Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210204

An Optimization of Several Distance Function on Fuzzy Subtractive Clustering

Sugiyarto SURONO[a,1] and Annisa Eka HARYATI[b], Joko ELIYANTO [b]
[a] *Department of Mathematics, FAST Ahmad Dahlan University Yogyakarta, Indonesia*
[b] *Magister Math. Education Ahmad Dahlan University Yogyakarta, Indonesia*

Abstract. Fuzzy Subtractive Clustering (FSC) is a technique of fuzzy clustering where the cluster to be formed is unknown. The distance function in the FSC method has an important role in determining the number of points that have the most neighbors. Therefore, this study uses several distance functions. The results obtained indicate that the DBI results indicate that the Euclidean distance has a good cluster evaluation result in the number of clusters 4. Meanwhile, for the PC value the combination of the Minkowski Chebysev distance produces a good PC value in the number of clusters 2.

Keywords. Fuzzy subtractive clustering, partition coefficient, Davies-Bouldin index

1. Introduction

The classification method by applying the similarities between data in certain dataset is called clustering technique. The intention of this process is to determine certain pattern in the dataset by separating data into few number of clusters [1]. Clustering has basic concept where data in the same cluster will have high level of similarities and low level of similarities towards data in different cluster [2].

Clustering method which based on the classification with the application of fuzzy set membership degree is known as fuzzy clustering. The probability of the data to be included in one cluster is not absolute, which means data can be included as the member of one or more clusters [3].

Fuzzy subtractive clustering is one of the fuzzy clustering method where the cluster created through this method is still unidentified [4]. This method does not apply the membership matrix which initialized randomly [5]. The basic concept of this method is determining the cluster centroid through searching the data coordinate which have the most density and number of neighboring coordinates. The density of the coordinates which already selected as the cluster centroid will be reduced so that this coordinate will be ineligible to become the next cluster centroid. Then, algorithm will run again to measure the density and the number of neighboring coordinate in each data to be selected as the next cluster centroid. This iteration will be applied until all of the data in the dataset is tested [6].

There are already many research which studied FSC as its main topics. One previous research which worked by [7] applied FSC to classify Wavelet adaptive nerves to improve low-cost and high speed INS/GPS navigation system. Then, FSC was also used

[1] Corresponding Author: Sugiyarto Surono; E-mail: sugiyarto@math.uad.ac.id

by [8] to do grouping and is used by [9] to perform fuzzy-based classification. Furthermore, FSC also applied by [10] to do a classification of ad hoc mobile network based on the akaike information criteria and used by [11] to make predictions on the stock market.

Fuzzy subtractive clustering method requires the distance similarity measurement to determine the number of coordinates which have the most neighboring dots. The most common applied distance parameter to determine the similarity measurement is Euclidean distance. Aside from Euclidean distance, Minkowski distance also had been applied by [12] to do an optimization. Furthermore, Hamming distance also had been applied by [13] to detect malware in android. In addition to that, Minkowski-Chebysev combination distance also had been worked on by [14] to do a classification with KNN, by [15] to do clustering with FCM on categorical data, and by [16] to do clustering with FCM by applying dimension reduction through PCA. Based on these statement, we decided to apply fuzzy subtractive clustering method with the application of few distance function parameter. Then, the results of the cluster evaluation are used to determine which cluster has good quality.

2. Method

The method applied in this research is Fuzzy Subtractive Clustering with the application of few distance functions, such as Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev combination distance. Data simulation applied in this research is worked through Phyton programming language.

2.1. Fuzzy Subtractive Clustering (FSC)

Fuzzy Subtractive Clustering in literature [17] defined as one of the clustering method where the applied algorithm is considered as supervised algorithm. The number of cluster which will be created is initially unidentified, so it is necessary to do a radius simulation for acquiring the expected clusters. There are two comparison factors applied in FSC, which are accept ratio and reject ratio valued between 0 to 1. Accept ratio is the lower limit where certain data coordinates considered as cluster centroid candidate is allowed to be defined as cluster centroid. In the other hand, reject ratio is the upper limit where certain data coordinates considered as cluster centroid candidate is prohibited to be defined as cluster centroid.

There are 3 criteria which can occur in FSC method, there are:

- If ratio > accept ratio, then the data coordinate is accepted as new cluster centroid.
- If reject ratio < ratio, then the data coordinate will be accepted as new cluster centroid only if this new data coordinate is located far enough with the other cluster centroid. This minimum distance requirement can be measured through the addition between the ratio and the closest distance of this data coordinate to other existing cluster centroid. If the result of the addition between the ratio and the closest distance of the data coordinate to other existing cluster centroid is < 1, then this analyzed data coordinate will not be accepted as the new cluster centroid. This data coordinate also won’t be reconsidered to become the new cluster centroid (the potential value of this data will be set to 0).

