ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse135862019.pdf
https://doi.org/10.30534/ijatcse/2019/135862019

The Implementation of a Hybrid Fuzzy Clustering on the Public Health Facility Data

Samingun Handoyo[1], Agus Widodo[2], Waego Hadi Nugroho[3], Imam N. Purwanto[4]

[1]Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia, samistat@ub.ac.id
[2]Department of Mathematics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia, prof.agus_widodo@yahoo.com
[3]Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia, whn@ub.ac.id
[4]Department of Mathematics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia, purwanto_imam@yahoo.com

ABSTRACT

The government has responsible to make a better life of their society in various aspects of life including a standard level of society public. The research aims to implement combining two fuzzy clustering methods that are Fuzzy Subtractive Clustering and Fuzzy C-Means for clustering public health facility data. The hybrid method can result in the cluster more compactness and more homogenous in membership degree accessed bases on Partition coefficient (PC), Classification Entropy (CE), and Modified of PC and CE (MPC) of cluster validity index.

Key words : fuzzy clustering, hybrid method, public health, validity index.

1. INTRODUCTION

Health is important for human survival. From the age of toddlers to adults, people need to regularly check their health. However, in reality, many people, generally in developing countries, especially in Batu City and Malang regency still do not realize the importance of health checks, so that health problems always occur in the community, especially for people who live in villages that are far from the Health Office. In addition to the lack of public awareness of the importance of health checks, the lack of health facilities both infrastructure and medical personnel are few factors that cause health problems in the region above. Therefore, it is necessary to highlight the diseases that often arise in villages, the number of medical personnel and health infrastructure facilities, so that the government is able to take appropriate action for health problems that are basically in every village is not the same.

In order for the government can improve public health in general, the status of villages must be mapped accurately. If the village profile is known absolutely, it will support the government in making policy related to the handling of health problems. The information of the village profile related to public health needs to identify, and the method which is used to group some entities having some features is called the clustering method [1]. The term clustering is popular in both multivariate statistic and machine learning called unsupervised learning [2]. The main challenge in the implementation of the clustering method is how to find the optimal clusters from the data set.

The clustering method which yields the information of the membership degree of the entity or object on a cluster is more preferable because it seems more fairly and more making sense for the human viewpoint. The clustering method that has the characteristics above related to fuzzy logic and the method often is called fuzzy clustering. In machine learning, there are 2 popular fuzzy clustering methods that ate the Fuzzy Subtractive Clustering (FSC) and the Fuzzy C-mean (FCM) [3-4]. Both methods have advantages and weaknesses with each other, and the optimal compromise should be found.

The fuzzy subtractive clustering appeal is very high for researchers in various fields of interest which ranges in social science to engineering. The most its advantage is that the method can generate automatic clustering without requiring to set the cluster number at the beginning [5-8]. In fact, the input of the radius magnitude which has an important role in Fsc is a big problem determined by trial and error. Marji, et al. [9] had observed the effect of variability on variables related to the magnitude radius in yielding the optimal cluster.

The clustering method that clusters entities become c number of the cluster as the input and magnitude of cluster degree as a component for an entity being a member of cluster is called the Fuzzy C-Means (FCM) [11]. The FCM method is very popular and easy to implement in many applications including the use of the FCM for generating the fuzzy rule bases which are the core of the fuzzy inference system in the resulting output

[12-15]. Determining how many cluster numbers is the optimal number needs any considerations and still a big problem in the application. An effort must be done to get the optimal cluster numbers which it carries out by combining the advantages of both FSC and FCM.

Based on the explanation in some previous paragraphs, the research is intended to implement the combining two fuzzy clustering methods that are FSC and FCM for the clustering of public health facilities in Batu City and Malang Regency of Indonesia and the accessing of the performance of the methods are used validity indexes [16].

2. LITERATURES REVIEW

Cluster analysis is part of a multivariate technique that aims to identify similar entities or objects base on their characteristics. The members of a cluster have several characteristics or features that are very similar with respect to selection criteria that were decided previously The method is a useful tool for data analysis in various situations including a researcher who has collected data through a questionnaire that faced a large number of insignificant observations unless classified into manageable groups which mean Cluster analysis can be used to carry out dimension reduction organizing information being specific small subgroups. In this way, the researcher has a shorter observation and easier understanding of population description with less information loss [1].

Besides that Cluster analysis is also useful if the researcher wants to develop hypotheses about the nature of the data or to test hypotheses that have been stated previously. For example, researchers might believe that attitudes toward the light and ordinary beer consumption can be used to separate beer consumption into logical segments. In this case, the cluster analysis can be used to group beer consumers by their attitudes. The resulting clusters can be profiled for demographic similarities and their differences [5].

2.1. Fuzzy Subtractive Clustering

The Fuzzy Subtractive Clustering (FSC) which clustering based on the size of the data density in a dimension space is one of the popular cluster methods in unsupervised learning [9]. The determining regions in a variable space that has a high density of the surrounding points is the FSC basic concept. The cluster center is chosen by a point having the highest number of neighbors. The selected point will be reduced in its density. Choosing another point that has the most neighbors to be the center of another cluster is the process will be repeated in the same way and a stopping condition is when all points have been checked [6].

