CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(1) (2024), Pages 106-118
p-ISSN: 2086-0382; e-ISSN: 2477-3344

KFCM-PSOTD : An Imputation Technique for Missing Values in Incomplete Data Classification

Muhaimin Ilyas*, Syaiful Anam, Trisilowati

Department of Mathematics, Faculty of Science and Mathematics, Brawijaya University, Malang, Indonesia

Email: chaimin101@gmail.com

ABSTRACT

Missing values are one of the main problems in data mining. Generally, this can be solved with mean or median imputation, as it is easy to implement, although it does not guarantee that the values represent the actual values. This research develops a novel approach to obtain better imputation values, namely Kernel Fuzzy C-Means optimized with Particle Swarm Optimizer with Two Differential Mutations (KFCM-PSOTD). KFCM imputation is applied to obtain better estimation values due to its proven ability to recognize patterns in data. In addition, the PSOTD algorithm is used as an optimization tool to improve the performance of KFCM. PSOTD is adopted because it has more balanced exploration and exploitation capabilities compared to classic PSO. The imputed dataset on KFCM-PSOTD is classified using the Decision Tree algorithm. The results are evaluated using accuracy, precision, recall, and the f1 score to determine the quality of the imputed values. The results show that the KFCM-PSOTD algorithm has better performance; even the difference in evaluation value obtained is up to 10% better than other imputation techniques, although it requires longer computation time. KFCM-PSOTD algorithm is recommended as a missing value imputation tool compared to commonly used techniques such as mean and median.

Keywords: Imputation; Incomplete Data; Kernel based Fuzzy C-Means; Missing Value; Particle Swarm Optimization

INTRODUCTION

Over the years, data has continued to be produced in large volumes, known as big data. Big data can revolutionize the way of businesses, science, and trends. Data utilization can be done with data mining [1]. Data mining (DM) is the method of collecting data to extract valuable information from the data. DM has become an exciting trend and has significantly increased, with the size, variety, and speed growing every year [2]. One commonly studied DM technique is the classification technique. Classification is the process of identifying one or more classes or categories to which a new observation belongs, based on a training dataset [3]. Classification has been used in various areas of research, including medical [4]-[6], biology [7] [8], economy [9] [10], etc. Even so, classification often faces some problems. One of them is incomplete data due to missing

Submitted: December 21, 2023 Reviewed: January 26, 2024 Accepted: May 10, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i1.25138

values. The study states that 5% or higher of a dataset are missing values [11]. Therefore, data mining has one important stage, namely data preprocessing [12].

Datasets are preprocessed by checking for missing values, noisy data, and other errors. Removing missing values from the dataset can be done with imputation techniques [13]. The mean or median value is commonly used to fill in missing values, although it potentially ignores the variance of the data, especially if there are a large number of missing values. Therefore, this technique often causes biased estimates [14]. Several imputation techniques have been developed to produce better estimates. One of the most widely developed imputation techniques is the clustering-based imputation technique, due to its ability to recognize patterns in the data.

Clustering is the process of dividing data into several clusters based on the similarity of patterns in the data. A widely used clustering algorithm is the Fuzzy C-Means (FCM) algorithm. FCM is a development of the K-Means algorithm by adding fuzzy concepts that make it more flexible. Hathaway and Bezdek proposed four FCM approaches to deal with missing data. These four strategies were widely adopted by later studies that developed FCM as an imputation tool [15]. Aristiawati used the FCM algorithm to impute missing values in the lung disease dataset [16]. Li et al. developed an imputation algorithm based on FCM combined with a vaguely quantified rough set [17]. Zhang and Chen proposed the use of the FCM algorithm by adopting the Gauss kernel to solve the problem of incomplete data [18]. The results showed that the Kernel FCM outperformed the standard FCM algorithm. Kernel-based FCM (KFCM) is preferred as it has a better approach to the data structure. However, the FCM or KFCM, algorithm has the weakness of sensitivity to the initial centroid.

The initial centroid problem in FCM or KFCM algorithms can be solved by optimization techniques, such as metaheuristic optimization. Particle Swarm Optimization (PSO) is one of the most well-known and frequently used heuristic algorithms due to its easy implementation. PSO algorithm can help FCM algorithm achieve a global solution [19]-[21]. Salleh and Samat proposed a combination of FCM and PSO as a tool for imputing missing data in the Framingham Heart Disease Dataset [13]. The imputed dataset is then classified using the Decision Tree algorithm. The results show that the proposed algorithm is able to outperform the mean algorithm, KNN, and FCM with a significant value.

Getting stuck in a local solution is a major problem in PSO. The PSO algorithm is constantly being improved in order to avoid local optima [22]. The Differential Evolution (DE) algorithm is often used to modify PSO due to its effectiveness and simplicity. Zhang and Xie introduced the DEPSO (Differential Evolution Particle Swarm Optimization) algorithm [23]. In addition, Wu et al. developed the PSOCA (Particle Swarm Optimization Cultural Algorithm) [24]. Chen et al. also developed a combination of PSO and DE, but with a slightly different approach [22]. The proposed algorithm is Particle Swarm Optimizer with Two Differential Mutations (PSOTD). PSOTD adopts two different mutation operators and divides the swarm into two subswarms to balance the capabilities of exploration and exploitation. Based on the research results, PSOTD is able to boost the performance of standard PSO.

This study proposes the kernel-based FCM (KFCM) algorithm for the imputation of incomplete datasets [18] optimized by the PSOTD [22]. Then the imputed dataset will be classified using the Decision Tree algorithm, which has been widely applied in classification study [13]. The KFCM-PSOTD algorithm is compared with several other algorithms, such as mean, median, KFCM, and KFCMPSO imputation techniques.

