EKSAKTA
Journal of Sciences and Data Analysis

**E-ISSN**: 2720-9326
**P-ISSN**: 2716-0459
DOI: 10.20885/EKSAKTA.vol2.iss2.art1

Research Article

# Implementation of Minkowski-Chebyshev Distance in Fuzzy Subtractive Clustering

**Annisa Eka Haryati[1] , Sugiyarto[2,*], Suparman[3]**

[1,3] Masters in Mathematics Education, Universitas Ahmad Dahlan Yogyakarta
[2] Mathematics Faculty of Applied Science and Technology Universitas Ahmad Dahlan Yogyakarta
[*] *Corresponding author: sugiyarto@math.uad.ac.id*

Received: 31 May 2021; Accepted: 6 June 2021; Published: 23 June 2021

**Abstract:** Clustering is a method of the grouping which is done by looking at the similarities between data in a data set. Fuzzy clustering is a clustering method that uses fuzzy set membership values as the basis for grouping data. Fuzzy Subtractive Clutering (FSC) is a fuzzy clustering method where the number of clusters to be formed is unknown. The concept of FSC is to determine the highest data density and the data with the most number of neighbors will be selected as the center of the cluster. Thus, the size of the proximity or distance between points is needed to determine the members of each cluster. The distance used in this study is a combination of the Minkowski and Chebyshev distances. The number of clusters formed will be evaluated using the Partition Coefficient (PC) value where the highest PC value indicates the best number of clusters. The results obtained indicate that the best clusters are three clusters with a PC value of 0.7422.

**Keywords**: Clustering, Fuzzy clustering, Fuzzy subtractive clustering

## Introduction

In everyday life, a lot of data comes from various observations and measurements. These data have different characteristics. However, these data sets can be grouped by looking at the similarity of each data. Therefore, there is a method, namely clustering, which is used to group data that has various characteristics.

Clustering is a grouping that is done by looking at the similarities between data in a data set. The purpose of clustering is to classify data into several groups by looking for patterns in the data set [1], [2]. The basic concept in clustering is to classify data in clusters that are very similar to other data in the same cluster, but different from data contained in other clusters. Therefore, the higher the similarity of data in one cluster, the better the cluster formed [3].

Fuzzy clustering is a clustering method that uses fuzzy set membership values as the basis for grouping data. Each record will have the possibility of being a member of several groups. This means that each data is not a member of one group [4]. There are several methods that can be used to perform fuzzy clustering, including Fuzzy C-Means and Fuzzy Subtractive Clustering.

Fuzzy Subtractive clustering (FSC) is a fuzzy clustering method where the number of clusters to be formed is unknown. The randomized initialized membership matrix is not used in this method, so the results will be more consistent [5], [6]. The basic concept of the FSC method is to determine the highest data density and the data with the most number of neighbors will be selected as the center of the cluster. Then, the data point that becomes the center of the cluster will be reduced in density and the algorithm will select another data point that has the most neighbors to become another cluster center [7], [8].

Several studies have been carried out using the Fuzzy Subtractive Clustering method, other studies have been conducted by [9] used to classify polymer candidates based on some similarity in interaction with the chemical targeted for sensing. The results obtained were compared with the Fuzzy C-Means method where the FSC method produced a better selection than the FCM method in this study. In 2019, fuzzy subtractive clustering is used by [10] which is combined with particle swarm optimization to perform classification. Other research was also carried out by [11] with the entropy-based fuzzy subtractive clustering method to identify the multi-model algorithm structure. Fuzzy subtractive clustering has also been used by [12] to make predictions on the stock market. Besides, fuzzy subtractive clustering is also used by [13] to analyze patterns of changes in value movements in sales data.

