j) \tag{24}$$

3.1 Updating Rules

Based on Eqs. (6-7), the cluster centroid z_{jl} of cluster j on attribute l and membership degree u_{jli} are updated using Eqs. (25) and (26) respectively:

$$z_{jl} = a_l^{(r)}, \; r = arg_{1 \le r \le |h_l|} max \left\{ \sum_{v: x_{vl} = a_l^r} u_{jvl}^m \right\} \tag{25}$$

and

$$u_{jli} = \left[\sum_{h=1}^{k} \sum_{s=1}^{m} \left(\frac{d(x_{il}, C_j)}{d(x_{is}, C_h)} \right)^{\frac{1}{\alpha - 1}} \right]^{-1} \tag{26}$$

The object-to-cluster distance $d(x_{il}, C_j)$ is calculated using Eq. (21); the weight attribute l (l_l) is calculated based on a fuzzy weighting method, f_{lj}^r and g_{lj}^r, using Eq. (10).

The proposed MFWKM-PD follows the following procedures:

Step 1: Initialization – Initialize multivariate fuzzy partition U^1 which contains n membership matrix $u_i = [u_{jli}]$, $(j = 1, \ldots, k; i = 1, \ldots, n; l = 1, \ldots, m)$ to satisfy the constraint. Generate the attribute weight W^1 with $w_l = 1/m$ for all attributes. Identify the centroid Z^1 such that cost $F(U^1, Z^1, W^1)$ is minimized. Set iteration t = 1.
For t = 1 to max iteration

Step 2: Fix Z^t and W^t and update U^{t+1}.
If $F(U^{t+1}, Z^t, W^t) = F(U^t, Z^t, W^t)$, then stop;
Else go to Step 3

Step 3: Fix W^t and U^{t+1} and update Z^{t+1}
If $F(U^{t+1}, Z^{t+1}, W^t) = F(U^{t+1}, Z^t, W^t)$, then stop
Else go to Step 4

Step 4: Fix U^{t+1} and Z^{t+1} and update W^{t+1}.
If $F(U^{t+1}, Z^{t+1}, W^{t+1}) = F(U^{t+1}, Z^{t+1}, W^t)$ or iteration t = max iteration, then stop.
Else return to Step 2
End for

4 Experimental Results

This study evaluated the proposed MFWKM-PD algorithm using three benchmark datasets from different fields, i.e., molecular biology, balance, and lymphography. Table 2 shows the details of the datasets, which have different numbers of instances (N), categorical attributes (d_c), and clusters (k) and were retrieved from the UCI machine learning repository.

Table 2 Benchmark datasets.

Datasets	N	d_c	k
Molecular Biology [16]	106	57	2
Balance [17]	625	18	3
Lymphography [18]	148	18	4

The molecular biology dataset (E. Coli promoter gene sequences) has 57 categorical attributes and each attribute has four levels of categories; the balance dataset has 18 categorical attributes and each attribute has five levels of categories. Moreover, the lymphography dataset, even though it has a similar number of attributes to the balance dataset, has various levels of categories. There are ten attributes comprising two levels, two attributes comprising three levels, four attributes comprising four levels, and two attributes comprising eight levels.

Accuracy (AC), normalized mutual information (NMI), and modified Rand index (ARI) were used to measure the performance of MFWKM-PD. The AC metric is defined in Eq. (29), where a_i represents the correctness of cluster assignment and n is the total number of objects, x_i.

$$AC = \frac{\sum_{i=1}^{n} a_i}{n}. \tag{29}$$

The NMI metric is the normalized version of mutual information (*MI*). Its value is in the interval [0;1], where 1 indicates the perfect labeling between the clustering result and the class label. NMI is given by Eq. (30):

$$NMI = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{30}$$

The entropies of X and Y are represented by $H(X)$ and $H(Y)$, respectively, and $I(X, Y)$ represents the mutual information between X and Y. Note that the attributes X and Y are random.

