HASIL CEK_C6 Prosiding Confast

by C6 Prosiding Confast Nursyiva

Indonesian community welfare levels clustering using the fuzzy subtractive clustering (FCM) method

To cite this article: P Laras Sakti Cahyaningrum and Nursyiva Irsalinda 2019 *J. Phys.: Conf. Ser.* **1373** 012036

View the article online for updates and enhancements.

Indonesian community welfare levels clustering using the fuzzy subtractive clustering (FCM) method

P Laras Sakti Cahyaningrum[1], Nursyiva Irsalinda[2]

[1,2]Department of Mathematics, University Ahmad Dahlan, Yogyakarta, Indonesia

*Email: [1] putrilarassakti16@gmail.com

Abstract. Clustering is a technique used to classify objects or cases into groups based on their similarity, called clusters or groups. Objects in each group tend to resemble each other and differ greatly (not the same) with objects from other clusters. Public welfare is a condition of fulfilling the material, spiritual and social needs of citizens in order to be able to live properly. Fuzzy Subtractive Clustering (FSC) method is a clustering algorithm that can form the number and centroid of clusters in accordance with data conditions. This study aims to determine the FSC results in grouping the level of welfare of the Indonesian people in 2017. The testing results of the cluster validity index show 2 values of Partition Entropy and Classification Entropy forming into 2 clusters that have the best value, indicating that the provincial group has a high welfare level and the provincial group has a low welfare level.

1. Introduction

Clustering a technique used to classify objects or cases into relatively homogeneous groups, called clusters or groups. Objects in each group tend to resemble each other and differ greatly (not the same) with objects from other clusters [1].

There are many methods that can be used in grouping, one of which is most often used is Fuzzy C-Means (FCM). Fuzzy C-Means method is a supervised clustering algorithm, because in FCM we need to know in advance the number of clusters to be formed [2]. The FCM method has a weakness, it requires the number of groups and the matrix of group membership predetermined [3]. Meanwhile, if the number of groups is not known beforehand, there are alternative grouping methods that can be used, namely the Subtractive clustering (SC) method. SC gets more consistent results compared to FCM [4].

Fuzzy subtractive clustering is an unsupervised clustering algorithm that can form the number and centroid of clusters that match the data conditions [5], [6], [7]. The basic concept of subtractive clustering is to determine the areas in a variable that have a high density of the surrounding points. Fuzzy subtractive clustering method produces very good estimation model accuracy in modelling non-linear systems [8], [9]. The community welfare is a condition of fulfilling the material, spiritual and social

needs of citizens to live properly and be able to develop themselves, so that they can carry out their social functions. The level of community welfare is caused by many factors including economic growth factors, education factors, health factors and so forth. Equitable distribution of people's welfare needs to be accelerated by taking into account employment growth and absorption, people's access to employment opportunities in parallel must be followed by efforts to improve their skills in order to compete in the free market [10],[11].

Therefore, researcher will use the Fuzzy subtractive clustering method to classify the level of welfare of Indonesian people. Thus the purpose of this study is to determine the results of the application of fuzzy subtractive clustering method to welfare level of Indonesian.

2. Problem characteristics and assumption

The data source in this study is secondary data. The data used in this study are the percentage of education level, average income level, percentage of good nutritional status, percentage of national road length and average floor area per capita. Data was obtained through the official website of the Central Statistics Agency (BPS), the Ministry of Public Works and Public Housing, and the Ministry of Health of the Republic of Indonesia

The steps of the FSC algorithm in this study are as follow:
1. Input the data to be clustered x_{ij} , with $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,m$, where n is amount of data and m is the amount of data attributes.
2. Set the parameter values, i.e r (radius each data attributes), q (squash factor), Accept ratio, Reject ratio, $XMin$ (minimum data), $XMax$ (maximum data).
3. Normalize each data using the S-curve membership function using Equation (1) as follows:

$$\mu(x) = \begin{cases} 0 & x \leq a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a < x < \frac{a+b}{2} \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b \\ 1, & x \geq b \end{cases} \quad (1)$$

with x are elements of the decision matrix (research data) and a and b are intervals of the growth S-curve membership function.

