Interval Kernel Fuzzy C-Means - Particle Swarm Optimizer with Two Differential Mutations (IKFCM-PSOTD) for Incomplete Data Clustering

Muhaimin Ilyas[1], Syaiful Anam[2,*], and Trisilowati Trisilowati[3]
[1,2,3] Brawijaya University, Malang, Indonesia
*syaiful@ub.ac.id

Abstract. Data processing and optimization are two major challenges in data analysis such as clustering. In practice, data often contains missing values that must be handled appropriately. This research provides an innovative approach to clustering incomplete data using Interval Kernel Fuzzy C-Means (IKFCM) technique optimized by Particle Swarm Optimizer with Two Differential Mutations (PSOTD). The proposed method solves the problem of incomplete data clustering by introducing interval imputation that allows more flexible handling of missing values. The interval value is obtained using the nearest neighbor method, which provides information on the similarity between an incomplete datum and its neighbors. Then, Kernel Fuzzy C-Means (KFCM) is applied due to its efficacy in handling outlier data and improving the accuracy of data representation in a high-dimensional feature space. In addition, Particle Swarm Optimization (PSO) algorithm adopting Differential Evolution (DE) technique with two different mutations is used to optimize the clustering algorithm to obtain better results. The addition of DE technique to PSO is believed to enhance global search capability and search efficiency. The proposed method is evaluated using the Partition Coefficient Index (PCI), Partition Entropy Index (PEI), and Apparent Error Rate (APER). Experimental results show that the proposed approach can efficiently cope with data incompleteness, resulting in more accurate clustering results than the comparison algorithms. In addition, the PSO algorithm enhanced with differential mutation makes the clustering result achieve a better solution.

Keywords: Incomplete data clustering, Interval imputations, Kernel based fuzzy c-means, Particle swarm optimization, Differential evolution

1 Introduction

Clustering is defined as the process of dividing a dataset into specific clusters according to the similarity of objects in the data. Objects that have similarities to each other will be collected in the same cluster. Research on clustering has been conducted frequently since the late 1980s and 1990s [1]. In fact, the monograph [2] in 1963 became the trigger for global research on clustering techniques. Clustering methods are useful for finding previously unknown groups in unlabeled data. Now, clustering has been widely used in various research sectors, including sports [3], medical [4],

K. Dharmawan and N. A. Sanjaya ER (eds.), *Proceedings of the First International Conference on Applied Mathematics, Statistics, and Computing (ICAMSAC 2023)*, Advances in Computer Science Research 110,
https://doi.org/10.2991/978-94-6463-413-6_18

economics [5], social networks [6], image [7], and others. One common type of clustering method is a fuzzy-based clustering algorithm.

Fuzzy sets were first presented by Zadeh in 1965 [8] as a development of crisp sets. The set is defined by x, which has a value between 0 and 1 Fuzzy C-Means (FCM) is a clustering algorithm that uses fuzzy concepts to handle uncertainty in data. FCM was introduced by Bezdek in 1973 [9] by improving the K-means algorithm proposed by Llyod [10]. FCM has been widely used for various studies such as education [11], medical [12], image [13], farm [14] and usually applied to complete datasets. Unfortunately, some problems are often found in the clustering process, one of which is the missing value problem that causes incomplete datasets. Generally, this may occur due to several factors, such as errors in data input. Research shows that at least 5% of the data are missing [15]. For example, suppose a datum $x_i = (3, ?, 5)$ with a value in the second variable x_{i2} is a missing value; if the value is not appropriately handled, the resulting conclusion may be inaccurate. In addition, the FCM algorithm cannot be used directly to cluster incomplete data [16].