The ARI metric considers all cluster pairwise combinations and contains values in the interval [-1;1], where 1 indicates that the clusters are identical [19]. ARI is defined in Eq. (31):

$$ARI = \frac{RI - expected(RI)}{max(RI) - expected(RI)}, \tag{31}$$

where the Rand index (*RI*) measures the similarity between two cluster results by considering the number of pairs of elements belonging to the same or different clusters.

4.1 Results

In this experiment, the parameter settings for all algorithms were set as follows:

1. The fuzziness component α was 1.1 instead of 2 so the experiment gained better performance. Some papers performed better with a smaller α value [9,14].
2. The initial weight w_l was initialized as $1/m$ for each attribute l.
3. The optimal value of the balancing parameter a between two weight components for most cases was close to 0 or 1. Therefore, the a value was set as 0.1 [14].

This study compared the performance of MFWKM-PD algorithm with several clustering algorithms, i.e., fuzzy *K*-modes (FKM), fuzzy weighted *K*-modes (FWKM), fuzzy *K*-modes with probabilistic distance (FKM-PD), fuzzy weighted *K*-modes with probabilistic distance (FWKM-PD), multivariate fuzzy *K*-modes (MFKM), multivariate fuzzy weighted *K*-modes (MFWKM), multivariate fuzzy *K*-modes with probabilistic distance (MFKM-PD), and multivariate fuzzy *K*-modes with probabilistic distance (MFWKM-PD).

Each algorithm was run 30 times with the same initial centroids for all algorithms and the average performance was calculated based on the average values. Therefore, the proposed algorithm had slightly modified steps regarding this experiment, using the random initial centroids instead of random multivariate membership. Tables 3, 4, and 5 present a summary of the computational results in terms of accuracy, NMI, and ARI, respectively. The standard deviation and average clustering accuracy for all algorithms are displayed in Table 3.

Table 3 Average of accuracy (AC) and standard deviation (SD).

Datasets	Index	FKM	FWKM	FKM-PD	FWKM-PD	MFKM	MFWKM	MFKM-PD	MFWKM-PD
Molecular	AC	55.031	56.824	58.491	60.377	58.742	60.126	60.912	61.509
Biology	SD	3.471	3.697	4.439	5.920	4.765	5.997	7.526	8.573
Balance	AC	54.352	54.352	54.469	54.784	54.208	54.389	55.109	55.163
	SD	4.235	4.235	4.493	4.652	5.320	4.128	4.490	5.438
Lympho	AC	65.698	65.811	65.450	68.311	63.941	65.766	55.856	55.901
graphy	SD	4.942	5.027	6.514	6.706	5.937	6.578	0.648	0.613

The proposed algorithms won over the other algorithms for the molecular biology and balance datasets. This was because those algorithms have the same level of categories. The proposed algorithm also outperformed the other algorithms in terms of the average of NMI and ARI scores for two out of three datasets. However, the standard deviation of accuracy, NMI, and ARI scores obtained by the MFWKM-PD algorithm was not always better than that of the other algorithms. In general, FKM-based with weight and probabilistic distance (PD) algorithms had better results than FKM algorithms for all datasets. This proves that weight and probabilistic distance can improve FKM performance.

Table 4 Average of NMI and standard deviation (SD).

Datasets	Index	FKM	FWKM	FKM-PD	FWKM-PD	MFKM	MFWKM	MFKM-PD	MFWKM-PD
Molecular Biology	NMI	0.011	0.018	0.029	0.046	0.031	0.044	0.090	0.099
	SD	0.012	0.016	0.029	0.049	0.027	0.046	0.079	0.086
Balance	NMI	0.028	0.028	0.028	0.030	0.028	0.028	0.030	0.032
	SD	0.023	0.023	0.027	0.029	0.029	0.023	0.024	0.031
Lympho graphy	NMI	0.127	0.126	0.130	0.157	0.125	0.128	0.091	0.094
	SD	0.055	0.054	0.064	0.065	0.048	0.060	0.043	0.040