METHODS

Kernel Fuzzy C-Means (KFCM) Algorithm

Kernelization supports a good approximation of the data structure [25]. The kernel-based FCM algorithm (KFCM) is a modified classic FCM algorithm designed to handle more complex data. KFCM minimizes an objective function in the form of equation (1) with the constraint of equation (2) [26]

$$\boldsymbol{J_{KFCM}}(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\phi(\boldsymbol{x}_j) - \phi(\boldsymbol{v}_i)\right\|^2, \tag{1}$$

$$\sum_{i=1}^{c} u_{ij}^{2} = 1, j = 1,2, \dots n, \tag{2}$$

$\boldsymbol{x}_j = [x_{j1}, x_{j2}, \dots, x_{js}]$ is a data point, $\boldsymbol{v}_i$ is the i-th cluster prototype in $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_c] \in \mathbb{R}^{s\times c}$, $u_{ij} \in [0,1]$ is the membership degree , c is the cluster amounts, n is the data amounts, m is fuzzy parameter, and

$$\left\|\phi(\boldsymbol{x}_j) - \phi(\boldsymbol{v}_i)\right\|^{\boldsymbol{2}} = K(\boldsymbol{x}_j, \boldsymbol{x}_j) + K(\boldsymbol{v}_i, \boldsymbol{v}_i) - \boldsymbol{2}K(\boldsymbol{x}_j, \boldsymbol{v}_i). \tag{3}$$

The adopted kernel function is the Gaussian kernel function shown in the equation (4)

$$K(\boldsymbol{x}, \boldsymbol{v}) = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{v}\|^2}{\sigma^2}\right). \tag{4}$$

If $x = v$, then $K(\boldsymbol{x}, \boldsymbol{x}) = 1$ so equation (3) can be written as

$$\left\|\phi(\boldsymbol{x}_j) - \phi(\boldsymbol{v}_i)\right\|^{\boldsymbol{2}} = 2\left(1 - K(\boldsymbol{x}_j, \boldsymbol{v}_i)\right). \tag{5}$$

The conditions required to minimize the equation (1) with constraint (2) can be found using the Lagrange multiplier technique. The results are shown in equations (6) and (7)

$$u_{ij} = \left[\sum_{t=1}^{c}\left(\frac{\|\phi(\mathbf{x}_{\mathrm{k}}) - \phi(\boldsymbol{v}_{\boldsymbol{i}})\|^2}{\|\phi(\mathbf{x}_{\mathrm{k}}) - \phi(\boldsymbol{v}_{\boldsymbol{t}})\|^2}\right)^{\frac{1}{m-1}}\right]^{-1}, i = 1,2, \dots, c; j = 1,2, \dots, n, \tag{6}$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} K(\boldsymbol{x}_j, \boldsymbol{v}_i)\boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^{m} K(\boldsymbol{x}_j, \boldsymbol{v}_i)}, i = 1,2, \dots, c. \tag{7}$$

The use of KFCM as an imputation technique can be applied at each iteration. The missing value estimate is obtained with the equation (8) [26]

$$x_{kj} = \frac{\sum_{i=1}^{c} u_{ij}^{m} K(\boldsymbol{x}_j, \boldsymbol{v}_i) v_{ik}}{\sum_{i=1}^{c} u_{ij}^{m} K(\boldsymbol{x}_j, \boldsymbol{v}_i)}, k = 1,2, \dots, s; j = 1,2, \dots, n. \tag{8}$$

Particle Swarm Optimizer with Two Differential Mutations (PSOTD) Algorithm

In general, PSO has some aspects that can be improved, mainly because the solution is often trapped in the local optimal region. Chen et al. added several operators of Differential Evolution (DE) to PSOTD, such as mutation, crossover, and selection [22]. PSOTD also divides the swarm in PSO into sub-swarms that focus on exploration and exploitation capabilities. The division is obtained using equations (9) and (10)

$$N = N_1 + N_2, \tag{9}$$

$$N_2 = \left\lfloor \tfrac{t}{Maxit} N \right\rfloor, \tag{10}$$

$Maxit$ is the number of iterations and $\lfloor\ \rfloor$ is the floor operation. N_1 denotes sub-swarm 1 while N_2 denotes sub-swarm 2. There are two vectors corresponding to the i-th particle $(i = 1,2,\dots,Np)$, which are the velocity vector $\boldsymbol{v_i} = (v_{i1}, v_{i2}, \dots, v_{is})$ and the position vector $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{is})$, where Np is the population size. Both are updated iteratively using equations (11) and (12).

$$\boldsymbol{v}_i^{t+1} = \omega \boldsymbol{v}_i^t + cr\left(\boldsymbol{p_{iwin}}_i^{\ t} - \boldsymbol{x}_i^t\right), \tag{11}$$

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \boldsymbol{v}_i^{t+1}, \tag{12}$$

ω is the inertial moment, c is cognitive constants, r is random numbers in the interval [0,1]. PSOTD has position and velocity vectors like the classic PSO algorithm, but there is a slight modification in the equation (11). $\boldsymbol{p_{iwin}}$ is the best position built in the PSOTD algorithm to find a global solution, as well as what differentiates it from the classic PSO algorithm. The $\boldsymbol{p_{iwin}}$ position has the same function as $\boldsymbol{p_{best}}$, but $\boldsymbol{p_{iwin}}$ is obtained through the processes of mutation, crossover, and selection. The three DE operators adopted by PSOTD are described below.