- If ratio ≤ reject ratio, then there will be no another data coordinate which will be considered as the new cluster centroid. This means that the iteration process will be terminated.

2.2. Fuzzy Subtractive Clustering Algorithm

- Determining the parameter value which will be applied, such as radius (*r*), squash factor (*q*), accept ratio (*ar*), and reject ratio (*rr*).
- Transforming data into fuzzy number through this following equation [18]:

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

Where *a* and *b* are the lowest and highest value from the data.

- Determining the potential value of each coordinate $D_i; i = 1,2,3, \dots, n$ through this following equation:

$$D_i = \sum_{k=1}^{n} e^{-4\left(\sum_{j=1}^{m}\left(\frac{d_{distance}}{r}\right)^2\right)} \tag{2}$$

Where D_i is potential data *i*.

- Measuring the maximum value on each iteration and the Ratio number (*R*) with this following equation:

$$R = \frac{Z}{M} \tag{3}$$

- Testing the appropriateness of the cluster centroid candidate with 3 criteria which already mentioned before. Specifically for condition 2, to determine whether the data coordinate is fit to be considered as cluster centroid or not, this following equation needs to be applied.

$$Sd_k = \sum_{j=1}^{m}\left(\frac{V_j - C_{kj}}{r}\right)^2 \tag{4}$$

With $k = 1,2, \dots, p$ and p = the number of clusters. Sd_k is the distance between the coordinate of the cluster centroid candidate and the previous cluster centroid. V_j and C_{kj} are the cluster centroid candidate and the cluster centroid of k in variable of j. If $(Md < 0)$ or $(Sd_k < Md)$, then $Md = Sd_k$.

$$Mds = \sqrt{Md};$$

Where Mds is the closest distance between the cluster centroid candidate and other existing cluster centroid. If $(ratio + Mds) \geq 1$, then the cluster centroid candidate is accepted as the new cluster centroid. While if jika $(ratio + Mds) < 1$, then the cluster centroid candidate is not allowed to be defined as the new cluster centroid and will not be reconsidered to be the new cluster centroid (the potential value of this data will be set to 0)

- When the new cluster centroid is acquired, the subtraction of the potential value of the data around the previous cluster centroid will be initiated through this following equation [17]:

$$D_i^t = D_i^{t-1} - Z * e^{-4\left[\sum_{j=1}^{m}\left(\frac{C_{kj}-x_{ij}}{r*q}\right)^2\right]} \quad (5)$$

D_i^t = Data potential of -i in iteration of-t.
D_i^{t-1} = Data potential of -i in iteration of-(t-1).
$D_{c_{ki}}$ = Data potential of -k in iteration of-i.
C_{kj} = Cluster centroid of -k in variable of –j.
x_{ij} = Data -i in variable -j.
r = Radius.
q = Squash factor.

- Measuring the value of sigma cluster on each variables by applying an equation as follows [17]:

$$\sigma_j = \frac{r * \left(X_{max_j} - X_{min_j}\right)}{\sqrt{8}}, j = 1,2, \dots, m \quad (6)$$

σ_j = Sigma in variable of-j.

X_{max_j} = The maximum value in variable -j.

X_{min_j} = The minimum value in variable -j.

- Measuring the membership degree by applying equation (9) as follows:

$$\mu_{k_i} = e^{-\sum_{j=1}^{m}\left(\frac{x_{ij}-C_{kj}}{\sqrt{2}\sigma_j}\right)} \quad (7)$$

Where μ_{k_i} is the membership value of cluster k on data i and x_{ij} is data of i on variable of j.

2.3. Distance Function

There are few distance function applied in this research which are:

- Euclidean distance
  Euclidean distance is the most common and used distance. This distance is defined for x and y coordinate as [19]:

$$d_{euclidean}(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (8)$$

Where x_k and y_k are the value of x and y on dimension n. This distance considered as the standard distance applied for fuzzy C-means clustering method.

- Minkowski distance
MInkowski distance is defined as [20]:

$$d_{min\ kowski}(p,x,y) = \sqrt[p]{\sum_{k=1}^{n} |x_k - y_k|^p} \tag{9}$$

With p is the minkowski parameter. In Euclidean distance, $(p = 2)$. As for in Manhattan distance and Chebysev distance, $(p = 1)$ and $(p \rightarrow \infty)$ respectively. The metric condition of this function is fulfilled as long as $p \geq 1$.