The fuzzy subtractive clustering method requires a similarity measure to determine the number of points that have the most neighbors. The distance most often used to determine the measure of similarity is the Euclidean distance. Therefore, this study will use a combination of the Minkowski and Chebyshev distances proposed by Rodrigues. This distance has been applied by [14] to perform classification using K-Nearest Neighbors (KNN) and produce a high degree of accuracy. The FSC method is used on data that does not have a certain class. In addition, [15] it also applies a combination of minkowski chebyshev distances for clustering cases using Fuzzy C-Means. Based on the description above, the researcher will conduct research using the fuzzy subtractive clustering method with a combination of Minkowski and Chebyshev distances.

## Methods

In this study, the approach used is a quantitative approach. A quantitative approach is an approach used in a study where the data to be analyzed is in the form of numbers (numeric). Besides, this research will also use more tables or graphs to display the results of data analysis. The subject used is one of the UCI Machine Learning data totaling 589 data.

The data used in this research is quantitative data. Quantitative data is data that can be measured directly or an explanation is expressed in numerical form. In this study, the quantitative data used were age $(X_1)$, gender $(X_2)$, ALB $(X_3)$, ALP $(X_4)$, ALT $(X_5)$, AST $(X_6)$, BIL $(X_7)$, CHE $(X_8)$, CHOL $(X_9)$, $GREA$ $(X_{10})$, GGT $(X_{11})$, $and\ PROT$ $(X_{12})$.

The method used in this study is fuzzy subtractive clustering using a combination of Minkowski and Chebyshev distances. The cluster evaluation that will be used is the Partition Coefficient (PC) [16]. Data processing was carried out with the help of Jupyter Notebook Software with the Python programming language.

Combination of Minkowski and Chebyshev distances with weights $w_1$ and $\mathrm{w}_2$ given in the equation of the distance function defined in the equation [14]:

$$d_{(w_1,w_2,p)}(x,y) = w_1 \sqrt[p]{\textstyle\sum_{m=1}^{k} |x_m - y_m|^p} + w_2\, max_{m=1}^{k} |x_m - y_m| \tag{1}$$

where $w_{1,}\mathrm{w}_2 > 0$ and $p \geq 1$.

The steps in this research are as follows [17]:

1. Determine the parameter values, namely radius (r), squash factor (q), accept ratio (ar), reject ratio (rr).
2. Converting natural numbers to fuzzy number form with the following equation [18]:

$$(x) = \begin{cases} 1 & x \leq a \\ \dfrac{e^{-\left(\frac{x-a}{b-a}\right)} - e^{-s}}{1-e^{-s}} & a \leq x \leq b \\ 0 & x \geq b \end{cases} \tag{2}$$

   where a and b are the smallest and greatest values of the data.
3. Determine $D_i; i = 1,2,3, \dots, n$ or potential each data point with the following steps:
   Step 1: calculate the distance for each data using the following equation:

$$Dist_{ij} = \left( \frac{w_1 \sqrt[p]{\sum_{m=1}^{k} |x_m - y_m|^p} + w_2\, max_{m=1}^{k} |x_m - y_m|}{r} \right) \tag{3}$$

   Minkowski and Chebyshev distances are used to calculate the distance between data points. Then, this distance will be used to calculate the potential data using step 2.

   Step 2: determine the initial potential of each data point using the following equation:

$$D_i = \textstyle\sum_{k=1}^{n} e^{-4\left(\sum_{j=1}^{m} Dist_{ij}^2\right)} \tag{4}$$

   where $D_i$ is the $i$-th data potential.
4. Determines the greatest potential value at each data point:
   $M = max\,[D_i | i = 1,2, \dots,];$ for the first iteration.

$Z = max[D_i|i = 1,2,\dots,]$; for the second, third iteration, and so on.

5. Calculating the ratio (R) of prospective cluster centers using the equation:

$$R = \frac{Z}{M} \tag{5}$$

$R$ = the ratio between the greatest potential in the first iteration with the greatest potential in the next iteration.
In the first iteration, Z = M.

