Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis

Sugiyarto Surono & Rizki Desia Arindra Putri

ONLINE FIRST

International Journal of Fuzzy Systems

ISSN 1562-2479

Int. J. Fuzzy Syst.
DOI 10.1007/s40815-020-00997-5

Springer

Int. J. Fuzzy Syst.
https://doi.org/10.1007/s40815-020-00997-5

Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis

Sugiyarto Surono[1] · **Rizki Desia Arindra Putri**[1]

Received: 21 August 2020 / Revised: 1 October 2020 / Accepted: 27 October 2020

Abstract Optimization is used to find the maximum or minimum of a function. In this research, optimization is applied to the objective function of the FCM algorithm. FCM is an effective algorithm for grouping data, but it is often trapped in local optimum solutions. Therefore, the similarity measure in the clustering process using FCM is very important. This study uses a new method, which combines the Minkowski distance with the Chebyshev distance which is used as a measure of similarity in the clustering process on FCM. The amount of data that is quite large and complex becomes one of the difficulties in providing analysis of multivariate data. To overcome this, one of the techniques used is dimensional reduction using Principal Component Analysis (PCA). PCA is an algorithm of the dimensional reduction method based on the main components obtained from linear combinations, which can help stabilize cluster analysis measurements. The method used in this research is dimensional reduction using PCA, clustering using FCM with a combination of Minkowski and Chebyshev distances (FCMMC), and clustering evaluation using the Davies Bouldin Index (DBI). The purpose of this research is to minimize the objective function of FCM using new distances, namely, the combination of Minkowski and Chebyshev distances through the assistance of dimensional reduction by PCA. The results showed that the cluster accuracy of the combined application of the PCA and FCMMC algorithms was 1.6468. Besides, the minimum value of the combined objective function of the two methods is also obtained, namely, 0.0373 which is located in the 15th iteration, where this value is the smallest value of the 100 maximum iterations set.

Keywords PCA · Fuzzy C-Means · Minkowski and Chebyshev distance · Davies Bouldin Index

1 Introduction

Optimization is everywhere, although it can mean different things from another perspective. From basic calculus, optimization can be easily used to find the maximum or minimum of a function [15]. Techniques in optimization with or without using gradient information, depending on the suitability of the problems that arise [15]. Unconstrained optimization is a problem consideration in performing objective functions on real variables without constraints on values. The simple unconstraint optimization problem allows the maxima or minima of a univariate function $f(x)$ to be $-\infty < x < +\infty$ (or the entire real domain $\mathbb{R}$); it can be written [15] as follows:

$$\max \text{ or } \min f(x), x \in \mathbb{R},$$

for an unconstrained optimization problem, optimally occurs at the critical point given the stationary condition $f'(x) = 0$. However, this stationary state is just a necessary condition, not a sufficient condition. If $f'(x_*) = 0$ and $f''(x_*) > 0$, this is the local minimum. Conversely, if $f'(x_*) = 0$ and $f''(x_*) < 0$, then this is the local maximum.

Whatever the problem in the real world, it is usually possible to formulate a constraint optimization problem in the general form [15]:

✉ Sugiyarto Surono
sugiyarto@math.uad.ac.id

[1] Departement of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia

$$\max / \min_{x \in \mathbb{R}} f(x), x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n,$$

subject to

$$h_i(x) = 0, (i = 1, 2, 3, \ldots, M),$$

$$g_j(x) \leq 0, (j = 1, \ldots, N),$$

where $f(x), h_i(x)$, and $g_j(x)$ are scalar functions of the real vector x. There are M equality constraints and N inequality constraints. In general, all function problems $[f(x), h(x)$, and $g(x)]$ are nonlinear, requiring sophisticated optimization techniques.

Data with large dimensions, when the algorithm is applied to the grouping method, usually take time to require. Therefore, the dimensional reduction can be used to overcome this. The dimension reduction method used in this study is Principal Component Analysis (PCA). PCA is an unsupervised multivariate analysis method that is effective for extracting high-dimensional data into low-dimensional space without data variables, using linear algebraic techniques to reduce dimensions that have interconnected variables into new data with unrelated ones called the main components [4]. Even though the data dimensions are smaller, there will not be much information because the variance of data is maintained at least 70–80% [13].