Several studies were conducted so that the FCM algorithm can be used on incomplete datasets. Hathaway and Bezdek proposed four strategies for handling incomplete data using the FCM algorithm, namely Whole Data Strategy (WDS), Partial Data Strategy (PDS), Optimal Completion Strategy (OCS), and Nearest Prototype Strategy (NPS) [17]. The first strategy (WDS) removes all datums that have missing values. This strategy may be recommended if the dataset contains a small number of missing values. The second strategy (PDS) approaches the missing values using partial distance equation instead of Euclidean distance in the standard FCM algorithm. The third (OCS) and fourth (NPS) strategies treat missing values as values to be imputed. OCS imputes them with a value that is considered optimal for obtaining a good estimate. Whereas NPS fills the missing value with a value that corresponds to the closest prototype. Zhang and Li [18] developed another approach for clustering incomplete data by incorporating information from the prior distribution of missing values. This information is then used in the FCM algorithm by involving the maximum expectation criterion. Balqis and Sadoq [19] developed a fuzzy self-organizing map algorithm by utilizing the OCS strategy at each iteration. Zhang and Chen [20] also developed research on the OCS strategy performed on the kernel-based FCM algorithm (KFCM). Kernel is a proven technique for dealing with non-linear data, as it supports a better approach to the data structure. The proposed algorithm is able to perform better clustering than the standard FCM algorithm. Rodrigues et al. [21] proposed the VKFCM-K-LP algorithm, which is a KFCM algorithm that considers metric kernelization with locally adaptive distance. This research adopts the WDS, PDS, and OCS strategies.

Intervals are one of the strategies often used to deal with incomplete datasets. Li et al. [16] proposed a new approach to estimate missing values using interval values. The value is obtained using the concept of nearest neighbor, which is based on the partial distance between datums. Interval can provide greater flexibility in estimated value. Khan et al. [22] proposed a missing value approach by considering the selection of shorter intervals, which proved to be more effective than the use of long intervals. An FCM algorithm with an interval re-construction strategy for incomplete data was

developed by Zhang et al. [23]. The proposed algorithm is also accompanied by optimization to improve clustering performance. The literature reviews above shows that strategies for handling incomplete datasets are very important, including interval strategies.

FCM algorithms, whether kernel-based or not, have the problem of being sensitive to the initial centroid [24]. Heuristic algorithms are often used to overcome this problem. Particle Swarm Optimization (PSO) is one of the most widely used heuristic algorithms because it is simple and easy to use. This algorithm was first introduced by Eberhart and Kennedy in 1995 [25]. Cura [26]conducted research on the use of PSO for clustering problems in general. [27], [28], and [29]proposed the use of PSO for FCM algorithm optimization to more easily obtain a global solution. Praseda and Shivakumar [30] proposed the Hybrid Kernel Distance-Based Possibilistic Fuzzy Local Information C-Means (HKD-PFLICM) algorithm for clustering customer churn data. The proposed algorithm is able to produce up to 95% accuracy. Salleh and Samat [31] combined PSO and FCM (FCMPSO) algorithms to handle missing values in heart disease datasets. The results showed that filling in missing values using FCMPSO obtained superior results compared to other methods, such as mean and median. Even so, the PSO algorithm also has a weakness, which is easily trapped in the local optimal solution [32]. Particle Swarm Optimizer with Two Differential Mutations (PSOTD) is one of the modifications of the PSO algorithm combined with the Differential Evolution (DE) algorithm [32]. DE is a simple, popular search algorithm and has been widely applied to other variants of the PSO algorithm, such as DEPSO (Differential Evolution Particle Swarm Optimization) [33] and PSOCA (Particle Swarm Optimization and Cultural Algorithm) [34]. DE has three common operators, namely, crossover, mutation, and selection. PSOTD adopts DE with the addition of two different mutation operations. The results show that PSOTD can improve standard PSO performance.

This article develops the Interval Kernel Fuzzy C-Means (IKFCM) algorithm proposed by [16], which has been proven to be able to perform clustering well on incomplete datasets. To improve the performance of the IKFCM algorithm, we add an optimization technique using the PSOTD algorithm [32] with the hope that the resulting performance will be better. There will be comparisons between the proposed algorithm and several other algorithms, such as mean KFCM, median KFCM, OCSKFCM, IKFCM, and IKFCMPSO.

2 Research Methods

This section outlines the techniques used in formulating the proposed methodology for incomplete data clustering. The approach taken is the optimization of the Interval Kernel Fuzzy C-Means (IKFCM) algorithm using the Particle Swarm Optimizer with Two Differential Mutations (PSOTD). Model performance is evaluated based on Partition Entropy Index (PEI), Partition Coefficient Index (PCI), and Apparent Error Rate (APER).