6. Checking the prospective cluster center to become a cluster center by considering the following conditions:
Condition 1: if the ratio> accept ratio, then the prospective cluster center is accepted as the new cluster center.
Condition 2: If the reject ratio <ratio ≤ accept ratio, then it will be checked the feasibility of the cluster center claon. If the prospective cluster center cannot become a cluster center, the iteration is terminated because there are no more data considered to become a cluster center candidate. The steps in condition 2, that is:
For $k = 1,2,\dots,p$, where p = number of clusters

$$Sd_k = \sum_{j=1}^{m}\left(\frac{V_j - C_{kj}}{r}\right)^2 \tag{6}$$

$Sd_k$ = the distance between the prospective cluster center and the previous cluster center.
$V_j$ = prospective cluster center.
$C_{kj}$ = the center of the k-th cluster in the j-th variable.
if $(Md < 0)$ or $(Sd_k < Md)$, then $Md = Sd_k$,

$$Mds = \sqrt{Md} \tag{7}$$

where Mds is the closest distance between the prospective cluster center data and the cluster center. If $(rasio + Mds) \geq 1$ ; the prospective cluster center is accepted as the new cluster center. Meanwhile, if $(ratio + Mds) < 1$ then the prospective cluster center is not accepted and will not be reconsidered as a new cluster center (data potential is set equal to zero).
Condition 3: If the ratio is ≤ reject ratio, then the iteration will stop and no more data will become the center of the cluster.

7. If a new cluster center has been obtained, then the data potential around the previous cluster center is reduced using equation (8)

$$D_i^t = D_i^{t-1} - D_{c_{ki}} \tag{8}$$

where $D_{c_{ki}}$ is as follows:

$$D_{c_{ki}} = Z * e^{-4\left[\sum_{j=1}^{m}\left(\frac{C_{kj}-x_{ij}}{r*q}\right)^2\right]} \tag{9}$$

$D_i^t$ = the potential of the i-th data in the t-iteration.
$D_i^{t-1}$ = the potential of the i-th data in the iteration (t-1).
$D_{c_{ki}}$ = potential k-data in the iteration.
$C_{kj}$ = the center of the k-th cluster in the j-th variable.
$x_{ij}$ = the ith data in the j-th variable.
$r$ = radius.
$q$ = squash factor.

8. Calculate the sigma cluster value for each variable using the equation (10):

$$\sigma_j = \frac{r*\left(X_{max_j} - X_{min_j}\right)}{\sqrt{8}}, j = 1,2,\dots,m \tag{10}$$

$\sigma_j$ = sigma in the jth variable.
$X_{max_j}$ = the largest value in the j-variable.
$X_{min_j}$ = the smallest value in the j-variable.

9. Calculating the value of the degree of membership using the equation (11):

$$\mu_{k_i} = e^{-\sum_{j=1}^{m}\left(\frac{x_{ij}-C_{kj}}{\sqrt{2}\sigma_j}\right)} \tag{11}$$

$\mu_{k_i}$ = the membership value of the k-th cluster in the i-data.
$x_{ij}$ = the ith data in the j-th variable.

10. Calculate the value of the partition coefficient or Partition Coefficient (PC) to determine the best number of clusters. The PC score evaluates the degree of membership regardless of the value of the data. The quality of the clusters will be semi-better if the PC value is getting bigger (closer to 1). This index measures the amount of overlap between groups. PC index equation by [16] sebagai berikut:

$$\mu_{k_i} = e^{-\sum_{j=1}^{m}\left(\frac{x_{ij}-C_{kj}}{\sqrt{2}\sigma_j}\right)} \tag{12}$$

where N is the number of research objects, K is the number of clusters, and μ_ij is the membership value of the ith object with the center of the j group.