Several studies related to the application of PCA to improve the performance of the clustering method have been carried out before, for example, a research conducted by Ref. [14]. The research resulted in high performance in determining the detail of fusion images. Another research was also conducted by Ref. [11], which was applied to the field of image processing and produced a level of 92.5%. The application of PCA which was then followed by the Fuzzy C-Means method was also carried out by previous researchers, namely Ref. [2]. This study discusses the algorithm for monitoring stations to monitor various sources of air pollution, as well as checking the zone in pollution. In this research, PCA is used as a method for monitoring the sources of air pollution, while the Fuzzy C-Means algorithm is used for the grouping stage of the monitoring stations. This study uses Euclidean distance as a measure of similarity and dissimilarity between objects in the dataset. There is also another research conducted by Ref. [16]. The research proposed an approach of mechanical failure information extraction and recognition in the early fault state variables, combined with the Principal Component Analysis (PCA) algorithm with FCM algorithm. The application results of this method to identify variables deviation fault rotor test bed are acceptable.

Research on the application of Principal Component Analysis (PCA) has been conducted by previous researchers, including research by Ref. [1]. This research resulted in a high accuracy rate of 96.07% by combining PCA and Levenberg–Marquardt backpropagation (LMBP). Also, there is other research about PCA by Ref. [8]. This study aims to examine network interference. The results obtained are as many as 10 main components with an accuracy rate of 99.7%. Another study was also conducted by Ref. [3]. This study uses PCA as a method in reducing the dimensions of features in animals so that the output of PCA will be used in the classification using the Logistic Regression (LR) method. The accuracy obtained from the research is 90%.

Fuzzy C-Means is a very effective algorithm, but the random values at the center point make iterative processes falling on the optimal local solution easily. The choice and size of inequality play an important role in the cluster structure of the data.

The use of the Fuzzy C-Means algorithm in the clustering process has been carried out by several researchers before, one of which is the research conducted by Ref. [6], who is engaged in the health sector to discuss the segmentation of brain tumors in patients using the Fuzzy C-Means and K-Way algorithms. The results obtained are that Fuzzy C-Means have better K-Means performance because the resulting segmentation is more precisely identified.

In this study, researchers used the new distance proposed by [12] by combining the Minkowski and Chebyshev distances. This distance has been applied to the classification using the K-Nearest Neighbors (KNN) method and produces a high degree of accuracy [12].

2 Preliminaries

2.1 Dimensional Reduction

Dimensional reduction is a method of reducing dimensions on a dataset by applying certain considerations [5], namely, maintaining the information contained in the dataset even though the data are undergoing a reduction process, and producing a smaller number of variables that will simplify and speed up the process computing and data visualization. Dimensional reduction can help stabilize the measurement for additional statistical analysis, such as regression analysis or cluster analysis.

PCA is an unsupervised multivariate analysis method that is effective for extracting high-dimensional data into low-dimensional space without losing data characteristics, using linear algebraic techniques to reduce dimensions that have interrelated variables into new data with unrelated variables called principal components [4]. Even though the data dimensions are smaller, it will not lose much information because data variance is maintained at least

70–80%. PCA is done to obtain the main components (principal components) that can explain most of the variance of data or can be said to be able to maintain most of the information measured using only a few variables which are the main components.

The calculation of PCA is based on the calculation of eigenvalues and eigenvectors which indicate the magnitude of the spread of data in a dataset. So, by using PCA, the initial variable is p variable will be reduced to k new variable which is called the principal component, with $k < p$. Even though it only uses k principal components, it is still able to produce the same value using the p variable [7]. The nature of the new variables formed from PCA is that the number of variables is less and has no correlation between the variables that are formed.

2.2 Clustering

Clustering is the process of grouping data based on the similarity between data in a dataset, with learning without direction or the so-called unsupervised. The clustered data will have the same level which is high in the same cluster but has a high degree of difference on different clusters. Grouping done on a dataset is very important to obtain information, and makes it easier to understand its elements. Clustering is divided into two methods, namely, hard clustering and fuzzy clustering. In this hard clustering algorithm, several k-clusters will be formed, where each data will only be one cluster member. While in fuzzy clustering, each data is assumed to have the possibility to be a member of several clusters with different degrees for each group.

There are two basic methods in fuzzy clustering, the first is Fuzzy C-Means (FCM), so named because this clustering algorithm will be formed as many c-clusters that have been previously determined. Furthermore, the second method is Fuzzy Subtractive Clustering (FSC), where in this method the number of clusters is not predetermined [9].