Table 5 Average of ARI and standard deviation (SD)

Datasets	Index	FKM	FWKM	FKM-PD	FWKM-PD	MFKM	MFWKM	MFKM-PD	MFWKM-PD
Molecular Biology	ARI	0.005	0.015	0.028	0.048	0.030	0.046	0.064	0.076
	SD	0.016	0.021	0.033	0.056	0.034	0.055	0.076	0.091
Balance	ARI	0.028	0.028	0.029	0.031	0.029	0.026	0.032	0.036
	SD	0.026	0.026	0.028	0.030	0.037	0.027	0.030	0.039
Lympho graphy	ARI	0.083	0.084	0.088	0.129	0.078	0.088	0.040	0.042
	SD	0.050	0.052	0.064	0.076	0.056	0.063	0.020	0.018

4.2 Statistical Test

This study also conducted a statistical test to analyze how significantly the proposed algorithm outperformed other algorithms. The null hypothesis for the statistical test was "*the proposed algorithm did not perform significantly differently from the other algorithms*". The significant level was set as 95% between the algorithms, which implied that α was set as 5%. Table 6 shows the *p*-value for all datasets in terms of AC, ARI, and NMI. The result shows that two of three of the datasets had a *p*-value less than 0.05. Therefore, the null hypothesis was rejected. This means that the proposed algorithm had significantly different results from the other algorithms for the molecular biology and lymphography datasets. Furthermore, the Bonferroni adjustment was implemented to make a pairwise comparison between the algorithms to show how they were grouped.

Table 6 The statistic results of all datasets (p-value).

Datasets	AC	ARI	NMI
Molecular Biology	0.000	0.000	0.000
Balance	0.975	0.988	0.999
Lymphography	0.000	0.000	0.000

The results of the Bonferroni pairwise comparisons can be seen in Tables 7, 8, and 9. Three datasets were used to examine each group in terms of AC, ARI, and NMI. For the Balance dataset, MFWKM-PD performed slightly better than the other algorithms for all performance metrics. Still, in the molecular biology dataset, MFWKM-PD had a significantly better performance compared to the FKM-based algorithm, except for the FKM algorithms with probabilistic distance and weight. As shown in Table 7, the FKM-based algorithms, especially FWKM-PD, had a significantly better result on the lymphography data, but the proposed algorithm did not perform well.

Table 7 Bonferroni pairwise comparison for AC.

Algorithm	N	Dataset: Molecular Biology			Dataset: Balance		Dataset: Lymphography		
		Mean	Grouping		Mean	Grouping	Mean	Grouping	
MFWKM-PD	30	61.5094	A		55.1627	A	55.9009		C
MFKM-PD	30	60.9119	A		55.1093	A	55.8559		C
FWKM-PD	30	60.3774	A		54.7840	A	68.3108	A	
MFWKM	30	60.1258	A		54.3893	A	65.7658	A	B
MFKM	30	58.7421	A	B	54.2080	A	63.9414		B
FKM-PD	30	58.4906	A	B	54.4693	A	65.4505	A	B
FWKM	30	56.8239	A	B	54.3520	A	65.8108	A	B
FKM	30	55.0314		B	54.3520	A	65.6982	A	B

Moreover, in terms of ARI and NMI, as shown in Tables 8 and 9, the MFWKM-PD algorithm performed significantly different from the other algorithms on the molecular biology dataset. In contrast, the FKM-based algorithms had lower performances. With the Balance dataset, all algorithms were only slightly different; therefore, there were no significant differences among their performances. On the other hand, with the lymphography dataset, even though the proposed algorithms had the lowest ARI value, FKM with weight and probabilistic distance had the best performance.

Table 8 Bonferroni pairwise comparison for ARI.