- Hamming distance

Hamming distance in [18] is defined as:

$$d_{hamming}(x,y) = |x_k - y_k|,\ k = 1,2,\dots,n \tag{10}$$

With x_k and y_k are the value for x dan y in n number of data.
- Minkowski-Chebysev combination distance

Minkowski-Chebysev combination distance is defined by [14], [16] as:

$$d_{\text{minkowski-chebysev}} = t_1 \sqrt[p]{\sum_{m=1}^{k} |x_m - y_m|^p} + t_2 max_{m=1}^{k} |x_m - y_m| \tag{11}$$

With t_1 and t_2 are the weight, and x_m and y_m are the value of x and y in m number of data.

2.4. Cluster Evaluation

Evaluation cluster is applied to measure how good the quality of the formed cluster centroid. The cluster evaluation methods applied in this research are described as:

- Partition Coefficient (PC)
This cluster evaluation is invented by [21] to evaluate the data membership value on each cluster. The higher the value of PC (close to 1) indicate that the quality of the formed cluster is better. The partition coefficient for this research is conducted through this following equation:

$$PC = \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K}\mu_{ij}^2\right) \tag{12}$$

Where N is the number of research objects, K is the number of clusters, and μ_{ij} is the membership degree in i with the cluster centroid j.

- Davies Bouldin Index (DBI)

Davies Bouldin Index (DBI) is one of the method to measure the cluster validity in clustering method. The objective of this method is to maximize the distance between clusters and minimize the distance between data in the same cluster. The formed cluster will have good quality of cluster if the DBI value is minimum or close to 0 [16]. The equation of this cluster evaluation method is defined as follows:

$$DBI = \frac{1}{c}\sum_{i=1}^{c} max_{i\neq j}\left(R_{ij}\right) \tag{13}$$

With

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,J}} \tag{14}$$

$$SSW_i = \frac{1}{m}\sum_{j=1}^{m_i} d\left(x_j, c_i\right) \tag{15}$$

$$SSB_{i,J} = d\left(c_i, c_j\right) \tag{16}$$

where:
m_i = the amount of data in the i-th cluster
c_i = center of cluster i
c_j = j cluster
$d\left(x_j, c_i\right)$ = data distance with cluster center

3. Result and Discussion

The data applied for the simulation in this research is a hypertension data taken from one of the health center in Yogyakarta. There are total of 100 data with 5 main variables, which are X_1 (age), X_2 (gender), X_3 (systolic pressure), X_4 (diastolic pressure), and X_5 (body weight). Clustering process with FSC will resulting in an output of few different formed clusters. The distance function applied in this research are Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev combination distance. The result of the dataset which transformed into fuzzy number is illustrated as follows:

Table 1. Fuzzy number result

X_1	X_2	X_3	X_4	X_5
0	0	0.4609	0.1289	0.4018
0.3913	1	0.5516	0.4785	0.1468

⋮	⋮	⋮	⋮	⋮
0.5734	0	0.3775	0.2862	0.6813

Next, the data in table 1 will be processed by FSC method. In this research, radius simulation is needed to determine how many clusters are able to be acquired. The simulation result from few number of radius is illustrated in table 2 below.

Table 2. The cluster resulf for each distance function

Distance	Radius	Number of formed cluster	Time
	1.07	2	0.47
Euclidean	0.83	3	0.57
	0.67	4	0.71
	1.55	2	0.42
Minkowski	1.25	3	0.63
	1.02	4	0.90
	0.97	2	0.39
Hamming	0.79	3	0.62
	0.72	4	0.85
	1.6	2	0.43
Minkowski-chebysev	1.31	3	0.66
	1.12	4	0.78

The application of Euclidean distance resulting in the creation of 2 clusters with radius 1.07, 3 clusters with radius 0.83, and 4 clusters with radius 0.63. In Minkowski distance, there are 2 clusters, 3 clusters, and 4 clusters formed with the radius 1.55, 1.25, and 1.02 respectively. As for Hamming distance, there are 2 clusters, 3 clusters, and 4 clusters formed with the radius 0.97, 0.79, and 0.72 respectively. Lastly, the application of Minkowski-Chebysev combination distance resulting in the formation of 2 clusters, 3 clusters, and 4 clusters with the radius 1.6, 1.31, and 1.12 respectively.