Fuzzy C-Means (FCM) was first introduced by Jim Bezdek in 1981 which is an algorithm of clustering using fuzzy grouping, so that data can be members of all classes or clusters with different degrees of 0 to 1. The concept of FCM first determines the cluster location that will mark the average for each cluster, with the initial conditions being inaccurate. Each data point has a minimum degree for each cluster. In this condition, the cluster center and each data level return repeatedly, so it can be seen that the cluster center will move to the right location. This iteration is based on minimizing the objective function assigned to the cluster center weighted by the degree of each data.

2.3 Combined Distances Between Minkowski and Chebyshev

Rodrigues [12] raises a new distance, namely, the combination of Minkowski and Chebyshev distances. The combination of Minkowski and Chebyshev distances is shown in the equation below:

$$d_{(w_1,w_2,p)}(x,y) = w_1 \sqrt[p]{\sum_{k=1}^{n} |x_k - y_k|^p} + w_2 \max_{k=1}^{n} |x_k - y_k|, \tag{1}$$

where x_k and y_k are the x and y values in the n dimension, respectively. When w_1 is greater than $w_{2,}$ the distance is more like Minkowski. Conversely, if w_2 is greater than $w_{1,}$ then the distance is more like Chebyshev.

2.4 Validity Index

Davies Bouldin Index (DBI) is one of the methods used to measure cluster validity in a clustering method. The purpose of measurement with DBI is to maximize the distance between clusters (inter-cluster) and to minimize the distance between data points (intra-cluster) in the same cluster. If the inter-cluster distance is maximum (large value), the similarity in characteristics between each cluster tends to be small, so that the differences between clusters are more pronounced. If the intra-cluster distance is minimum, then each data in the cluster has a high degree of characteristic similarity. A cluster will be considered to have an optimal clustering scheme if it has a minimal Davies Bouldin Index (DB) (close to 0) [10].

$$\mathrm{SSW}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d\left(x_j, c_i\right), \tag{2}$$

$$\mathrm{SSB}_{i,j} = d\left(c_i, c_j\right), \tag{3}$$

$$R_{i,j} = \frac{\mathrm{SSW}_i + \mathrm{SSW}_j}{\mathrm{SSB}_{i,j}}, \tag{4}$$

$$\mathrm{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(R_{i,j}\right). \tag{5}$$

3 Results and Discussion

The FCMMC algorithm is applied to the result of dimensional reduction using PCA which has been calculated previously. Dimensional reduction using PCA was applied to hypertensive patient data with 5 original variables, resulting in 3 principal components (main components) as a linear combination between loading values and standardized data. Furthermore, the values of each data contained in the 3 main components are used as input in the clustering process using the Fuzzy C-Means method, a combination of the Minkowski and Chebyshev distance (FCMMC). Input data consisted of 100 patient data samples with 3 variables, namely, PC_1, PC_2, and PC_3. The data will be grouped into 3 classes, namely, class 1 for mild hypertension type, class 2 for moderate hypertension type, and class 3 for severe hypertension type. The following will display the data from the dimensional reduction by PCA (Table 1).

The principle of the Fuzzy C-Means algorithm is to minimize an objective function. The basic concept of Fuzzy C-Means first determines the center of the cluster which will mark the average location for each cluster. At initial conditions, the center of this cluster is still inaccurate. Each data point has a degree for each cluster. By repeating the central cluster and the degree of each data point, it can be seen that the center of the cluster will move toward the right location. This iteration is based on minimizing the objective function which describes the distance from a given data point to the center of the cluster which is weighted by the degree of the data point [9]. The objective functions used in FCM with a combination of Minkowski and Chebyshev distances are as follows:

$$\min J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\left(w_1\sqrt[p]{\sum_{j=1}^{l}\left|x_{kj}-v_{ij}\right|^p} + w_2\max_{j=1}^{l}\left|x_{kj}-v_{ij}\right|\right),$$

subject to $\sum_{i=1}^{c}(\mu_{ik})^m = 1,$

$0 < \mu_{ik} < 1,$

where $J_m(U,V)$ is the objective functions with respect to U and V; c is the number of clusters; n is the amount of data; m is the weighter rank, $m \in [1,\infty)$; U is the initial partition matrix; V is the cluster center matrix; μ_{ik} is the degree of kth data membership in the ith cluster; w_1 is the Minkowski weights; w_2 is the Chebyshev weights; x_{kj} is the kth data in the j variable; v_{ij} is the center of the ith cluster in the jth variable; p is the Minkowski parameters.