## Results and Discussions

Fuzzy Subtractive Clustering algorithm is simulated on one of the data obtained from UCI Machine Learning which contains laboratory values from blood donors and hepatitis C patients The data used were 589 with 12 variables. The variables used were age $(X_1)$, gender $(X_2)$, ALB $(X_3)$, ALP $(X_4)$, ALT $(X_5)$, AST $(X_6)$, BIL $(X_7)$, CHE $(X_8)$, CHOL $(X_9)$, $GREA$ $(X_{10})$, GGT $(X_{11})$, $and$ $PROT$ $(X_{12})$. The minkowski and chebyshev distance approaches were used by [19] and [15] to do clustering with the FCM method. Therefore, because FSC is also a clustering method and uses the distance function to determine data potential, the Minkowski and Chebysev distances are applied in this study. Data processing was carried out using Python programming language.

In this study, the value of the radius of the data points around the center of the cluster to which the potential data reduction will be measured is 1.25 or also known as the squash factor (q) value. There is an accept ratio value which indicates that the lower limit value of the data point that is a candidate for the cluster center and the reject ratio which indicates that the upper limit value of the data point that is a candidate for the cluster center is not allowed to become a cluster center [13], [20]. The value of accept ratio, reject ratio and weight used in this study are 0.8, 0.2, 3.0 and 4.0. Meanwhile, the radii are r = 1.62,1.98 and 2.04. The first step is to convert the data to fuzzy numbers using equation (2) and the results are as shown in Table 1.

**Table 1**. Fuzzy numbers

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.7571 | 0 | 0.5321 | 0.8471 | 0.9672 | 0.9430 | 0.9499 | 0.5134 | 0.6896 | 0.8617 | 0.9815 | 0.3035 |
| 0.7571 | 0 | 0.5321 | 0.7857 | 0.9188 | 0.9304 | 0.9766 | 0.2435 | 0.4690 | 0.9055 | 0.9731 | 0.1577 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 0.2584 | 1 | 0.7010 | 0.7295 | 0.9874 | 0.6112 | 0.6791 | 0.9749 | 0.6293 | 0.9156 | 0.8604 | 0.0663 |

The cluster centers obtained are as follows:

$$C_{2.04} = \begin{bmatrix} 0.7571 & 0 & 0.5321 & 0.8471 & 0.9672 & 0.9430 & 0.9499 & 0.5134 & 0.6896 & 0.8617 & 0.9815 & 0.3035 \\ 0.1863 & 1 & 0.6450 & 0 & 0.9758 & 0.5689 & 0.6670 & 0.6174 & 0.2941 & 0.9311 & 0 & 0.3142 \end{bmatrix}$$

$$C_{1.98} = \begin{bmatrix} 0.7571 & 0 & 0.5321 & 0.8471 & 0.9672 & 0.9430 & 0.9499 & 0.5134 & 0.6896 & 0.8617 & 0.9815 & 0.3035 \\ 0.1863 & 1 & 0.6450 & 0 & 0.9758 & 0.5689 & 0.6670 & 0.6174 & 0.2941 & 0.9311 & 0 & 0.3142 \\ 0.2302 & 1 & 0.8999 & 0.3913 & 0 & 0.4431 & 0.9543 & 0.6368 & 0.4792 & 0.9649 & 0.4264 & 0.7145 \end{bmatrix}$$

$$C_{1.62} = \begin{bmatrix} 0.7571 & 0 & 0.5321 & 0.8471 & 0.9672 & 0.9430 & 0.9499 & 0.5134 & 0.6896 & 0.8617 & 0.9815 & 0.3035 \\ 0.1863 & 1 & 0.6450 & 0 & 0.9758 & 0.5689 & 0.6670 & 0.6174 & 0.2941 & 0.9311 & 0 & 0.3142 \\ 0.2302 & 1 & 0.8999 & 0.3913 & 0 & 0.4431 & 0.9543 & 0.6368 & 0.4792 & 0.9649 & 0.4264 & 0.7145 \\ 0.5308 & 1 & 0.6269 & 0.7566 & 0.9864 & 0.7880 & 0.0257 & 0.9687 & 0.4240 & 0.8846 & 0.6884 & 0.2846 \end{bmatrix}$$

The center of the cluster above is the result obtained from several radius. $C_{1.62}$ represents cluster centers of radius 1.62, $C_{1.98}$ indicates cluster centers of radius 1.98 and $C_{2.04}$ denotes cluster centers of radius 2.04. The number of clusters is indicated by the number of rows and the number of menu columns, indicating the number of variables used.