Lisangan, et al stated if there are N observations: $X_1, X_2, \ldots, X_N$ and assuming that the data are in a normal distribution, then the density of the point X_k can be calculated as follows [7]:

$$D_k = \sum_{j=1}^{N} exp\left(-\frac{\|X_k - X_j\|}{(r/2)^2}\right) \quad (1)$$

Where Dk is density of point k, after the density on each point is improved, then the second cluster center (X_{c2}) will be searched. After X_{c2} was obtained, the size of the density of each data point will be fixed again, and so on.

The FSC algorithm considers two kinds of ratio that are Accept and Reject ratio. The ratios have values between 0 and 1. The lower boundary in which a data point is a center candidate of the cluster and it is allowed to be the cluster center is called the Accept ratio. While the upper boundary in which a data point is a center candidate of the cluster and it is not allowed to be the cluster center is called the Reject ratio. There are 3 conditions that will occur in an iteration process as follows [9]:

1) The data point is accepted as the center of the new cluster If the ratio > accept ratio.
2) The data point will only be accepted as the new cluster center if the reject ratio < ratio ≤ accept ratio, if and only if the data point is located at a considerable distance from the other cluster center (the sum of the ratio and the closest distance of the data point to the other existing cluster centers ≥1). On another hand, the data point will not be accepted as the cluster center if the results of the sum between the ratio and the longest distance of the data point with the other existing cluster centers < 1. It will no longer be considered to be the center of the new cluster.
3) There are no more data points that will be considered to be candidate of a cluster center if the ratio ≤ reject ratio, and the iteration is stopped.

2.2. The Fuzzy C-Means (FCM) clustering

Fuzzy C-Means (FCM) that was firstly introduced by Bezdek, et al [11] is a clustering technique in which the existence of each data point in a cluster is determined by its membership degree.
The basic concept of the FCM is to determine the cluster center that will mark the average location for each cluster. Each data point was assigned randomly a membership degree on each cluster through improving the cluster center and was updated its membership degree repeatedly. It will be seen that the cluster center will move to the right location. The iteration is based on the objective function minimization that describes the distance between the data point and the cluster center weighted by its membership degree [10].

In order to resulting the optimal cluster number, the FCM algorithm is presented as follows [11]:

a. Setting some parameter including cluster number(c), fuzzifier (m), iteration number (nIter), the expectation of objective function shifted(ε), the initial value of objective function (P0 = 0).
b. Initializing membership degree (Uik) of partition matrix U is generated by random numbers, where i is a

record number and k is a cluster number.

c. calculating the cluster center (Pi) of the i-th cluster is given by the formula as follows:

$$p_i = \frac{\sum_{k=1}^{N} (u_{ik})^m X_k}{\sum_{k=1}^{N} (u_{ik})^m} \quad (2)$$

Where uik is the membership degree of the k-th object on the center of the i-th cluster, Xk is the k-th data object, N is the number of research objects, and m is the fuzzifier

d. Meanwhile, the calculating of the objective function in the t-iteration uses the formula:

$$(P, U, X, c, m) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m d_{ik}^2 (x_k, p_i) \quad (3)$$

Where the desired cluster number (c), the number of the entities (N). The uik is the membership degree of the k-th entity on the i-th cluster which is an element of the U matrix partition, m is the fuzzifier , and the $d_{ik}^2(x_k, p_i)$ is the distance between the k-th record and the center of the i-th cluster.

e. The calculating of amount change of the membership degree in partition matrix of U is given by the formula as follows:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2} \right)^{\frac{1}{m-1}}} \quad (3)$$

2.3. The Subtractive Fuzzy C-Mean (SFCM)

The Subtractive Fuzzy C-Mean (SFCM) is a clustering method that hybrid the FSC and FCM methods. This method was used by Li, et al. [17] which their research concluded that SFCM generally provides a better solution than FCM and provides a higher rate of speed in terms of objective functions convergence. Other advantages of the SFCM method are that it can increases speed, reduces the iteration numbers, produce a more stable and more accurate data partition because the FSCM uses the Partition Coefficient (PC) validity index [16]

The SFCM method considered each data point as a cluster center candidate where its potential value is calculated by the following equation: [17]

$$P_i = \sum_{j=1}^{n} e^{-\frac{4}{r_\alpha^2} \|x_i - x_j\|^2} \quad (4)$$

Pi is the potential value for the i-th data point, the term of ‖. ‖ denotes the second norm. The positive constant rα is the radius defined the distance between the cluster center and neighborhood points around a cluster center. The potential value of each data point can be updated by using the formula as follows

$$P_i = P_i - P_1^* e^{-\frac{4}{r_\beta^2} \|x_i - x_1^*\|^2} \quad (5)$$

The formula used to calculate the initial membership degree matrix in FCM based on the cluster center obtained through the FSC method is as follows:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2} \right)^{\frac{1}{m-1}}} \quad (6)$$

If there is a distance d_{ik}^2 which is zero then the degree of membership of the kth data point in the i-th group (Uik) will be 1 and the degree of membership of the k-th data point in the other group will be 0.