1. Mutation : This operation serves to create algorithms with more diversity and exploration.
   a. DE/rand/1

$$m_{ik} = p_{r1k} + F\left(p_{r2k} - p_{r3k}\right), \tag{13}$$

   b. DE/current to best/1

$$m_{ik} = p_{ik} + F\left(pg_{ik} - p_{ik}\right) + F\left(p_{r1k} - p_{r2k}\right), \tag{14}$$

   $r_1, r_2,$ 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 swaps some $\boldsymbol{m}$ components with $\boldsymbol{p}$ to make a new vector $\boldsymbol{q}$.

$$q_{\boldsymbol{ik}} = \begin{cases} m_{ik} & \text{if } r_2 \le CR \text{ atau } k = k_{rand} \\ p_{ik} & \text{otherwise,} \end{cases} \tag{15}$$

   r_2 is a random number in the interval [0,1]. CR is the Crossover Rate. $k_{rand}$is an integer random number in$[1, k]$ which guarantees that at least one element in vector $\boldsymbol{q}$ differs from vector $\boldsymbol{p}$.
3. Selection: This operation ensures that vectors with higher fitness values are kept in the next generation.

$$\boldsymbol{P_{iwin_i}} = \begin{cases} \boldsymbol{q_i} & \text{if } f(\boldsymbol{q_i}) \le f(\boldsymbol{p_i}) \\ \boldsymbol{p_i} & \text{otherwise.} \end{cases} \tag{16}$$

These three processes guarantee that each particle iteratively improves until it reaches a global solution.

Proposed Algorithm: KFCM-PSOTD

KFCM-PSOTD is a KFCM algorithm optimized by PSOTD. The optimized part is the centroid of KFCM. The following are the steps of the KFCM-PSOTD algorithm.

Step 01 : Randomly generate the initial velocity matrix $\boldsymbol{V}(0)$, initial position (centroid) $\boldsymbol{X}(0)$, and initial membership $\boldsymbol{U}(0)$.

Step 02 : Calculate $\boldsymbol{U}(1)$ using equation (6).

Step 03 : Calculate the initial fitness value using the formula $\epsilon = Max|\boldsymbol{U}(1) - \boldsymbol{U}(0)|$, then determine $\boldsymbol{p_{best}}, \boldsymbol{g_{best}}, \boldsymbol{ug_{best}}$.

Step 04 : Split the swarm into two parts, which are sub-swarm 1 and sub-swarm 2, using equations (9) and (10).

Step 05 : For every iteration $t = 1,2,\dots,Maxit$, do mutation operation $\boldsymbol{M}(t)$ with equations (13)and (14), crossover $\boldsymbol{Q}(t)$ with equation (15), and selection $\boldsymbol{p\_iwin}$ (t) with equation (16).

Step 06 : Update $\boldsymbol{V}(t)$ with equation (11), $\boldsymbol{X}(t)$ with equation (12), and $\boldsymbol{U}(t+1)$ with equation (6).

Step 07 : Calculate the fitness value with the formula $\epsilon = Max|\boldsymbol{U}(t+1) - \boldsymbol{U}(t)|$, then determine $\boldsymbol{p_{best}}(t)$, $\boldsymbol{g_{best}}(t)$, dan $\boldsymbol{ug_{best}}(t)$. Execute and repeat steps 05-07 until the PSOTD stopping criterion is satisfied, then $\boldsymbol{g_{best}}$ and $\boldsymbol{ug_{best}}$ solutions are obtained.

Step 08 : Use $\boldsymbol{g_{best}}$ and $\boldsymbol{ug_{best}}$ as the initial centroid values $\boldsymbol{v}(0)$ and $\boldsymbol{u}(0)$ in the KFCM algorithm.

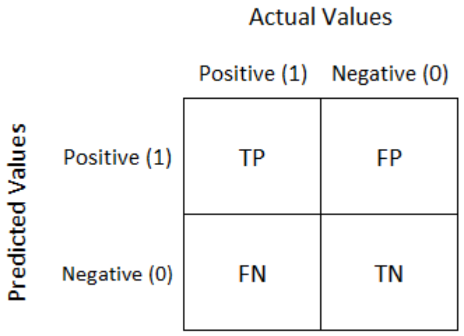
Step 09 : For every iteration $it = 1,2,\dots,Maxit$, update $\boldsymbol{v}(it)$ with equation (7) and $\boldsymbol{u}(it)$ with equation (6).

Step 10 : For each missing value, estimate the missing value using equation (8). Execute and repeat steps 09-10 until the stopping criteria are satisfied.

Step 11 : Extract the complete imputed dataset for the classification process using C4.5 Decision Tree [28] with a proportion of 80% training data and 20% testing data .

Evaluation Tools

The final step in data mining is to evaluate the performance of the algorithms used. This study utilizes the widely used evaluation methods of accuracy, precision, recall, and f1-score. The value can be calculated using the confusion matrix's components, which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as shown in Figure 1.



Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Figure 1. Confusion Matrix

The accuracy, precision, recall, and f1-score are obtained using equations (18)–(21)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{18}$$

$$precision = \frac{TP}{TP + FP}, \tag{19}$$

$$recall = \frac{TP}{TP + FN}, \tag{20}$$

$$f1\ score = \frac{2 \times recall \times precision}{recall + precision}. \tag{21}$$

Experimental Design

The test datasets used in this study are Iris dataset and Wholesale Customers dataset. The dataset actually a complete data sets but organized into incomplete dataset with missing value levels of 10%, 15%, and 20%. The Iris dataset has attributes about the characteristics of iris flowers and has 150 datums consisting of four attributes and three classes. The wholesale customers dataset contains information related to the clients of wholesale distributors. It has 440 datums, six attributes, and two classes. The research design of this study can be seen in Figure 2.