2.1 Interval Kernel Fuzzy C-Means

Given an incomplete dataset $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\} \subset \boldsymbol{R^s}$, if $\boldsymbol{x_b}$ is an incomplete datum, then the nearest neighbor relationship of $\boldsymbol{x_b}$ can be obtained using the partial distance approach [16]. The partial distance of $\boldsymbol{x_b}$to $\boldsymbol{x_p}$ can be found using equation (1).

$$d_{pb} = \sqrt{\frac{s}{\sum_{j=1}^{s} I_j} \sum_{j=1}^{s} \left(x_{jb} - x_{jp}\right)^2 I_j}, \tag{1}$$

where x_{jb} and x_{jp} are the j th attributes of $\boldsymbol{x_b}$ and $\boldsymbol{x_p}$, then

$$\boldsymbol{I_j} = \begin{cases} 1 & if\ \mathrm{x_{jb}}, x_{jp} \in \boldsymbol{x_p} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $p, b = 1,2, \dots, n, j = 1,2, \dots, s$, and $\boldsymbol{x_p}$ is a datum that has a value. Based on the concept of nearest neighbor, a datum containing missing values and its neighbors have similar features. Therefore, the range of missing values x_{jb} can be estimated using the minimum and maximum values of q nearest neighbors and can be written as the interval $\left[x_{jb}^{-}, x_{jb}^{+}\right]$. Suppose $\overline{\boldsymbol{X}} = \{\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, \dots, \overline{\boldsymbol{x}}_n\}$ is an interval-valued dataset,, where $\overline{\boldsymbol{x}}_k = \left(\overline{x}_{j1}, \overline{x}_{j2}, \dots, \overline{x}_{js}\right), \forall j, k: \ \overline{x}_{jk} = \left[\overline{x}_{jk}^{-}, \overline{x}_{jk}^{+}\right]$. The objective function of interest in the IKFCM algorithm is:

$$J(\boldsymbol{U}, \bar{V}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| \Phi\left(\overline{\boldsymbol{x}}_j\right) - \Phi\left(\overline{\boldsymbol{v}}_i\right) \right\|^2, \tag{3}$$

where

$$\left\| \Phi\left(\overline{\boldsymbol{x}}_j\right) - \Phi\left(\overline{\boldsymbol{v}}_i\right) \right\|^2 = 2\left(1 - K\left(\overline{\boldsymbol{x}}_j, \overline{\boldsymbol{v}}_i\right)\right) \tag{4}$$

The necessary conditions to minimize equation (3) with constraints $\sum_{i=1}^{c} u_{ij} = 1; j = 1,2, \dots, n$ can be obtained by the Lagrange multiplier, which results in equations (5) and (6).

$$u_{ij} = \left[\sum_{t=1}^{c} \left(\frac{\left\| \Phi\left(\overline{\boldsymbol{x}}_j\right) - \Phi\left(\overline{\boldsymbol{v}}_i\right) \right\|^2}{\left\| \Phi\left(\overline{\boldsymbol{x}}_j\right) - \Phi\left(\overline{\boldsymbol{v}}_t\right) \right\|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \quad \text{where } i = 1,2, \dots, c; j = 1,2, \dots, n. \tag{5}$$

$$\overline{\boldsymbol{v}}_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} K\left(\overline{\boldsymbol{x}}_j, \overline{\boldsymbol{v}}_i\right) \overline{\boldsymbol{x}}_j}{\sum_{j=1}^{n} u_{ij}^{m} K\left(\overline{\boldsymbol{x}}_j, \overline{\boldsymbol{v}}_i\right)} \text{ where } i = 1,2, \dots, c, \tag{6}$$

where

$$K\left(\overline{\boldsymbol{x}}_j, \overline{\boldsymbol{v}}_i\right) = exp\left(- \frac{\left(\overline{\boldsymbol{x}}_j^{-} - \overline{\boldsymbol{v}}_i^{-}\right)^T \left(\overline{\boldsymbol{x}}_j^{-} - \overline{\boldsymbol{v}}_i^{-}\right) + \left(\overline{\boldsymbol{x}}_j^{+} - \overline{\boldsymbol{v}}_i^{+}\right)^T \left(\overline{\boldsymbol{x}}_j^{+} - \overline{\boldsymbol{v}}_i^{+}\right)}{\sigma^2} \right) \tag{7}$$