4. Determine the initial potential of each data point $(D_i; i = 1,2,\ldots,n)$. The steps are as follows:
   a. Calculate the distance of each data for T_j with formula:

$$Dist_{ij}(x_i) = \left(\frac{T_j - x_{ij_{norm}}}{r}\right) \quad (2)$$

   where $T_j = x_{ij_{norm}}; j = 1,2,\ldots,m$. To calculate the $Dist_{ij}(x_1)$, value T_j used is $x_{1j_{norm}}$ with value T_j not changing until $i = n$. And so on until $Dist_{ij}(x_n)$.
   b. Determine the initial potential of each data point using equation (3) for $m = 1$ and equation (4) for $m > 1$.

$$D_i = \sum_{i=1}^{n} e^{-4(Dist_{ij}(x_i)^2)} \quad (3)$$

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

5. Search the highest potential value:.

$$M = \max[D_i | i = 1,2,\ldots,n]; \quad \text{for first iteration.} \tag{5}$$
$$Z = \max[D_i | i = 1,2,\ldots,n]; \quad \text{for the next iteration.} \tag{6}$$

6. Calculate the value of prospective cluster centroid ratio by using Equation (7) as follows:

$$Rasio = \frac{Z}{M} \tag{7}$$

Particularly in the first iteration $Z = M$.

7. Checking the eligibility of prospective cluster centroids into cluster centroid, there are 3 conditions that can occur as follows:

    a. If Ratio > Accept ratio the prospective cluster centroid can be accepted as a new cluster centroid

    b. If Reject ratio < Ratio ≤ Accept ratio, then checking the credibility of the prospective cluster centroid can be accepted as a new cluster centroid, if its existence is far enough from the existing cluster centroid.

    The procedure in this situation is as follows:

    For example : $Md = -1$;

    For $k = 1$ to $k = p$; $p =$ the amount of $cluster$.

$$Sd_k = \sum_{j=1}^{m} \left(\frac{V_j - C_{kj}}{r}\right)^2 \tag{8}$$

If $(Md < 0)$ or $(Sd < Md)$, then $Md = Sd_l$, if $(Sd_l > Md)$, then Md not changing.
$$Mds = \sqrt{Md};$$

with $M\,ds$ is the closest distance from the prospective cluster data centroid to the cluster centroid. If $(Rasio + Mds) \geq 1$, prospective cluster centroid is accepted as a new cluster centroid. If $(Rasio + Mds) < 1$, prospective cluster centroids are not accepted and will not be reconsidered as new cluster centroid and the potential data set is zero. If Ratio ≤ Reject ratio, then there are no more data points to be considered to be candidates for the cluster centroid, the iteration is stopped.

8. If a prospective cluster centroid can be accepted as a new cluster centroid, a reduction in the potential of the data around the first cluster centroid is performed with Equation (9) as follows:

$$D_i^t = D_i^{t-1} - D_{C_{kj}} \tag{9}$$

To calculate value $D_{C_{lj}}$ using Equation (10).

$$D_{C_{kj}} = M \times e^{-4\left[\sum_{j=1}^{m}(S_{ij})^2\right]} \tag{10}$$

$$S_{ij} = \frac{C_{kj} - x_{ij_{norm}}}{r * q} \tag{11}$$

9. Returns the cluster centroid from the normalized form to the original form using the Equation (12).

$$C_{kj_{denorm}} = C_{kj} * \left(X_{max_j} - X_{min_j} \right) + X_{min_j} \tag{12}$$