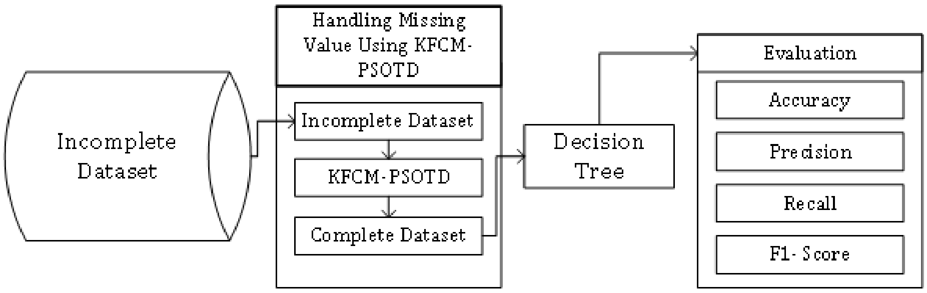


Figure 2. Experimental Design

Table 1 shows the parameters used in this study. The parameters are adopted from [26](KFCM) and [22] (PSOTD).

Table 1. Parameter Settings

KFCM Parameter	Value	PSOTD Parameter	Value
ϵ	10^{-10}	Np	50
m	2	C	1.496
$Maxiter$	1000	CR_1	0.025
σ^2 iris	1	CR_2	0.9
σ^2 wholesale customers	0.7	F	0.5
c iris	3	$Maxiter$	1000
c wholesale customers	2	Testing Amount	25

RESULTS AND DISCUSSION

The experimental results of the KFCM-PSOTD algorithm along with four other algorithms on Iris and Wholesale Customers dataset with missing rates of 10%, 15%, and 20% are shown in Table 2 and 3.

Table 2. Results on incomplete Iris dataset

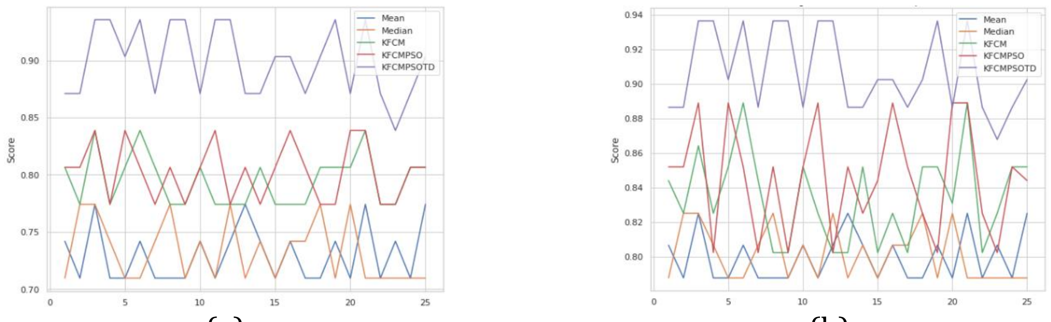
Missing Rate	Mean		Median		KFCM		KFCM-PSO		KFCM-PSOTD	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Accuracy (Acc)										
10%	0.730	0.024	0.733	0.027	0.794	0.022	0.802	0.027	**0.899**	0.031
15%	0.790	0.028	0.774	0.000	0.808	0.054	0.809	0.041	**0.965**	0.013
20%	0.790	0.016	0.804	0.030	0.823	0.024	0.827	0.024	**0.919**	0.041
Precision (Pre)										
10%	0.800	0.014	0.801	0.015	0.835	0.026	0.845	0.031	**0.907**	0.032
15%	0.801	0.023	0.789	0.000	0.819	0.046	0.820	0.034	**0.963**	0.018
20%	0.809	0.024	0.810	0.034	0.825	0.024	0.826	0.025	**0.913**	0.038
Recall (Re)										
10%	0.778	0.019	0.780	0.021	0.832	0.020	0.840	0.023	**0.899**	0.023
15%	0.796	0.026	0.782	0.000	0.810	0.052	0.812	0.040	**0.949**	0.015
20%	0.787	0.015	0.805	0.029	0.824	0.025	0.828	0.024	**0.926**	0.039
F1 Score										
10%	0.752	0.024	0.754	0.026	0.808	0.017	0.816	0.025	**0.900**	0.030
15%	0.792	0.028	0.776	0.000	0.811	0.051	0.812	0.038	**0.954**	0.016
20%	0.783	0.014	0.805	0.029	0.819	0.027	0.824	0.027	**0.911**	0.043

Time (s)										
10%	**0.761**	0.171	1.011	0.125	1.209	0.206	77.36	14.24	107.0	9.11
15%	**0.490**	0.030	0.494	0.044	1.426	0.464	78.67	18.75	106.4	25.99
20%	**0.523**	0.059	0.537	0.053	1.229	0.690	84.93	2.57	100.3	16.14

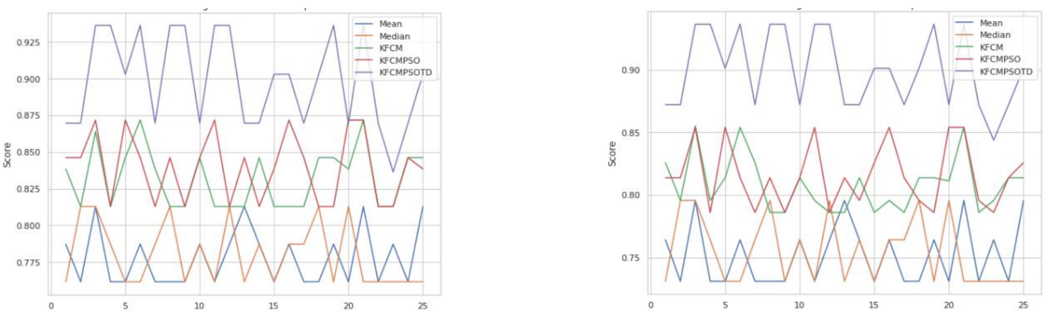
Table 3. Results on incomplete Wholesale Customers dataset