2.2 Particle Swarm Optimizer with Two Differential Mutations (PSOTD)

PSO algorithm only uses the global and personal best position iteratively, so the algorithm is less balanced in exploration and exploitation ability in the search space region. The addition of crossover, mutation, and selection principles from the DE algorithm can help PSO overcome these problems. [32] proposed the PSOTD algorithm,

which is a modification of the PSO algorithm by adding DE operators to improve the position of the personal best. This algorithm has two mutation operators that have different characteristics; one excels in exploration and the other excels in exploitation. PSOTD still has the principle of speed and position, like the PSO algorithm. The velocity and position equations are shown in equations (8) and (9).

$$v_i^{t+1} = \omega v_i^t + cr\left(p_{iwin_i^t} - x_i^t\right) \tag{8}$$

$$x_{id}^{t+1} = x_i^t + v_i^{t+1}, \tag{9}$$

p_{iwin} is the main focus built by PSOTD to find the global solution, and this is what distinguishes it from standard PSO. P_{iwin} determines the direction of solution search with the help of the DE algorithm to prevent premature convergence and improve search performance. The following is an explanation of the three DE principles used in PSOTD algorithm.

1. Mutation. There are two DE mutation operators used.
   a. DE/rand/1

$$m_{i,d} = p_{r1,d} + F\left(p_{r2,d} - p_{r3,d}\right) \tag{10}$$

   b. DE/current-to-best/1

$$m_{i,d} = p_{i,d} + F\left(pg_{i,d} - p_{i,d}\right) + F\left(p_{r1,d} - p_{r2,d}\right), \tag{11}$$

   where $r_1, r_2,$ dan r_3 are integer random numbers in the interval $[1, n]$, while F is a positive control parameter. p and pg denote personal best and global best.

2. Crossover. This operation aims to swap some components of $\boldsymbol{m}$ with $\boldsymbol{p}$ to form a new vector $\boldsymbol{q}$.

$$q_{\boldsymbol{i,d}} = \begin{cases} m_{i,d} & \text{if } r_2 \le CR \text{ atau } d = d_{rand} \\ p_{i,d} & \text{otherwise,} \end{cases} \tag{12}$$

   where r_2 is a random number in the interval [0,1]. *CR* is the Crossover Rate. d_{rand} is an integer random number in $[1, d]$ which guarantees that at least one element in vector $\boldsymbol{q}$ differs from vector $\boldsymbol{p}$.

3. Selection. This operation aims to ensure that vectors that have better fitness values will be retained in the next generation.

$$\boldsymbol{P_{iwin_i}} = \begin{cases} \boldsymbol{q_i} & \text{if } f(\boldsymbol{q_i}) \le f(\boldsymbol{p_i}) \\ \boldsymbol{p_i} & \text{otherwise.} \end{cases} \tag{13}$$

Thus, the particles will continue to evolve with each generation. In addition, the PSOTD algorithm also divides the swarm into two sub-swarms. The division is done iteratively with the following equation:

$$N = N_1 + N_2 \tag{14}$$

$$N_2 = \left\lfloor \frac{t}{Maxit} N \right\rfloor, \tag{15}$$

where $Maxit$ is the maximum iteration and $\lfloor \quad \rfloor$ is the floor operator. N_1 and N_2 denote the number of swarms in sub-swarm1 and sub-swarm2.

2.3 Proposed Method: IKFCM-PSOTD

The structure of the proposed method is shown in Figure 1. There are four main steps: 1) Imputation using intervals; 2) IKFCM-PSOTD; 3) Performance evaluation.

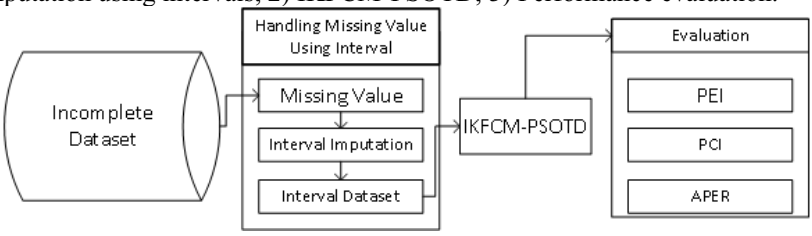

Fig. 1. Structure of Proposed Method

The detailed procedure of the IKFCM-PSOTD algorithm for incomplete data can be seen in the following explanation:

Step 1 : Find as many as q nearest neighbors of each datum containing missing values using equation (1), then transform the incomplete dataset into a complete interval-valued dataset.