10. Calculate the sigma cluster value of each data attribute by using the Equation (13).

$$\sigma_j = \frac{r * \left(X_{max_j} - X_{min_j} \right)}{\sqrt{8}} \tag{13}$$

11. Calculate the degree value of membership using the gauss function with Equation (14).

$$\mu_{ki} = e^{-\sum_{j=1}^{m} \frac{(x_{ij} - c_{kj})^2}{2\sigma_j^2}} \tag{14}$$

2.1 Validity Index

1. **Partition Coefficient** (PC)

$$PC = \frac{1}{N} \left(\sum_{i=1}^{N} \sum_{j=1}^{K} \mu_{ij}^2 \right) \tag{15}$$

with N is the number of research objects, K is the number of clusters, and μ_{ij} is the value of membership of the i-th object with the centroid of the j-th group.

2. **Partition Entropy** (PE)

$$PE = -\frac{1}{N} \left(\sum_{i=1}^{N} \sum_{j=1}^{K} \mu_{ij}{}^2 \log \mu_{ij} \right) \tag{16}$$

3. **Coefficient Entropy** (CE)

$$CE = -\frac{1}{N} \left(\sum_{i=1}^{N} \sum_{j=1}^{K} \mu_{ij} \log(\mu_{ij}) \right) \tag{17}$$

with N is the amount of research objects, K is the amount of clusters, and μ_{ij} is the value of membership of the i-th object with the centroid of the j group.

3. Result and Discussion

Inputting Data

Data is made by a matrix with size 34 ×5, The size of this matrix shows that the sample amounts are 34 provinces and there are 5 variables used in the analysis, namely the percentage of Education Level, Average Income Level, Percentage of Good Nutrition Status, Percentage of National Road Length and Average Floor Area Per Capita.

Determining Parameter Values

a. In testing the effect of the radius's value in this study there were 5 values of the radius used, namely 1.35, 1.2, 0.95, 0.92 and 0.85. Researchers took 5 values of this radius because based on experiments using laboratory of math software if taking a value of radius> 1.35 will produce a very small group whereas taking a value of radius <0.85 produces a very large group.

b. Squash factor used in this study was 1.25. Squash factor of 1.25 shows that the radius of the data points around the centroid of the cluster to be measured decreases in potential data is 1.25.

c. Accept ratio used in this study is 0.5, indicating that the lower boundary value of the data points that are prospective cluster centroids to be allowed to become cluster centroid is 0.5.

d. The rejection ratio used in this study is 0.15, indicating that the upper boundary value from the data points that become prospective cluster centroids is not allowed to be the cluster centroid is 0.15.

e. Minimum Data: V_{1min} =5.385; V_{2min} =2010062; V_{3min} =71.7; V_{4min} =0.54; V_{5min} =13.54. The minimum data on the first attribute is 5,385 indicates, showing the lowest level of education, the minimum data on the second attribute is 2010062, showing the lowest average income, the minimum data on the third attribute of 71.7 shows the lowest level of good nutrition, the minimum data on the fourth attribute of 0.54 shows the lowest national road length presentation and the minimum data on the fifth attribute of 13.54 shows the lowest average floor area.

f. Maximum data: V_{1max} =11.755; V_{2max}=4089123; V_{3max}=91.4; V_{4max}=8.3658; V_{5max} =34.13. The maximum data on the first attribute of 11,755 shows the highest level of education, the maximum data on the second attribute of 4089123 shows the highest level of income, the maximum data on the third attribute of 91.4 shows the highest level of good nutrition, the maximum data on the fourth attribute of 8.3658 shows the long presentation the highest national road and the maximum data on the fifth attribute of 34.13 shows the highest average floor area.

Normalizing the data
In normalizing data the researcher uses the growth S-curve membership function in the equation (1).