Missing	Mean		Median		KFCM		KFCM-PSO		KFCM-PSOTD	
Rate	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Accuracy (Acc)										
10%	**0.834**	0.017	0.802	0.013	0.788	0.028	0.791	0.028	**0.833**	0.032
15%	0.775	0.018	0.798	0.010	0.769	0.024	0.769	0.022	**0.849**	0.025
20%	0.750	0.001	0.768	0.013	0.774	0.026	0.777	0.017	**0.815**	0.022
Precision (Pre)										
10%	0.766	0.020	0.717	0.016	0.733	0.031	0.736	0.031	**0.787**	0.033
15%	0.738	0.015	0.792	0.011	0.709	0.027	0.708	0.027	**0.803**	0.029
20%	0.739	0.011	0.762	0.015	0.713	0.030	0.719	0.019	**0.767**	0.022
Recall (Re)										
10%	0.798	0.015	0.750	0.016	0.762	0.034	0.766	0.034	**0.829**	0.030
15%	0.776	0.016	0.776	0.013	0.732	0.030	0.730	0.032	**0.845**	0.036
20%	0.724	0.001	0.741	0.014	0.731	0.034	0.740	0.025	**0.811**	0.024
F1 Score										
10%	0.798	0.019	0.750	0.016	0.743	0.032	0.746	0.032	**0.829**	0.034
15%	0.747	0.017	0.782	0.012	0.717	0.028	0.716	0.029	**0.818**	0.031
20%	0.729	0.001	0.747	0.014	0.720	0.031	0.726	0.020	**0.781**	0.024
Time (s)										
10%	1.315	0.136	**1.259**	0.270	2.349	0.231	109.2	60.07	197.8	46.13
15%	1.589	1.345	**0.951**	0.164	2.778	0.467	131.7	46,81	210.1	6.701
20%	**1.522**	0.146	1.639	1.122	3.376	1.779	120.5	55.70	206.3	4.920

Based on Table 2, the KFCM-PSOTD algorithm obtained the best evaluation value on each evaluation tool except time. Note that the KFCM, KFCMPSO, and KFCM-PSOTD imputation techniques outperform the mean and median techniques that are often used easily in data analysis. This indicates that instead of using the mean or median technique, the KFCM-based imputation technique is recommended. The use of PSOTD can improve the performance of the KFCM algorithm by up to 10% more (significantly better than PSO). Figure 3-5 shows the comparison graph of evaluation scores on iris data with a 10%, 15%, and 20% missing rate.

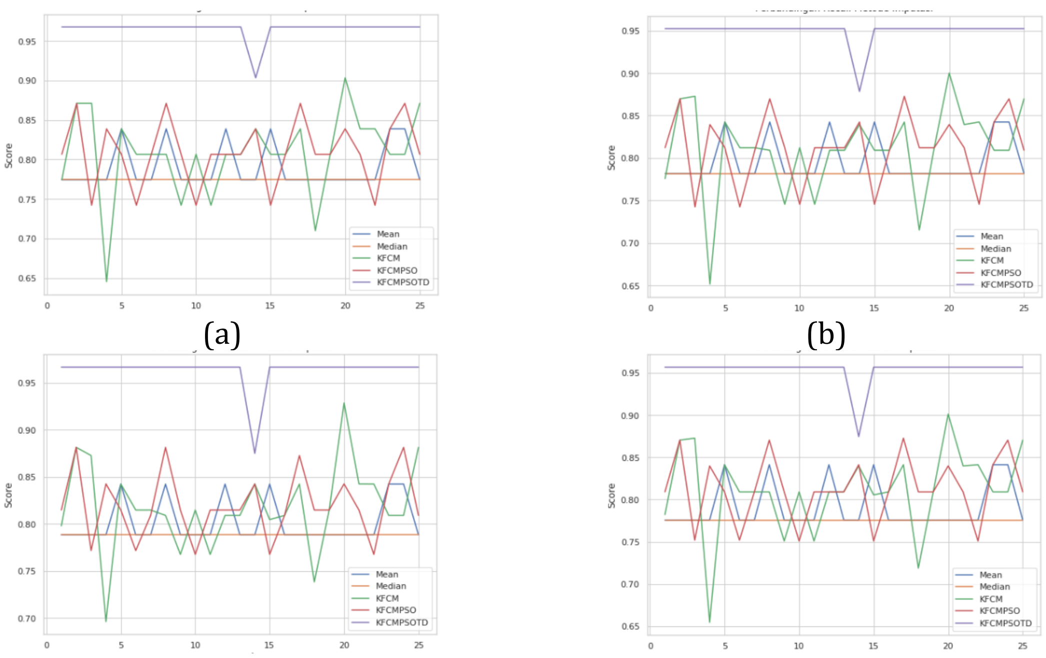


(a) (b)



(c) (d)