Step 2 : Initialize IKFCM and PSOTD Parameters

Step 3 : Initialize the initial membership matrix $\boldsymbol{U}(0)$, the initial centroid matrix $\boldsymbol{X}(0)$ as the initial position in PSOTD and the initial velocity $\boldsymbol{V}(0)$.

Step 4 : Calculate the initial membership matrix $\boldsymbol{U}(1)$ using equation (5).

Step 5 : Calculate $\epsilon = Max|\boldsymbol{U}(1) - \boldsymbol{U}(0)|$ and determine the personal best $\boldsymbol{P}$, global best $\boldsymbol{p_g}$, and global best membership $\boldsymbol{u_g}$.

Step 6 : Divide the swarm into sub-swarms 1 and 2 using equations (14) and (15).

Step 7 : For each iteration $t = 1,2, \ldots, Maxit$, perform mutation $\boldsymbol{M}(t)$, crossover $Q(t)$, and selection $\boldsymbol{P_{iwin}}(t)$ using equations (10), (11), (12), and (13).

Step 8 : Update the centroid $\boldsymbol{X}(t)$, velocity $\boldsymbol{V}(t)$, and membership $\boldsymbol{U}(t+1)$ using equations (8), (9), and (5).

Step 9 : Calculate $\epsilon = Max|\boldsymbol{U}(t+1) - \boldsymbol{U}(t)|$ and personal best $\boldsymbol{P}(t)$, global best $\boldsymbol{p_g}(t)$, dan global best membership $\boldsymbol{u_g}(t)$. Repeat steps 7 to 9 until the PSOTD stopping condition is satisfied and get the global best $\boldsymbol{p_g}$ and memberhship global best $\boldsymbol{u_g}$.

Step 10 : Convert $\boldsymbol{p_g}$ and $\boldsymbol{u_g}$ into $\boldsymbol{v}(0)$ and $\boldsymbol{u}(0)$ in the IKFCM algorithm.

Step 11 : For each iteration $i = 1,2,\dots,Maxit$, update $\boldsymbol{u}(i)$ and $\boldsymbol{v}(i)$ with equations (5) and (6). Perform step 11 until the IKFCM stopping condition is met and get the solution $\boldsymbol{u}$ and $\boldsymbol{v}$.

2.4 Evaluation Tools

The last step is the process of evaluating the performance of the algorithm. Evaluation can be done with various methods. This study uses three clustering evaluation tools, namely Partition Entropy Index (PEI), Partition Coefficient Index (PCI), and Apparent Error Rate (APER).

PEI measures the level of data uniformity within a cluster [35].

$$PCI = -\frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}\, log_2\, u_{ij}\right) \tag{16}$$

PCI measures the level of diversity between clusters [35].

$$PCI = -\frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^2\right) \tag{17}$$

APER calculates the clustering error rate [36].

$$APER = \frac{number\ of\ incorrect\ objects}{total\ number\ of\ objects} \times 100\% \tag{18}$$

3 Experimental Study

3.1 Datasets

In this research, we use two different datasets, Iris and Wholesale Customers. Both datasets are actually complete datasets, but we make them incomplete datasets with several missing rates, namely 10%, 15%, and 20%.

1. Iris dataset is a data containing information about iris flowers. The dataset has four features and consists of 150 datums. There are three classes in this data, namely Setosa, Versicolor, and Virginica.
2. The Wholesale Customers dataset consists of 440 datums that have 6 attributes related to the clients of wholesale distributors. This data has 2 classes that can be seen in the channel index.

Each dataset has different values of q,c, dan σ^2 parameters in IKFCM. The differences are shown in Table 1.

Table 1. The values q and σ^2

Parameter	Iris	Wholesale Customers
q	5	7
c	3	2

σ^2	1.0	0.7

3.2 Parameter Settings

The parameter configuration of the IKFCM-PSOTD algorithm can be seen in Table 2. The parameters used in the IKFCM algorithm are taken from [16], while the parameters used in the PSOTD algorithm are taken from [32]. These parameters are also used in five comparison algorithms, namely: Mean KFCM, Median KFCM, OCSKFCM [17], IKFCM [16], and IKFCMPSO.