Algorithm	N	Dataset: Molecular Biology				Dataset: Balance		Dataset: Lymphography			
		Mean	Grouping			Mean	Grouping	Mean	Grouping		
MFWKM-PD	30	0.0762132	A			0.0359719	A	0.041697			C
MFKM-PD	30	0.0637444	A	B		0.0320827	A	0.040150			C
FWKM-PD	30	0.0479689	A	B	C	0.0310023	A	0.129494	A		
MFWKM	30	0.0460753	A	B	C	0.0262616	A	0.087775	A	B	
MFKM	30	0.0304458		B	C	0.0289763	A	0.078398		B	C
FKM-PD	30	0.0276115		B	C	0.0285140	A	0.088116	A	B	
FWKM	30	0.0146631			C	0.0282712	A	0.083588		B	C
FKM	30	0.0054138			C	0.0282712	A	0.082640		B	C

Table 9 Bonferroni pairwise comparison for NMI.

Algorithm	N	Dataset: Molecular Biology			Dataset: Balance		Dataset: Lymphography		
		Mean	Grouping		Mean	Grouping	Mean	Grouping	
MFWKM-PD	30	0.0992461	A		0.0315774	A	0.093781		B
MFKM-PD	30	0.0902149	A		0.0297288	A	0.090752		B
FWKM-PD	30	0.0461485		B	0.0299344	A	0.157197	A	
MFWKM	30	0.0438051		B	0.0282079	A	0.128155	A	B
MFKM	30	0.0309028		B	0.0281922	A	0.124625	A	B
FKM-PD	30	0.0294507		B	0.0284125	A	0.130448	A	B
FWKM	30	0.0178571		B	0.0277592	A	0.126420	A	B
FKM	30	0.0108815		B	0.0277592	A	0.127219	A	B

4.3 Computational Time

Table 10 presents the average computational time for all algorithms for each iteration. It reveals that the proposed MFWKM-PD algorithm needed more computational time for almost all datasets, except the lymphography dataset, where the MFKM algorithm with weight and probabilistic distance converged very fast..

Table 10 Computational time (in seconds).

Datasets	FKM	FWKM	FKM-PD	FWKM-PD	MFKM	MFWKM	MFKM-PD	MFWKM-PD
Molecular Biology	0.372	0.124	0.169	0.874	2.453	5.874	4.870	7.101
Balance	0.044	0.042	0.281	0.451	3.628	2.165	2.184	3.361
Lymphography	0.167	0.163	0.134	0.567	8.222	3.521	3.200	3.195

The molecular biology dataset having 57 categories compared to other datasets, which only have 18 categories, was the reason that the proposed algorithm needed more computational time. The impact of the number of attributes led to more computational time for the MFKM-based algorithm than the number of clusters.

Although the proposed MFWKM-PD did not have the lowest computational time, it could provide better performance in terms of AC, NMI, and ARI for the molecular biology and balance datasets

5 Conclusions

The MFWKM-PD algorithm, which is based on a multivariate approach, uses a Gini impurity weight that considers the distribution of attributes based on information from within and between clusters. This proposed algorithm also adopts probabilistic distance instead of Hamming distance to reduce the noise sensitivity. Evaluation results from three datasets showed that MFWKM-PD had better accuracy, ARI, and NMI performance, particularly with the molecular biology dataset. The proposed algorithm works well on datasets with the same category level regarding the number of categorical attributes and clusters; even compared to FKM-based algorithms with weight and probabilistic distance, it can perform well. This makes it useful for applications in molecular biology, such as identifying gene groups with similar functions or for clustering patients by molecular profiles.

However, the centroids being initialized randomly can lead to unstable results, as indicated by higher standard deviations. To address this, the initial centroids can be optimized using metaheuristic approaches, such as a genetic algorithm or a particle swarm optimization algorithm. Additionally, future studies can focus on determining the optimal number of clusters, which may be used to address the unsupervised problem in real-world applications.

Acknowledgment

The current study was financially supported by the National Science and Technology Council of the Taiwanese government under contract number: MOST 111-2221-E-011-076-MY3. It is appreciated.