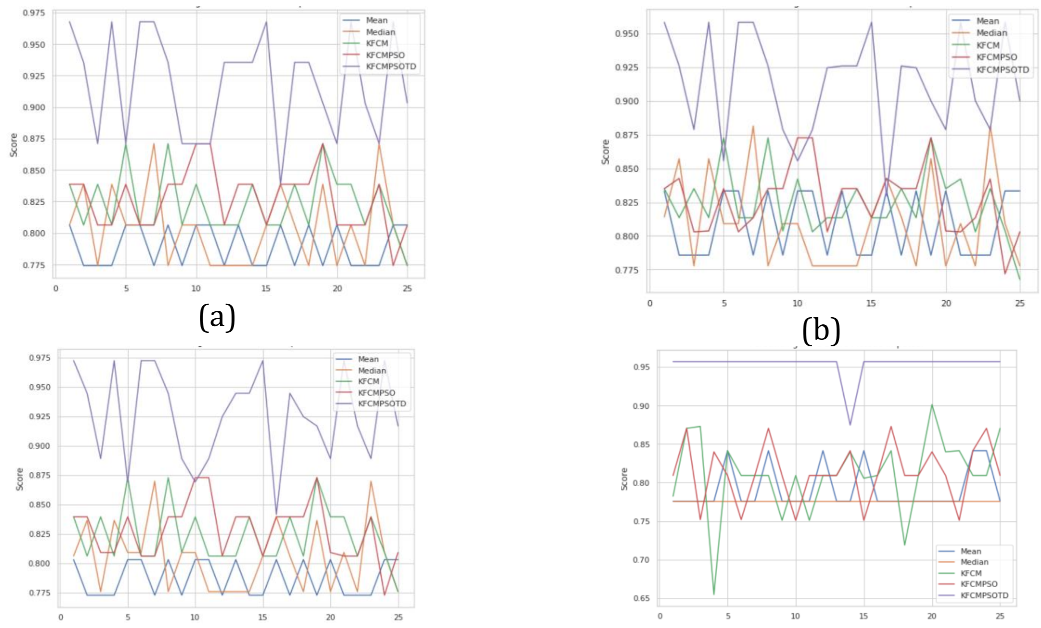
Figure 3. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Iris 10%



(a) (b)

(c) (d)

Figure 4. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Iris 15%



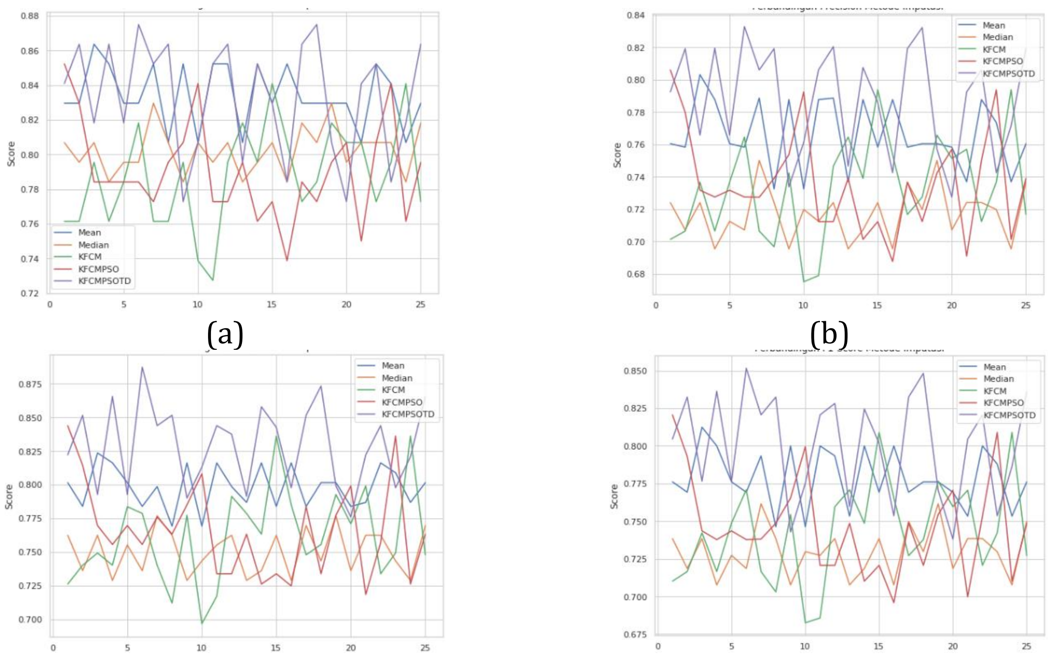
(a) (b)

(c) (d)

Figure 5. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Iris 20%

Based on Figures 3–5, it is clear that the KFCM-PSOTD algorithm is superior to the other four algorithms in 25 trials. According to Table 3, the KFCM-PSOTD algorithm has the highest evaluation score in almost every test. Although not as significant as the experiments on Iris data, PSOTD still performs well by producing a score difference of up to 7% from other algorithms. In addition, all tested algorithms have relatively small standard deviation values (approximately 0.0). Its indicate that the KFCM-PSOTD algorithm obtained consistent evaluation scores in 25 trials. Although superior in the evaluation of accuracy, precision, recall and f-1 score on Iris and Wholesale Customers data experiments, the KFCM-PSOTD algorithm has the disadvantage of being less efficient in computational time. Experimental results show that the KFCM-PSOTD algorithm takes much longer than other algorithms, especially those that are not optimized.