Table 2. Parameter Settings

Parameter	Value
IKFCM Algorithm	
ϵ	10^{-10}
m	2
Maxiter	1000
PSOTD Algorithm	
Np	50
C	1.496
CR_1	0.025
CR_2	0.9
F	0.5
$Maxiter$	1000

3.3 Experimental Results

3.3.1 Experimental results and comparisons on Iris dataset

The experimental results of the proposed algorithm and five comparison algorithms on the iris dataset can be seen in Table 3.

Table 3. Experimental results for incomplete Iris dataset.

Missing Rates	Mean KFCM	Median KFCM	OCSKFCM	IKFCM	IKFCMPSO	IKFCMPSOTD
			Mean Result of PEI			
10%	0.3042	0.3408	0.3569	0.2641	0.2501	**0.2465**
15%	0.2914	0.3668	0.3233	0.2758	0.2738	**0.2622**
20%	0.2794	0.3431	0.2886	0.2759	0.2673	**0.2584**
			Mean Result of PCI			
10%	0.8368	0.8161	0.8055	0.8555	0.8651	**0.8660**
15%	0.8435	0.8035	0.8265	0.8536	0.8561	**0.8591**
20%	0.8448	0.8120	0.8465	0.8024	0.8598	**0.8632**
			Mean Result of Aper			
10%	0.1776	0.1711	0.1711	0.1503	0.1503	**0.1447**
15%	0.1842	0.1776	0.1711	0.1579	0.1503	**0.1382**
20%	**0.1645**	0.1909	0.1830	0.1503	0.1830	0.1710

As seen in Table 3, the IKFCMPSOTD algorithm is able to obtain the best PEI and PCI values on incomplete data with various missing rates. This shows that IKFCMPSOTD is able to produce the best clustering performance among all tested algorithms. A low PEI value indicates that the clusters in the partition have a high degree of homogeneity,

meaning that the data in one cluster has similar characteristics. A high PCI value indicates that the clusters in the partition have significant differences. In addition, the IKFCMPSOTD algorithm also obtained the best score on the Iris dataset with a 10% and 15% missing rate, while Mean KFCM obtained the best APER score with a 20% missing rate. This shows that the imputation quality using intervals provides a good value that results in a low error rate.

3.3.2 Experimental results and comparisons on Wholesale Customers dataset
The experimental results of the proposed algorithm along with five comparison algorithms on the Wholesale Customers dataset can be seen in Table 4.

Table 4. Experimental results for incomplete Wholesale Customers dataset.

Missing Rates	Mean KFCM	Median KFCM	OCSKFCM	IKFCM	IKFCMPSO	IKFCMPSOTD
Mean Result of PEI						
10%	0.2315	0.3004	**0.2288**	0.2723	0.2398	0.2394
15%	0.2405	0.3037	0.2226	0.2819	0.2283	**0.2203**
20%	0.2749	0.3209	0.2491	0.2642	0.2128	**0.1765**
Mean Result of PCI						
10%	0.8614	0.8172	**0.8626**	0.8339	0.8559	0.8564
15%	0.8559	0.8137	0.8660	0.8263	0.8650	**0.8698**
20%	0.8832	0.8012	0.8502	0.8405	0.8727	**0.8980**
Mean Result of Aper						
10%	0.4068	0.4568	0.1636	0.1818	0.1772	**0.1591**
15%	0.1523	0.2045	0.1659	0.1500	0.1932	**0.1432**
20%	0.4818	0.4614	0.1455	0.1432	0.1409	**0.1341**

As seen in Table 4, the IKFCMPSOTD algorithm obtained the best accuracy or APER value at each missing rate. This shows that imputation using intervals is able to produce values that are close to the original. Meanwhile, IKFCMPSOTD also obtained the best PEI and PCI scores at missing rates of 15% and 20%, while OCSKFCM obtained the best scores in the remaining trials. PEI and PCI are used to measure the quality of the resulting clusters. So, it can be said that the IKFCMPSOTD algorithm produces good clustering performance. Based on Table 3 and 4, the IKFCMPSOTD algorithm has better results than the IKFCM and IKFCMPSO algorithms. This indicates that the PSOTD algorithm used as an optimization algorithm successfully improves the performance of the IKFCM algorithm and is more effective than the standard PSO.