Running time for each distance has a different time. Based on Table 2, in general the Euclidean distance has less running time than the other distances. This can be seen in the number of clusters 3 and 4 at the Euclidean distance each takes 0.57 and 0.71.

Furthermore, all of the cluster which already formed through the distance function are evaluated through the application of Partition Coefficient (PC) and Davies Boulding Index (DBI). This evaluation process is necessary to observe which cluster can be considered as high quality cluster. The output value of PC and DBI evaluation process can be observed in Figure 1 to Figure 3.

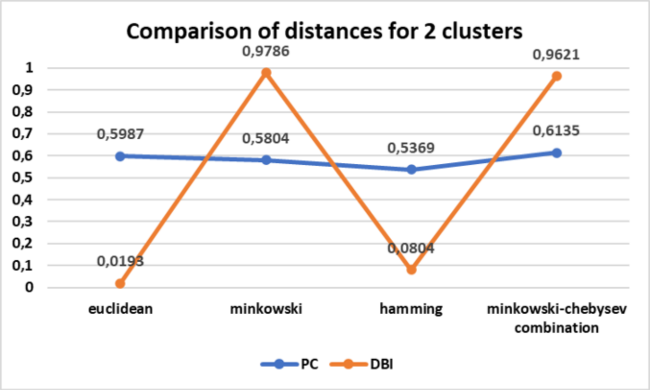

Figure 1. PC and DBI value comparison for 2 formed clusters.

Figure 1 illustrate the comparison between the PC and DBI value for 2 formed clusters in each distance function. Based on the figure 1 above, the PC value for Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev distance are 0.5987, 0.5804, 0.5369, and 0.6135 respectively. While the DBI value for each distance function are 0.0193, 0.9786, 0.0804, and 0.9621 respectively.

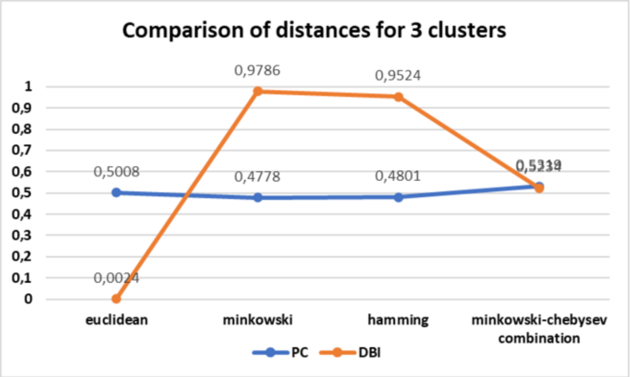

Figure 2. PC and DBI value comparison for 3 formed clusters.

Comparison of the PC and DBI value for 3 formed clusters in each of the distance function is depicted in figure 3. The PC value in the application of Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev distance are 0.5008, 0.4778, 0.4801, and 0.5319. As for the DBI value in these four distance function are 0.0024, 0.9786, 0.9524, and 0.5234 respectively.

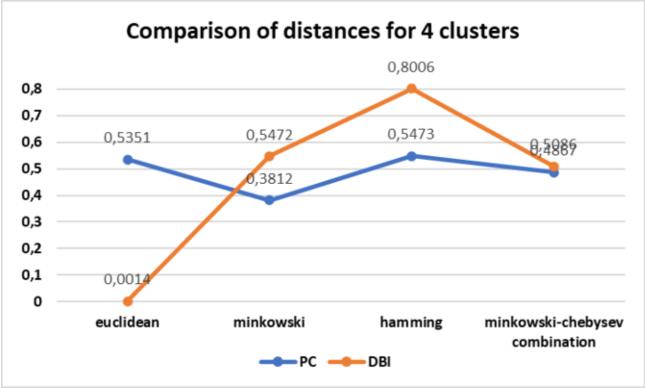

Figure 3. PC and DBI value comparison for 4 formed clusters.

Figure 3 shows the comparison of PC and DBI value for 4 formed cluster in each of the distance function. The PC value in the application of Euclidean distance is 0.5351, Minkowski distance is 0.3182, Hamming distance is 0.5473, and Minkowski-Chebysv distance is 0.4867. Furthermore, the DBI value for Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev distance are 0.0014, 0.5472, 0.8006, and 0.5086 respectively.

In this study, the proposed method will produce clusters with different radius for each distance used. The radius values must be simulated one by one to get the desired number of clusters.