Finding Cluster Centroid Values
Analysis with equations r = 1.35 produces 2 clusters, r = 1.2 produces 3 clusters, r = 0.95 produces 4 clusters, r = 0.92 produces 5 clusters and r = 0.85 produces 6 clusters. The results of the cluster centroid values obtained are as follows:

$$C_{r=1.35} = \begin{bmatrix} 0.3230 & 0.1211 & 0.5251 & 0.0092 & 0.3554 \\ 0.9946 & 0.2338 & 0.1044 & 0.1581 & 0.3958 \end{bmatrix}$$

$$C_{r=1.2} = \begin{bmatrix} 0.2407 & 0.0501 & 0.5051 & 0.1270 & 0.3579 \\ 0.9946 & 0.2338 & 0.1044 & 0.1581 & 0.3958 \\ 0.5855 & 1.0000 & 0.8497 & 0.7486 & 0.3050 \end{bmatrix}$$

$$C_{r=0.95} = \begin{bmatrix} 0.2407 & 0.0501 & 0.5051 & 0.1270 & 0.3579 \\ 0.9946 & 0.2338 & 0.1044 & 0.1581 & 0.3958 \\ 0.0580 & 0.9413 & 0.3811 & 0.4072 & 0.2863 \\ 0.5754 & 0.5414 & 0.7687 & 1.0000 & 0.6213 \end{bmatrix}$$

$$C_{r=0.92} = \begin{bmatrix} 0.2407 & 0.0501 & 0.5051 & 0.1270 & 0.3579 \\ 0.9946 & 0.2338 & 0.1044 & 0.1581 & 0.3958 \\ 0.0580 & 0.9413 & 0.3811 & 0.4072 & 0.2863 \\ 0.5754 & 0.5414 & 0.7687 & 1.0000 & 0.6213 \\ 0.9983 & 0.0885 & 0.8384 & 0.1626 & 0.2775 \end{bmatrix}$$

$$C_{r=0.85} = \begin{bmatrix} 0.2407 & 0.0501 & 0.5051 & 0.1270 & 0.3579 \\ 0.7847 & 0.0671 & 0.0249 & 0.1486 & 0.3329 \\ 0.0580 & 0.9413 & 0.3811 & 0.4072 & 0.2863 \\ 0.9122 & 0.1982 & 0.5918 & 0.1413 & 0.3604 \\ 0.0004 & 0 & 0.6364 & 0.2351 & 0.9844 \\ 0.5754 & 0.5414 & 0.7687 & 1.0000 & 0.6213 \end{bmatrix}$$

The number of rows in the C matrix is the number of clusters formed. The first row shows the centroid of the first cluster, the second line shows the centroid of the second cluster, and so on. Whereas the first column shows the Education Level value, the second column shows the Income Level value, the third column shows the Good Nutrition value, the fourth column shows the National Road Length value and the fifth column shows the Floor Area value.

Returning the Cluster Centroid from the Normalized Form to the Original Form

In returning the cluster centroid from the normalized form to the original form we use equation (12).

Table 1. The Cluster Centroid Value by $r = 1.35$

Cluster	Education	Income	Good Nutrition	PJN	Floor Area	Cluster Average	Cluster Rating
1	7.4426	2261753.95	82.04	0.61	20.86	452372.981	2
2	9.0504	3135754.06	86.84	8.37	26.33	627176.9306	1

Table 2. Cluster Centroid Value by $r = 1.2$

Cluster	Education	Income	Good Nutrition	PJN	Floor Area	Cluster Average	Cluster Rating
1	7.595	2339093	81.6	2.512	22.25	422863.505	3
2	11.425	2720928	76.2	2.740	22.7	499257.0158	2
3	17.71	4089123	86	5.591	21.58	817849.3543	1

Table 3. Cluster Centroid Value by $r = 0.95$

Cluster	Education	Income	Good Nutrition	PJN	Floor Area	Cluster Average	Cluster Rating
1	7.595	2339093	81.6	2.512	22.25	422863.505	4
2	11.425	2720928	76.2	2.740	22.7	499257.0158	3
3	6.47	3732904	80.3	4.071	21.33	793432.9007	1
4	8.82	3093605	84.7	8.366	25.17	627176.9306	2