4 Conclusion

We propose the IKFCMPSOTD algorithm, which is used on incomplete datasets. The IKFCMPSOTD algorithm is a KFCM algorithm optimized with the PSOTD algorithm. In addition, the interval technique is used to overcome the problem of missing values in incomplete datasets. The results show that the IKFCMPSOTD algorithm produces the best PEI and PCI evaluation scores in most experiments so that the resulting cluster can be said to have good quality. Meanwhile, the IKFCMPSOTD

algorithm also obtained the best APER value in almost all experiments. This shows that the use of intervals as an imputation method is highly recommended, especially when compared to commonly used methods such as mean and median. The results also show that the use of PSOTD can improve the performance of the IKFCM algorithm and is superior to the standard PSO algorithm.

Acknowledgements

We would like to thank the High-Performance Computing (HPC) AI-Center Brawijaya University for providing this research facility.

References

[1] H. H. Bock, "Clustering Methods: A History of k-Means Algorithms," in *Data Analysis and Classification*, Berlin, Heidelberg, Springer, 2007.

[2] P. H. Sneath and R. R. Sokal, Numerical Taxonomy : The Principles and Practice of Numerical Classification, San Fransisco: W. H. Freeman, 1973.

[3] L. Zhang and L. Guo, "Application of CLustering and Recommendation Algorithm in Sports Competition Pressure Source," *Scientific Programming,* 2022.

[4] C. Regan, C. Fehily, E. Campbell, J. Bowman, J. Faulkner, C. Oldmeadow and K. Bartlem, "Clustering of Chronic Disease Risks Among People Accessing Community Mental Health Services," Preventive Medicine Reports, 2022.

[5] Q. Wen, "Application of Clustering Algorithm in Corporate Strategy and Risk," *Computational Intelligence and Neuroscience,* 2022.

[6] I. Skrjanc, G. Andonovski, J. A. Iglesias, M. P. Sesmero and A. Sanchis, "Evolving Gaussian Online Clustering in Social Network Analysis," in *International Conference on Information and Communication Technology (IconICT)*, 2022.

[7] A. G. Oskouei, M. Hashemzadeh, B. Asheghi and M. A. Balafar, "CGFFCM: Cluster-weight and Group-local Feature-weight Learning in Fuzzy C- Means Clustering Algorithm for Color Image Segmentation," *Applied Soft Computing,* 2021.

[8] L. A. Zadeh, Fuzzy Sets and Fuzzy Information Granulation Theory, Beijing: Normal University Press, 2000.

[9] J. C. Bezdek, "Cluster Validity with Fuzzy Sets," *Cybernetics,* vol. 3, no. 3, pp. 58-73, 1973.

[10] S. P. Lloyd, "Least squares quantization in PCM," Technical Report RR-5497, Bell Lab, 1957.

[11] E. R. Syahputra, Y. A. Dalimunthe and Irvan, "Application of fuzzy C-Means Algorithm for Determining Field of Interest in Information System Study STTH Medan," *Expert Systems With Applications,* no. 207, 2017.

[12] D. Krasnov, D. Davis, K. Malott, Y. Chen, X. Shi and A. Wong, "Fuzzy C-Means Clustering: A Review of Applications in Breast Cancer Detection," *Entrophy,* no. 25, 2023.

[13] N. Yuwen, S. Xiaoyuan, Y. Jingong and X. Duowen, "Application of Fuzzy C-Means Clustering Method in The Analysis of Severe Medical Images," *Journal of Intelligent & Fuzzy Systems,* vol. 38, no. 4, pp. 3635-3645, 2020.

[14] A. Halder, "Kernel Based Rough Fuzzy C-Means Clustering Optimized Using Particle Swarm Optimization," *International Symposium on Advanced Computing and Communication (lSACC),* 2015.

[15] H. Wang and S. Wang, "Mining Incomplete Survey Data Through Classification," *Knowl Inf Syst,* vol. 24, pp. 221-223, 2010.

[16] T. Li, L. Zhang, W. Lu, H. Hou, X. Liu and W. Pedrcz, "Interval kernel Fuzzy C-Means Clustering of Incomplete Data," *Neurocomputing,* vol. 237, pp. 316-331, 2017.

[17] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-Means Clustering of Incomplete Data," *IEEE Transanctions on System, Man, and Cybernetics—Part B: Cybernetics,* vol. 31, no. 5, pp. 735-744, 2001.