References

[1] Aljobouri, H.K., Jaber, H.A., Koçak, O.M., Algin, O. & Çankaya, I., *Clustering fMRI Data with a Robust Unsupervised Learning Algorithm for Neuroscience Data Mining*, J Neurosci Meth, **299**, pp. 45-54, Apr. 1, 2018. DOI: 10.1016/j.jneumeth.2018.02.007.

[2] Bushel, P.R., Wolfinger, R.D. & Gibson, G., *Simultaneous Clustering of Gene Expression Data with Clinical Chemistry and Pathological Evaluations Reveals Phenotypic Prototypes*, Bmc Syst Biol, **1**, Feb. 23, 2007. DOI: 10.1186/1752-0509-1-15.

[3] Kuo, R.J., Potti, Y. & Zulvia, F.E., *Application of Metaheuristic based Fuzzy K-modes Algorithm to Supplier Clustering*, Comput. Ind. Eng., **120**, pp. 298-307, 2018.

[4] Chen, L.F., Wang, S.R., Wang, K.J. & Zhu, J.P., *Soft Subspace Clustering of Categorical Data with Probabilistic Distance*, PATTERN RECOGNITION, **51**, pp. 322-332, MAR. 2016. DOI: 10.1016/j.patcog.2015.09.027.

[5] Jain, A.K., Murty, M.N. & Flynn, P.J., *Data Clustering: A Review*, ACM Comput. Surv., **31**, pp. 264-323, 1999.

[6] MacQueen, J., *Some Methods for Classification and Analysis of Multivariate Observations*, 1967.

[7] Huang, Z., *Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values*, Data Mining and knowledge discovery, **2**(3), pp. 283-304, 1998.

[8] Bezdek, J.C., Ehrlich, R. & Full, W.E., *FCM: The Fuzzy C-means Clustering Algorithm*, Computers & Geosciences, **10**, pp. 191-203, 1984.

[9] Huang, J.Z. & Ng, M.K., *A Fuzzy K-modes Algorithm for Clustering Categorical Data*, IEEE Trans. Fuzzy Syst., **7**, pp. 446-452, 1999.

[10] Maciel, D.B.M., Amaral, G.J.A., Souza, R. & Pimentel, B.A., *Multivariate Fuzzy K-modes Algorithm*, Pattern Analysis and Applications, **20**, pp. 59-71, 2017.

[11] Pimentel, B.A. & Souza, R.M.C.R.D., *A Multivariate Fuzzy C-means Method*, Appl. Soft Comput., **13**, pp. 1592-1607, 2013.

[12] Pimentel, B.A. & de Souza, R.M., *Multivariate Fuzzy C-means Algorithms with Weighting*, Neurocomputing, **174**, pp. 946-965, 2016

[13] Cao, F., Liang, J., Li, D. & Zhao, X., *A Weighting K-modes Algorithm for Subspace Clustering of Categorical Data*, Neurocomputing, **108**, pp. 23–30, 05/01 2013. DOI: 10.1016/j.neucom.2012.11.009.

[14] Kim, K., *A Weighted K-modes Clustering using New Weighting Method based on Within-cluster and Between-cluster Impurity Measures*, J. Intell. Fuzzy Syst., **32**, pp. 979-990, 2017.

[15] Sen, P., *Gini Diversity Index, Hamming Distance, and Curse of Dimensionality*, Metron - International Journal of Statistics, LXIII, pp. 329-349, 02/01 2005.

[16] Harley, C., Reynolds, R. & Noordewier, M. *Molecular Biology (Promoter Gene Sequences)*, UCI Machine Learning Repository, DOI: 10.24432/C5S01D.

[17] Siegler, R. *Balance Scale*, UCI Machine Learning Repository, DOI: 10.24432/C5488X.

[18] Zwitter, M.A.S.M., *Lymphography*, UCI Machine Learning Repository, DOI: 10.24432/C54598.

[19] Hubert, L.J. & Arabie, P., *Comparing Partitions*, Journal of Classification, **2**, pp. 193-218, 1985.