Table 4. Cluster Centroid Value by $r = 0.92$

Cluster	Education	Income	Good Nutrition	PJN	Floor Area	Cluster Average	Cluster Rating
1	7.595	2339093	81.6	2.512	22.25	422863.505	5
2	11.425	2720928	76.2	2.740	22.7	499257.0158	3
3	6.47	3732904	80.3	4.071	21.33	793432.9007	1
4	8.82	3093605	84.7	8.366	25.17	627176.9306	2
5	11.57	2447512	85.8	2.771	21.21	438853.7075	4

Table 5. Cluster Centroid Value by r= 0.85

Cluster	Education	Income	Good Nutrition	PJN	Floor Area	Cluster Average	Cluster Rating
1	7.595	2339093	81.6	2.512	22.25	422863.505	5
2	9.665	2390924	73.9	2.673	21.94	429941.2914	4
3	6.47	3732904	80.3	4.071	21.33	793432.9007	1
4	10.42	2664486	82.5	2.620	22.28	484432.7932	3
5	5.48	2010062	83	3.223	32.31	402037.5625	6
6	8.82	3093605	84.7	8.366	25.17	627176.9306	2

Calculate the sigma cluster

Calculate the sigma cluster using Equation 13 and the results obtained as in Table 6 below:

Table 6. Sigma Cluster Values

	Sigma Cluster				
r	Education	Income	Good Nutrition	PJN	Floor Area
1.35	3.04	992329.74	9.40	3.74	9.83
1.2	2.70	882070.88	8.36	3.32	8.74
0.95	2.13	698306.11	6.62	2.63	6.92
0.92	2.07	676254.34	6.41	2.56	6.70
0.85	1.91	624800.21	5.92	2.35	6.19

Calculate the Membership Degree

In calculating the degree of membership using the Gauss function in equation (14), the degree of membership of each data can be found in each cluster.

Determine the Location of Clusters

The location of clusters in each province is determined based on the value of the membership's degree. The highest degree of membership indicates a high tendency for a province to join the cluster.

Calculating Cluster Validity

The Partition Coefficient (PC) is calculated using Equation (15) and produces the following values:

Table 7. Index of PC Value Validity Results of the FSC Grouping

Cluster Amount	PC Value
2	0.3252
3	0.3963
4	0.2982
5	0.3073
6	0.3316

From Table 7 the highest PC value is 0.3963 so it shows the formation of 3 clusters is better than the others. Partition Entropy (PE) is calculated using Equation (16) and produces values in Table 7:

Table 8. Index of PE Value Validity Results of the FSC Grouping

Cluster Amount	PE Value
2	0.7502
3	0.9093
4	0.8444
5	0.9606
6	1.0670

From Table 8 the smallest PE value is 0.7502, so it shows the formation of 2 clusters is better than the others. Classification Entropy (CE) is calculated using equation (17) and produces the following

values: **Table 9. Index of CE Value Validity Results of the FSC Grouping**

Cluster Value	CE Rating
2	0 .5200
3	0.6303
4	0.5853
5	0.6659
6	0.7396

from Table 9 the smallest CE value obtained is 0.5200, so it shows the formation of 2 clusters is better than the others.

Based on Table 6 it can be seen that table 8 has the formation of 2 clusters for better grouping at the level of Indonesian welfare if compared to 3,4,5 or 6 clusters. From the validity index it can be seen that 2 clusters have 2 levels of validity that are better than other clusters with the best PE value of 0.7502 and the CE validity index obtained the best value of 0.5200. The formation of 3 clusters has only one good validity index namely the PC value of 0.3963.