The initial defined values for the grouping of data resulting from the reduction of dimensions are as follows:

Table 1 Data from PCA dimension reduction

No.	PC_1	PC_2	PC_3
1	1.0952	1.3483	− 0.1778
2	− 1.3153	0.1893	1.7954
3	0.5015	− 0.7913	− 0.4487
⋮	⋮	⋮	⋮
99	− 0.3543	0.9746	0.3117
100	− 0.3945	− 0.0928	− 0.9341

$c = 3, w_1 = 3,\ w_2 = 2, p = 4, m = 2,\ \text{maxiter} = 100, \varepsilon = 10^{-5}$. Then, we obtain the initial partition random matrix that meets the constraints $\sum_{i=1}^{c}(\mu_{ik})^m = 1$. The partition matrix is shown as follows:

$$U^{(0)} = \begin{bmatrix} 0.1789 & 0.0786 & 0.7425 \\ 0.4496 & 0.3524 & 0.1980 \\ 0.1252 & 0.2773 & 0.5975 \\ \vdots & \vdots & \vdots \\ 0.2688 & 0.0819 & 0.6493 \\ 0.5351 & 0.4130 & 0.0520 \\ 0.7809 & 0.1588 & 0.0602 \end{bmatrix}.$$

Each data object has a certain degree of membership in each cluster. The value of the largest degree of membership shows the tendency of the data object to become a member of the cluster.

Furthermore, the cluster center is obtained for the first iteration, namely,

$$v_{ij}^{(1)} = \begin{bmatrix} -0.0550 & 0.0775 & 0.0152 \\ 0.0798 & -0.1368 & 0.0179 \\ 0.1184 & -0.0334 & -0.0164 \end{bmatrix}.$$

where $i = 1,2,3$ and $j = 1,2,3$.

The objective function that results from the first iteration is

$$F^{(1)} = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\left(w_1\sqrt[p]{\sum_{j=1}^{l}\left|x_{kj}-v_{ij}\right|^p} + w_2\max_{j=1}^{l}\left|x_{kj}-v_{ij}\right|\right) = 330.3307.$$

It is found that $\left|F^{(1)}_{\text{FCMMC}}(U,V) - F^{(0)}_{\text{FCMMC}}(U,V)\right| = 330.3307 > \varepsilon$; then proceed to iteration 2. Update the new partition matrix and calculate the center cluster in the second iteration using a sequential equation:

$$\mu_{ik} = \frac{\left(\frac{1}{\left(w_1\sqrt[p]{\sum_{k=1}^{n}\left|x_{kj}-v_{ij}\right|^p}+w_2\max_{k=1}^{n}\left|x_{kj}-v_{ij}\right|\right)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{n}\left(\frac{1}{\left(w_1\sqrt[p]{\sum_{k=1}^{n}\left|x_{kj}-v_{ij}\right|^p}+w_2\max_{k=1}^{n}\left|x_{kj}-v_{ij}\right|\right)}\right)^{\frac{1}{m-1}}},$$

$$v_{ij} = \frac{\sum_{k=1}^{n}\mu_{ik}^m x_{kj}}{\sum_{k=1}^{n}\mu_{ik}^m}.$$

The iteration process will stop when it meets the stop condition, when $t > \text{maxiter}$ or when $\left|F^{(t)}_{\text{FCMMC}}(U,V) - F^{(t-1)}_{\text{FCMMC}}(U,V)\right| < \varepsilon$. The application of the FCMMC method, which is carried out on the result of dimensional reduction with PCA, undergoes an iteration process stop when t = 100, this happens because it has met the stopping condition, namely, $t > \text{maxiter}$.