Then, the membership value of each data will be calculated using equation (10) and the following results are obtained.

**Table 2**. Membership values of $r = 2.04$

| The μ value in the cluster to | |
|---|---|
| 1 | 2 |
| 1 | 0.0391 |
| 0.8633 | 0.0399 |
| ⋮ | ⋮ |
| 0.1937 | 0.1830 |

In Table 2, the first data tends to be included in cluster 1 because the largest degree of membership in the first data is located in cluster 1. The second data tends to enter cluster 1 because the largest membership value in the second data is in cluster 1, and so on until the 589th data.

**Table 3**. Membership values of $r = 1.98$

| The μ value in the cluster to | | |
|---|---|---|
| 1 | 2 | 3 |
| 1 | 0.0320 | 0.0477 |
| 0.8556 | 0.0328 | 0.0401 |
| ⋮ | ⋮ | ⋮ |
| 0.1751 | 0.1649 | 0.1512 |

Table 3 shows that the first data goes to the first cluster. This is because the largest membership value in the first data can be in the first cluster, and so on until the 589th data.

**Table 4**. Membership values of $r = 1.62$

| The μ value in the cluster to | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 1 | 0.0059 | 0.0106 | 0.0389 |
| 0.7921 | 0.0061 | 0.0082 | 0.0256 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 0.0741 | 0.0677 | 0.0595 | 0.2707 |

Similar to Table 4, the first and second data are included in the second cluster. This is because the largest membership value in the first and second data is found in the second cluster. This was done until the 589th data.

Then, the number of clusters that have been obtained will be evaluated using equation (11). The results of the Partition Coefficient value of the number of clusters formed can be seen in the following figure.

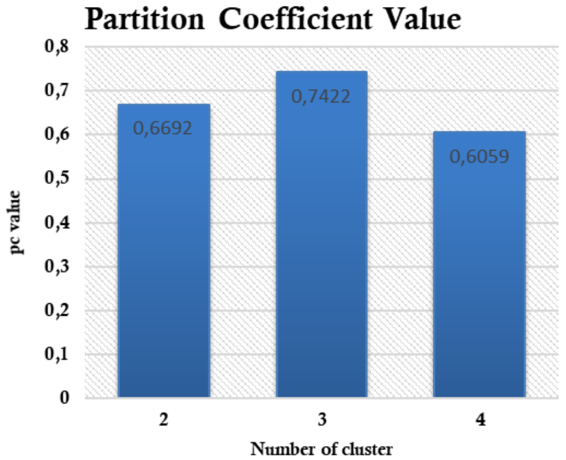

**Figure 1**. Partition Coefficient Value

Figure 1 shows the PC value of each number of clusters obtained. The PC value is used to evaluate clusters to determine which number of clusters is the best. PC value for the number of clusters 2 with a radius of 2.04 is 0.6692, the number of clusters 3 with a radius of 1.98 is 0.7422 and the number of clusters 4 with a radius of 1.62 is 0.6059.

Based on the PC value above, it can be seen that the largest PC value is 0.7422. Therefore, the number of cluster 3 has the best cluster quality because it has the largest PC value. This corresponds to [16], [21] where the greater the PC value means that the quality of the clusters obtained is getting better.

## Conclusion

This study presents a modified Fuzzy Subtractive Clustering using a combination of Minkowski and Chebysev distances. This method is used to find the best group by looking at the Partition Coefficient (PC) value. Based on the proposed method, the results obtained indicate that the best cluster is 3 clusters with a PC value of 0.7422.

## References

[1] K. M. Bataineh, M. Naji, and M. Saqer, A comparison study between various fuzzy clustering algorithms, Jordan J. Mech. Ind. Eng., 5 (4) (2011) 335–343.