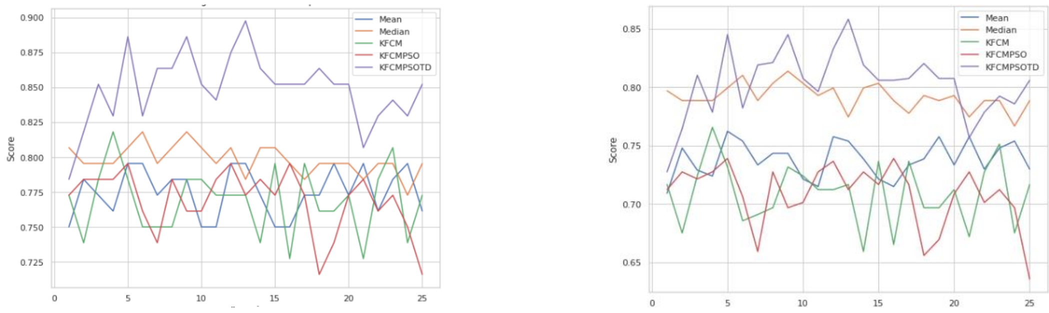
Figure 6-8 shows the comparison graph on Wholesale Customers data with a 10%, 15%, and 20% missing rate. Based on Figure 6, it can be seen that the algorithm consistently obtains better values, although the difference is not too significant. Figures 7 and 8 show that KFCM-PSOTD algorithm outperforms the other four algorithms on the wholesale customer dataset experiments.



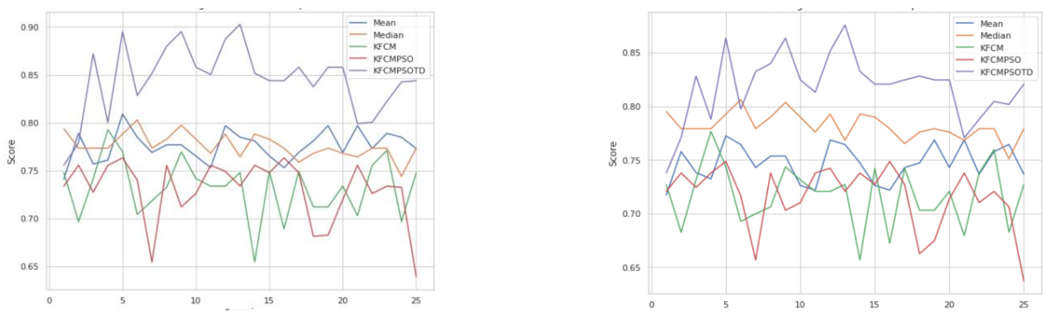
(a) (b)

(c) (d)

Figure 6. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Wholesale 10%

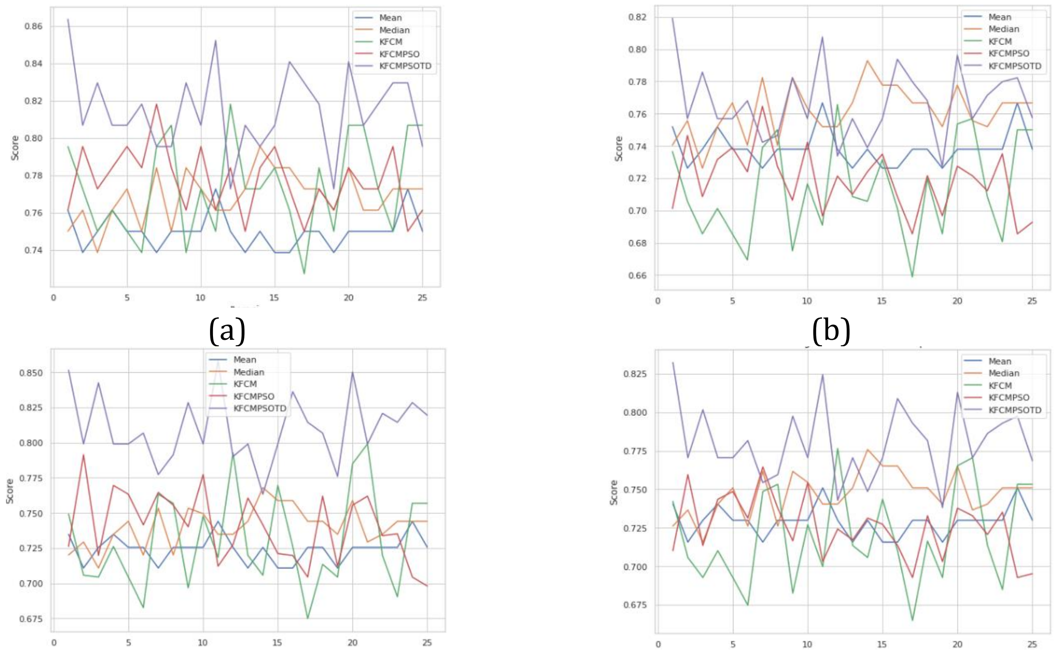


(a) (b)



(c) (d)

Figure 7. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Wholesale 15%



(a) (b)

(c) (d)

Figure 8. Comparison graph: (a) Acc (b) Pre (c) Re (d) F1 Score on Wholesale 20%

CONCLUSIONS

In this study, a KFCM clustering algorithm optimized by PSOTD has been proposed. This algorithm is applied as a missing value imputation tool in incomplete datasets. PSOTD is used to optimize initial centroid for the KFCM algorithm. The imputed dataset is used in the classification process using the Decision Tree algorithm to determine the quality of the imputation that has been performed.

The results indicate that the KFCM-PSOTD algorithm obtained the best evaluation value in almost all experiments. In fact, the resulting value far outperforms other imputation techniques, which are mean, median, KFCM, and KFCM-PSO. The PSOTD algorithm as an optimization tool also has a significant impact by increasing the accuracy, precision, recall, and f1 score compared to KFCM and KFCM-PSO. However, the proposed algorithm requires much longer computational time than other non-optimized algorithms. Based on the results obtained, the KFCM-PSOTD algorithm is recommended as a missing value imputation tool compared to commonly used techniques such as mean and median.