[18] A. Z. Zhang and J. Z. Li, "Interval Estimation for Aggregate Queries on Incomplete Data," *Journal Of Computer Science and Technology,* vol. 34, no. 6, pp. 1203-1216, 2019.

[19] A. Balqis and B. Y. Sadok, "A New Algorithm for Fuzzy Clustering Handling Incomplete Dataset," *Int. J. Artif. Intell. Tools,* vol. 23, no. 4, 2014.

[20] D. Q. Zhang and S. C. Chen, "Clustering Incomplete Data Using Kernel- Based Fuzzy C-means Algorithm," *Neural Processing Letters,* vol. 18, pp. 155-162, 2003.

[21] A. K. G. Rodrigues, R. Ospina and M. R. P. Ferreira, "Adaptive Kernel Fuzzy Clustering for Missing Data," *Plos One,* vol. 16, no. 11, 2021.

[22] H. Khan, X. Wang and H. Liu, "Missing Value Imputation Through Shorter Interval Selection Driven by Fuzzy C-Means Clustering," *Computers and Electrical Engineering,* vol. 93, 2021.

[23] L. Y. Zhang, W. Lu, X. D. Lin, W. Pedrycz, C. Q. Zhong and L. Wang, "A Global Clustering Approach Using Hybrid Optimization for Incomplete Data Based on Interval Reconstruction of Missing Value," *Int. J. Intell. Syst,* vol. 31, no. 4, pp. 297-313, 2016.

[24] Y. Ding and X. Fu, "Kernel-based Fuzzy C-Means Clustering Algorithm based on Genetic Algorithm," *Neurocomputing,* vol. 188, pp. 233-238, 2016.

[25] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, 1995.

[26] T. Cura, " particle swarm optimization approach to clustering," *Expert Systems with Applications,* vol. 39, no. 1, pp. 1582-1588, 2012.

[27] H. Izakian and A. Abraham, "Fuzzy C-Means and Fuzzy Swarm for Fuzzy Clustering Problem," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1835-1838, 2011.

[28] S. Sengupta, S. Basak and R. A. Peters, "Data Clustering Using a Hybrid of Fuzzy C-Means and Quantum-behaved Particle Swarm Optimization," in *Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2017.

[29] T. M. Silva Filho, B. A. Pimentel, R. M. C. R. Souza and A. L. I. Oliveira, "ybrid Methods for Fuzzy Clustering Based on Fuzzy C-Means and Improved Particle Swarm Optimization," *Expert Systems with Applications,* vol. 42, no. 17-18, pp. 6315-6328, 2015.

[30] C. K. Praseeda and B. L. Shivakumar, "Fuzzy Particle Swarm Optimization (FPSO) Based Feature Selection and Hybrid Kernel Distance Based Possibilistic Fuzzy Local Information C-Means (HKD-PFLICM) Clustering for Churn Prediction in Telecom Industry," *SN Appl Sci,* 2021.

[31] M. N. M. Salleh and N. A. Samat, "FCMPSO: An Imputation for Missing Data Features in Heart Disease Classification," in *IOP Conf. Series: Materials Science and Engineering*, 2017.

[32] Y. Chen, L. Li, H. Peng, J. Xiao, Y. Yang and Y. Shi, "Particle Swarm Optimizer with Two Differential Mutation," *Applied Soft Computing,* vol. 61, pp. 314-330, 2017.

[33] W. Zhang and X. Xie, "DEPSO: Hybrid Particle Swarm with Differential Evolution Operator," in *IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance*, 2003.

[34] Y. Wu, X. Z. Gao, X. L. Huang and K. Zenger, "A Hybrid Optimization Method of Particle Swarm Optimization and Cultural Algorithm," in *Sixth International Conference on Natural Computation*, 2010.

[35] S. Mashfuufah and D. Istiawan, "Penerapan Partition Entropy Index, Partition Coefficient Index dan Xie Beni Index untuk Penentuan Jumlah Klaster Optimal pada Algoritma Fuzzy Cmeans dalam Pemetaan Tingkat Kesejahteraan Penduduk Jawa Tengah," in *Proceeding of the 7th University Research Colloquium 2018: Mahasiswa (student paper presentation)*, Surakarta, 2018.

[36] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, Upper Saddle River: Pearson Education, Inc, 2007.