4. Conclusions

This research provides the study on the application of Fuzzy Subtractive Clustering which implemented on 4 different distance function, Euclidean distance, Minkowski distance, Hamming distance, and Minkowski-Chebysev distance. The acquired clusters then evaluated through the application of Partition Coefficient (PC) method and Davies Bouldin Index (DBI) method. In this research, the DBI result in the application of Euclidean distance provides good cluster evaluation value for the 4 formed clusters with DBI 0.0014. This conclusion is suitable with the criteria in DBI where the lower the DBI value (close to 0), the better the quality of the formed cluster. From the PC cluster evaluation, the best quality cluster created in the application of Minkowski-Chebysev distance for the two formed clusters with 0.06135 PC value. This conclusion also based on the criteria where the higher the PC value (close to 1), the better the quality of the formed cluster. In future research, it can be done by simulating at another distance with several data sets.

References

[1] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*. Wiley, 2012.
[2] R. Sharma and K. Verma, "Fuzzy shared nearest neighbor clustering," *Int. J. Fuzzy Syst.*, vol. 21, no. 8, pp. 2667–2678, 2019, doi: 10.1007/s40815-019-00699-7.
[3] 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.*, vol. 42, no. 10, pp. 1482–1484, Oct. 1997, doi: 10.1109/TAC.1997.633847.
[4] 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.*, vol. 137, no. 1–2, pp. 31–43, 2019, doi: 10.1007/s00704-018-2576-4.
[5] 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).*, vol. 484, pp. 350–366, 2019, doi: 10.1016/j.ins.2019.01.071.
[6] M. Ghane'i Ostad, H. Vahdat Nejad, and M. Abdolrazzagh Nezhad, "Detecting overlapping communities in LBSNs by fuzzy subtractive clustering," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1007/s13278-018-0502-5.
[7] E. S. Abdolkarimi and M. R. Mosavi, "Wavelet-adaptive neural subtractive clustering fuzzy inference system to enhance low-cost and high-speed INS/GPS navigation system," *GPS Solut.*, vol. 24, no. 2, pp. 1–17, 2020, doi: 10.1007/s10291-020-0951-y.
[8] S. M. M. Alam and M. H. Ali, "A new subtractive clustering based ANFIS system for residential load forecasting," *2020 IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. ISGT 2020*, 2020, doi: 10.1109/ISGT45199.2020.9087653.
[9] H. Salah, M. Nemissi, H. Seridi, and H. Akdag, "Subtractive Clustering and Particle Swarm Optimization Based Fuzzy Classifier," *Int. J. Fuzzy Syst. Appl.*, vol. 8, no. 3, pp. 108–122, 2019, doi: 10.4018/ijfsa.2019070105.
[10] L. Banteng, H. Yang, Q. Chen, and Z. Wang, "Research on the subtractive clustering algorithm for mobile ad hoc network based on the akaike information criterion," *Int. J. Distrib. Sens. Networks*, vol. 15, no. 9, 2019, doi: 10.1177/1550147719877612.
[11] S. K. Chandar, "Stock market prediction using subtractive clustering for a neuro fuzzy hybrid approach," *Cluster Comput.*, vol. 22, no. s6, pp. 13159–13166, 2019, doi: 10.1007/s10586-017-1321-6.
[12] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on Minkowski distance for many-objective optimization," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3968–3979, 2019, doi: 10.1109/TCYB.2018.2856208.
[13] R. Taheri, M. Ghahramani, R. Javidan, M. Shojafar, Z. Pooranian, and M. Conti, "Similarity-based Android malware detection using Hamming distance of static binary features," *Futur. Gener. Comput. Syst.*, vol. 105, pp. 230–247, 2020, doi: 10.1016/j.future.2019.11.034.
[14] O. Rodrigues, "Combining minkowski and cheyshev: new distance proposal and survey of distance metrics using k-nearest neighbours classifier," *Pattern Recognit. Lett.*, vol. 110, pp. 66–71, 2018, doi: 10.1016/j.patrec.2018.03.021.
[15] P. Noviyanti, *Fuzzy c-Means Combination of Minkowski and Chebyshev Based for Categorical Data Clustering*. Yogyakarta: Universitas Ahmad Dahlan, 2018.
[16] 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, doi: 10.1007/s40815-020-00997-5.
[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," vol. 16, no. 6, pp. 159–176, 2019.
[19] G. Gan, C. Ma, and J. Wu, "Data clustering: theory, algorithms, and applications," *Soc. Ind. Appl. Math.*, 2020.
[20] T. Brunello, D. Bianchi, and E. Enrico, *Introduction To Computational Neurobiology and Clustering*. Singapore: World Scientific Publishing, 2007.
[21] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.