4. Conclusion

It can be concluded that Community Welfare in Indonesia can be grouped into 2 clusters. The first cluster for the second group level will consist of the provinces of Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Kepulauan Bangka Belitung, West Java, Central Java, East Java, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua and Papua. Whereas the second Cluster for the 1st group level will consist of Kepulauan Riau province, DKI Jakarta, DI Yogyakarta, Banten, Bali and North Sulawesi.

References

[1]   Supranto. J 2004 *Analisis Multivariat Arti dan Inferensi* (Jakarta: Rieka)
[2]   Kusumadewi. S and Purnomo. H 2010 *Aplikasi logika Fuzzy untuk Pendukung Keputusan* (Yogyakarta: Graha Ilmu)

[3]   Le. T and Altman. T 2011 *Proc. International Conference on Information and knowledge Enginering Las Vegas USA* **1** 144-150

[4]   Bataineh. K M, Naji. M and Saqer. M 2011 *Jordan Journal of Mechanical and Industrial Engineering* **5** 335-343

[5]   Bezdek. J 1981 *Pattern Recognition with Fuzzy Objective Function Algorithm* (New York: Plenum)

[6]   Chamzini. A Y, Razani. M and Yakhchali. S H 2013 *Automation in Construction* **35** 113

[7]   Chiu. S L 1994 *Fuzzy Model Identification Based on Cluster Estimation Rockwell Science Center* (California)

[8]   Haqiqi. B N and Kurniawan. R 2015 *Statistika* **8**

[9]   Hossen. J, Rahman. A, Sayeed. S, Samsuddin. K, and Rokhani. F 2011 *Australian Journal of Basic and Applied Sciences* **5** 674-681

[10]   Kusumadewi. S and Hartiati. S 2006 *Neuro Fuzzy Integrasi Sistem Fuzzy dan Jaringan Syaraf* (Yogyakarta: Graha Ilmu)

[11]   Kusumadewi. S and Purnomo. H 2004 *Aplikasi Logika Fuzzy Untuk Pendukung Keputusan* (Yogyakarta: Graha Ilmu)

HASIL CEK_C6 Prosiding Confast

12% SIMILARITY INDEX

9% INTERNET SOURCES

7% PUBLICATIONS

5% STUDENT PAPERS

PRIMARY SOURCES

1	eprints.uad.ac.id Internet Source	2%
2	pokameame.com Internet Source	2%
3	pt.scribd.com Internet Source	1%
4	"Intelligent Systems Technologies and Applications", Springer Science and Business Media LLC, 2018 Publication	1%
5	Submitted to Universitas International Batam Student Paper	1%
6	insightsociety.org Internet Source	1%
7	Abdul Aziz, Muhammad Maarif Farid, Esti Suryani. "Floyd warshall algorithm with FIS Sugeno for search evacuation route optimization", 2017 International Seminar on Application for Technology of Information and Communication (iSemantic), 2017	1%

16   Anif Hanifa Setyaningrum, Praditya Megananda Swarinata. "Weather prediction application based on ANFIS (Adaptive neural fuzzy inference system) method in West Jakarta region", 2014 International Conference on Cyber and IT Service Management (CITSM), 2014
Publication

<1%

17   Khairul Saleh. "ANALISIS RULE INFERENSI FUZZY MAMDANI DALAM MENENTUKAN IPK AKHIR", JURNAL TEKNOLOGI INFORMASI, 2017
Publication

<1%

18   Sri Endah Moelya Artha, Hasbi Yasin, Budi Warsito, Rukun Santoso, Suparti. "Application of Wavelet Neuro-Fuzzy System (WNFS) method for stock forecasting", Journal of Physics: Conference Series, 2018
Publication

<1%

19   M. Sato-Ilic. "On fuzzy clustering based regression models", IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS '04., 2004
Publication

<1%

20   www.msc-les.org
Internet Source

<1%

21   Construction Innovation: Information, Process, Management, Volume 12, Issue 3

<1%

(2012-06-30)

Publication

Exclude quotes	On		Exclude matches	Off
Exclude bibliography	On			