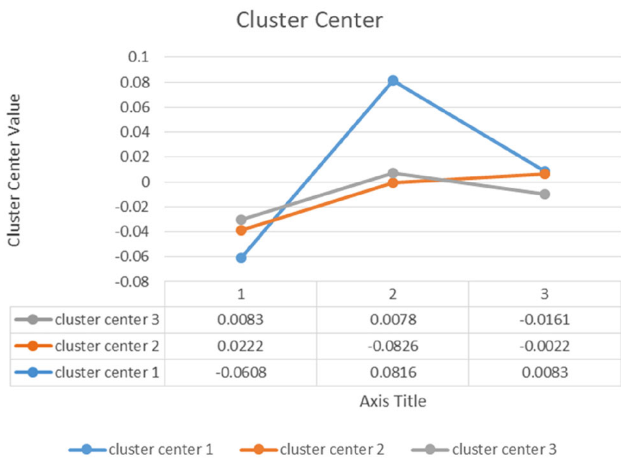

Fig. 1 Centroid at t = 100

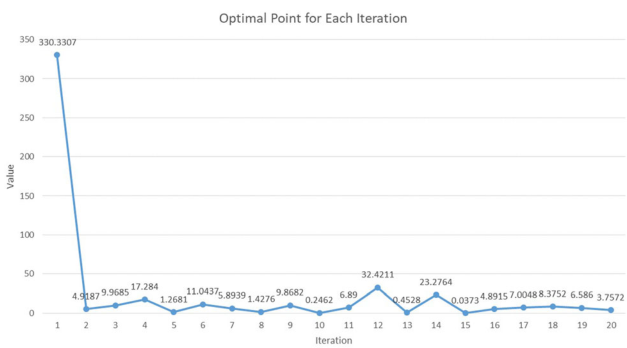

Fig. 2 Value of the objective function of each iteration

The cluster center output in the last iteration when t = 100 was as follows:

$$v_{ij}^{(100)} = \begin{bmatrix} -0.0608 & 0.0816 & 0.0083 \\ 0.0222 & -0.0826 & 0.0078 \\ 0.0406 & -0.0022 & -0.0161 \end{bmatrix}, \quad \begin{matrix} i = 1,2,3 \\ j = 1,2,3. \end{matrix}$$

The graph of the cluster center in the last iteration, when t = 100, is shown in Fig. 1.

Fuzzy C-Means has a weakness that is trapped in an optimum local solution. Of the 100 iterations carried out, the clustering process reached the minimum point at the time of the 15th iteration, with a value of 0.0373. The graph of the objective function values for each iteration is shown in Fig. 2.

As for the new partition matrix generated in the last iteration, when t = 100 is

$$U^{(0)} = \begin{bmatrix} 0.3496 & 0.3142 & 0.3364 \\ 0.3386 & 0.3340 & 0.3273 \\ 0.2892 & 0.3760 & 0.3348 \\ \vdots & \vdots & \vdots \\ 0.3341 & 0.3208 & 0.3451 \\ 0.3715 & 0.2997 & 0.3289 \\ 0.3319 & 0.3288 & 0.3394 \end{bmatrix}.$$

So, from the new partition matrix generated in the last iteration when t = 100, the following clustering is obtained (Table 2).

The results of clustering using the Fuzzy C-Means algorithm with a distance between Minkowski and Chebyshev (FCMMC), namely, the data in group 1 are included in the mild hypertension class, data in group 2 are included in the moderate hypertension class, and data contained in group 3 are included in class severe hypertension.

Each data resulted from the FCMMC clustering process, each grouped according to the resulting cluster. The SSW values are obtained as follows:

$$SSW_1 = 5.4171$$
$$SSW_2 = 3.9744$$
$$SSW_3 = 7.6456.$$

Furthermore, the value of SSB obtained from the calculation results is

$$SSB_{12} = 6.3353$$
$$SSB_{13} = 7.5549$$
$$SSB_{23} = 8.5845.$$

Thus, the DBI value obtained is 1.6468. This DBI value is a measure of the accuracy of the clustering results.

4 Conclusion

In this paper, we introduce a new distance, namely, the combination of Minkowski and Chebyshev distances as a similarity measure applied to the clustering process using

Table 2 FCMMC clustering results

Data	Group
1, 2, 4, 6, 7, 8, 9, 10, 11, 14, 19, 22, 25, 26, 27, 29, 34, 35, 38, 41, 48, 49, 51, 54, 57, 58, 59, 63, 66, 67, 68, 69, 70, 71, 73, 76, 77, 79, 80, 81, 82, 83, 84, 87, 93, 94, 95, 96, 97, 99	1
33, 36, 37, 42, 56, 92	2
3, 5, 12, 13, 15, 16, 17, 18, 20, 21, 23, 24, 30, 31, 32, 39, 40, 43, 44, 45, 46, 47, 50, 52, 53, 55, 60, 61, 62, 64, 65, 72, 74, 75, 78, 85, 86, 88, 89, 90, 91, 98, 100	3

Fuzzy C-Means. Besides, this paper also combines the PCA algorithm into Fuzzy C-Means with the Minkowski and Chebyshev Distance Combination, as an effort to optimize the clustering process. The applications of these methods are new because previously there were no researchers who combined the two methods.