Future research may be improved, especially with the development of faster optimization tools, so that the proposed algorithm becomes more efficient in computational time. Applications on real and high-dimensional datasets can also be implemented to find out how good the proposed algorithm is on various datasets. In addition, various parameter settings will show the stability of the proposed algorithm in various applications.

REFERENCES

[1] D. Dietrrich, B. Heller and B. Yang, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Hoboken: John Wiley & Sons, 2015.

[2] J. Dean, Big Data, Data Mining, and Machine Learning, Hoboken: John Wiley & Sons, 2014.

[3] D. Lamba, W. H. Hsu and M. Alsadhan, "Predictive Analytics and Machine Learning for Medical Informatics: A Survey of Tasks and Techniques," in *Intelligent Data-Centric Systems, Machine Learning, Big Data, and IoT for Medical Informatics,*, Academic Press, 2021, pp. 1-35.

[4] S. Loghavi, R. Kanagal-Shamanna, J. D. Khoury, L. J. Medeiros, K. N. Naresh, R. Nejati and M. M. Patnaik, "5th Edition of The World Health Classification of Tumors of The Hematopoietic And Lymphoid Tissues," *Modern Pathology,* 2023.

[5] B. W. v. Rhijn, A. E. Hentschel, J. Bründl, E. M. Compérat, V. Hernández, O. Čapoun, H. M. Bruins, D. Cohen, M. Rouprêt, S. F. Shariat, A. H. Mostafid, R. Zigeuner, J. L. Dominguez-Escrig and M. Burge, "Prognostic Value of the WHO1973 and WHO2004/2016 Classification Systems for Grade in Primary Ta/T1 Non–muscle-invasive Bladder Cancer: A Multicenter European Association of Urology Non–muscle-invasive Bladder Cancer Guidelines Panel Study," *European Urology Oncology,* vol. 4, no. 2, pp. 182-191, 2021.

[6] A. R. Fielder, G. E. Quinn, R. P. Chan, G. E. Holmström, M. F. Chiang, A. Berrocal, G. Binenbaum, M. Blair, J. P. Campbell, A. Capone, Y. Chen, S. Dai, A. Ells, B. Fleck, W. V. Good and M. E. Hartnet, "Retinopathy of Prematurity Classification Updates: possible Implications for Treatment," *Journal of American Association for Pediatric Ophthalmology and Strabismus,* vol. 26, no. 3, pp. 109-112, 2022.

[7] S. Zhou, H. Cai, X. He, Z. Tang and S. Lu, "Enzyme-mimetic Antioxidant Nanomaterials for ROS Scavenging: Design, Classification, and Biological Applications," *Coordination Chemistry Reviews,* vol. 500, 2023.

[8] B. Wei, K. Hao, L. Gao, X.-s. Tang and Y. Zhao, "A Biologically Inspired Visual Integrated Model for Image Classification," *Neurocomputing,* vol. 405, pp. 103-113, 2020.

[9] N. Dhakate and R. Joshi, "Classification of Reviews of e-Healthcare Services to Improve Patient Satisfaction: Insights from An Emerging Economy," *Journal of Business Research,* vol. 164, 2023.

[10] M. d. l. Paz-Marín, P. A. Gutiérrez and C. Hervás-Martínez, "Classification of Countries' Progress Toward a Knowledge Economy Based on Machine Learning Classification Techniques," *Expert Systems with Applications,* vol. 42, no. 1, pp. 562-572, 2015.

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

[12] A. T. S. Dhevi, "Imputing Missing Values Using Inverse Distance Weighted Interpolation for Time Data Series," in *Sixth International Conference on Advanced Computing (ICoAC)*, 2014.

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

[14] M. Jamshidian and M. Mata, "Advances in Analysis of Mean and Covariance Structure when Data are Incomplete," in *Handbook of Computing and Statistics with Applications*, North-Holland, Handbook of Latent Variable and Related Models, 2007, pp. 21-44.

[15] R. J. Hathaway and J. C. Bezdek, "Fuzzy C-Means Clustering of Incomplete Data," *IEEE Trans. Syst. Man Cybern,* vol. 31, no. 5, pp. 735-744, 2001.

[16] K. Aristiawati, T. Siswantining, D. Sarwinda and S. M. Soemartojo, "Missing Values Imputation Based on Fuzzy C-Means Algorithm for Classification of Chronic Obstructive Pulmonary Disease (COPD)," in *Proceedings of the 8th SEAMS-UGM International Conference on Mathematics and Its Applications 2019: Deepening Mathematical Concepts for Wider Application through Multidisciplinary Research and Industries Collaborations*, 2019.

[17] H. Z. D. Li, T. Li, A. Bouras, X. Yu and T. Wang, "Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set," *IEEE Transactions on Fuzzy Systems,* vol. 30, no. 5, pp. 1396-1408, 2022.

[18] 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.

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

[20] 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, 2017.

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

[22] 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.

[23] 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.

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

[25] J. S. Rojo-Alvarez, M. Martinez-Ramon, J. Munoz-Mari and G. Camps-Valls, Digital Signal Processing with Kernel Methods, Hoboken: John Wiley & Sons, 2018.

[26] 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.

[27] M. A. Tuegeh, A. Soeprijanto and M. H. Purnomo, "Optimal Generator Scheduling Based on Particle Swarm Optimization," in *Seminar Nasional Informatika (SemnasIF)*, 2009.

[28] J. R. Quinlan, "Induction of Decision Tree," *Machine Learning,* pp. 81-106, 1986.