[2] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*. Wiley, 2012.

[3] R. Sharma and K. Verma, Fuzzy shared nearest neighbor clustering, Int. J. Fuzzy Syst., 21 (8) (2019) 2667–2678.

[4] J. S. R. Jang, C. T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence [Book Review], IEEE Trans. Automat. Contr., 42 (10) (1997) 1482–1484.

[5] K. Benmouiza and A. Cheknane, Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting, Theor. Appl. Climatol., 137 (1–2) (2019) 31–43.

[6] S. Zeng, S. M. Chen, and M. O. Teng, Fuzzy forecasting based on linear combinations of independent variables, subtractive clustering algorithm and artificial bee colony algorithm, Inf. Sci. (Ny)., 484 (2019) 350–366.

[7] M. Ghane'i Ostad, H. Vahdat Nejad, and M. Abdolrazzagh Nezhad, Detecting overlapping communities in LBSNs by fuzzy subtractive clustering, Soc. Netw. Anal. Min., 8 (1) (2018) 1–11.

[8] R. S. Kamath and R. K. Kamat, Earthquake magnitude prediction for andaman-nicobar Islands: adaptive neuro fuzzy modeling with fuzzy subtractive clustering approach, J. Chem. Pharm. Sci., 10 (3) (2017) 1228–1233.

[9] T. Sonamani Singh, P. Verma, and R. D. S. Yadava, Fuzzy subtractive clustering for polymer data mining for saw sensor array based electronic nose, Adv. Intell. Syst. Comput., 546 (2017) 245–253.

[10] H. Salah, M. Nemissi, H. Seridi, and H. Akdag, Subtractive Clustering and Particle Swarm Optimization Based Fuzzy Classifier, Int. J. Fuzzy Syst. Appl., 8 (3) (2019) 108–122.

[11] X. Zhao and G. Yang, An entropy-based online multi-model identification algorithm and generalized predictive control, J. Intell. Fuzzy Syst., 32 (3) (2017) 2339–2349.

[12] S. K. Chandar, Stock market prediction using subtractive clustering for a neuro fuzzy hybrid approach, Cluster Comput., 22 (s6) (2019) 13159–13166.

[13] I. Sangadji, Y. Arvio, and Indrianto, Dynamic segmentation of behavior patterns based on quantity value movement using fuzzy subtractive clustering method, J. Phys. Conf. Ser., 974 (1) (2018) 0–7.

[14] O. Rodrigues, Combining minkowski and cheyshev: new distance proposal and survey of distance metrics using k-nearest neighbours classifier, Pattern Recognit. Lett., 110 (2018) 66–71.

[15] S. Surono and R. D. A. Putri, Optimization of fuzzy c-means clustering algorithm with combination of minkowski and chebyshev distance using principal component analysis, Int. J. Fuzzy Syst., (2020)

[16] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.

[17] S. Kusumadewi and H. Purnomo, *Aplikasi logika fuzzy untuk pendukung keputusan*. Yogyakarta: Graha Ilmu, 2010.

[18] K. Rezaei and H. Rezaei, New distance and similarity measures for hesitant fuzzy soft sets, 16 (6) (2019) 159–176.

[19] P. Noviyanti, *Fuzzy c-Means Combination of Minkowski and Chebyshev Based for Categorical Data Clustering*. Yogyakarta: Universitas Ahmad Dahlan, 2018.

[20] N. Azizah, D. Yuniarti, and R. Goejantoro, Penerapan Metode Fuzzy Subtractive Clustering (Studi Kasus: Pengelompokkan Kecamatan di Provinsi Kalimantan Timur Berdasarkan Luas Daerah dan Jumlah Penduduk Tahun 2015), J. EKSPONENSIAL, 9 (2) (2018) 197–206.

[21] V. Utomo and D. Marutho, Measuring hybrid sc-fcm clustering with cluster validity index, 2018 Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2018, (C) (2018) 322–326.