Based on the results and discussion of this study, it can be concluded that the PCA calculation produces 3 new variables as the main component (principal component) of the 5 variables studied, where the main component is the factor that most influences one's hypertension. Furthermore, the clustering accuracy obtained for the FCMMC which is applied to the dimensional reduction data with PCA has an accuracy value of 1.6468 and the minimum point of 100 iterations in the FCMMC process is achieved at the time of the 15th iteration which is 0.0373.

References

1. Adiwijaya, W.U., Lisnawati, E., et al.: Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. J. Comput. Sci. **14**(11), 1521–1530 (2018)
2. Dogruparmak, S.C., et al.: Using principal component analysis and fuzzy C-means clustering for the assesment of air quality monitoring. Atmos. Pollut. Res. **5**(4), 656–663 (2014)
3. Doraikannan, S., Prabha, S., Vijay, K.B.: Principal component analysis for dimensionality reduction for animal classification based on LR. IJITEE (2019). https://doi.org/10.35940/ijitee.J8805.0881019
4. Farjo, J., et al.: Does principal component analysis improve cluster-based analysis? In: 2013 IEEE sixth international conference on software testing, verification and validation workshops, 2013. IEEE, New York (2013)
5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, Burlington (2006)
6. Jose, A., Ravi, S., Sambath, M.: Brain tumor segmentation using K-means clustering and fuzzy C-means algorithms and its area calculation. IJIRCCE (2014). https://doi.org/10.13140/RG.2.1.3961.5841
7. Johnson, A.R., Dean, W.W.: Applied Multivariate Statistical Analysis. Pearson Practice Hall, New Jersey (2007)
8. Keerthi Vasan, K., Surendiran, B.: Dimensionality Reduction Using Principal Component Analysis for Network Instrusion Detection. Elsevier, Amsterdam (2016)
9. Kusumadewi, S., Purnomo, H.: Aplikasi Logika Fuzzy Untuk Pendukung Keputusan. Graha Ilmu, Graha Ilmu (2010)
10. Mustofa, Z., Imam, S.S.: Algoritma clustering K-medoids Pada e-government Bidang information and communication technology Dalam Penentuan status Edgi. Jurnal Teknologi Informasi dan Komunikasi **9**(1), 1 (2018)
11. Panna, M.B.: Human face detection based on combination of linear regression, PCA and fuzzy C-means clustering. IJCSIS **17**(7), 57–62 (2019)
12. Rodrigues, E.O.: Combining Minkowski and Chebyshev: New Distances Proposal and Survey of Distances Metrics Using K-Nearest Neighbours Classifier. Elsevier, Amsterdam (2018)
13. Timm, N.H.: Applied Multivariate Analysis. Springer, New York (2002)
14. Vijarayan, R., Muttan, S.: Fuzzy C-means clustering based principal component averaging fusion. Int. J. Fuzzy Syst. **16**(2), 153 (2014)
15. Yang, X.S.: Optimization Techniques and Applications with Example. Wiley, New York (2018)
16. Zhao, Q.: The study on rotating machinery early fault diagnosis based on principal component analysis and fuzzy C-means algorithm. JSW **8**(3), 709–715 (2013)

Sugiyarto Surono completed his bachelor degree in Mathematics in Universitas Gadjah Mada. He has also completed his master degree in Mathematics and Statistics in the same university. He has completed his doctoral degree in Universiti Teknologi Malaysia, majored in optimization. Currently, he is working as a lecturer in Faculty of Science and Applied Technology Ahmad Dahlan University while doing several research in fuzzy for machine learning and deep learning.

Rizki Desia Arindra Putri, was born on 26/12/1997 at Yogyakarta, Indonesia. She obtained her Bachelor's degree in Mathematic from Ahmad Dahlan University, Yogyakarta, Indonesia in 2020. Her research interest includes Fuzzy logic, Mathematical modelling with numerical approach.