2.4. The Cluster validity index

This study uses several types of validation indexes which include the Partition coefficient (PC), Classification Entropy(CE, and Modified Partition Coefficient (MPC). The partition coefficient is used to measure the amount of overlap between groups. Classification of entropy measures the fuzziness of a group partition. Modified Partition Coefficient aims to overcome the shortcomings of PCs and CE [16].

Table 1. The formula of cluster validity index

Validity name	Formula
Partition coefficient	$PC = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^2$
Classification Entropy	$CE = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^2 \ln U_{ij}$
Modified of PC and CE	$MPC = 1 - \frac{c}{c-1}(1 - PC)$

Where n is sample number, c is cluster number, u_{ij} is membership degree of the i-th data point on the j-th cluster number, m is a fuzzyfier, $\|x_k - v_i\|$ is euclidean distance between a data point and a cluster center, $\|x_k - v_i\|$ is the a euclidean distance between 2 cluster centers.

3. DATA AND RESEARCH METHOD

3.1. Research Variables

The data used in the study are secondary data obtained from the Central Statistics Bureau of Batu City and Malang Regency in 2016. The variables used in the study are as follows:

V1: The number of doctors
V2: Number of midwives
V3: Number of nurses and other health workers
V4: Number of community health centers
V5: Many public hospitals
V6: Number of village health centers
V7: Number of cases of dengue fever sufferers
V8: Number of cases of diarrhea sufferers

3.2. Stages in Data Analysis

The analysis method of the research is divided into 4 steps as follows:

1) Conduct a clustering using the FCM and SFCM methods on the health data of Batu City and Malang Regency in 2016
2) Calculate three indexes of the validity of the results of the FCM and SFCM method
3) Determine the optimal cluster number of the FCM and SFCM based on three validities
4) Compare and discuss the results of the FCM and SFCM clustering for the optimal cluster number.

4. RESULTS AND DISCUSSION

The result of the summary statistic on the data of 8 villages features on 44 villages in Batu city and Malang regency is given in Table 2 as follows:

Table 2. Summary Statistic of the characteristic 44 villages

Features	Min.	Max.	Mean	Spead
V1	0	16	1.98	2.22
V2	1	5	1.91	0.87
V3	0	20	3.57	3.99
V4	0	1	0.39	0.38
V5	0	3	0.25	0.47
V6	0	22	3.77	3.77
V7	0	26	3.68	4.86
V8	51	242	142.07	40.27

Table 2 shows that the average value of the doctor number was 1.98, with the highest number of doctors in the Ngaglik village of 16 people and there were some villages which do not a doctor including Tlekung, Torongrejo, Bulukerto, Sumberbrantas, Sumbergondo, Oro- oro Ombo, and Sumberejo. The standard deviation of the number of doctors is 2.22. The average value of the number of midwives is 1.91, the most midwives are in Temas and Ngaglik villages of 5 midwives and the lowest was in the villages of Tlekung, Mojorejo, Bulukerto, Bumiaji, Giripurno, Pandanrejo, Sumberbrantas, Sumbergondo, Sidomulyo and Songgokerto of 1 midwife. The value of the standard deviation of the number of midwives is 0.87. The average value of the number of paramedics and other health workers is 3.57, with the largest number in the Pesanggrahan and Ngaglik villages of 20 paramedics and other health workers while Pendem, Tlekung, Bulukerto, Bumiaji, Pandanrejo, Sumberbrantas, Tulungrejo, and Songgokerto Villages do not have spells and other health workers. The amount of deviation from the number of spells and other health workers is 3.99.

The largest number of hospitals is in Panesari Village with 3 hospitals, while in Sisir Village, Pesanggrahan, and Tlekung there is one hospital. Large deviations of the hospital by 0.47. Beji Village, Junrejo, Bumiaji, Giripurno, Gunungsari, Pandanrejo, Sumberbrantas, Ngaglik, Pesanggrahan, Sisir, and Temas have 1 community health centers or auxiliary community health centers, while in other villages there are no community health centers or auxiliary community health centers. Deviations in the number of health centers by 0.38. Most polindes are in Sisir Village with 22 polindes, while Beji and Junrejo villages do not have polindes. The deviation value for Polindes is 3.77.

The highest number of cases of Dengue Hemorrhagic Fever (DHF) sufferers is 26 cases of DHF in Mulyoagung Village. Mostly there are 1 to 5 sufferers in each village. The magnitude of the deviation in the number of cases of dengue patients in Batu and surrounding areas is 4.86 with an average of 3.68. The least number of diarrhea cases in Batu City and its surroundings were 51 cases in the Oro-Oro Ombo Village and the most were 242 cases in Bumiaji Village. Most of them have 142 cases of diarrhea in each village. The magnitude of the deviation in the number of cases of diarrhea sufferers in Batu City and surrounding areas with an average of 40.27.

4.1. The result of the FCM method

The clustering uses the FCM method by setting the number of clusters to 3, 4, 5, and 6 obtained the number of members in each cluster is presented in table 3 as follows:

Table 3. The member of each cluster resulted by FCM method

Amount member of each cluster for 3 cluster					
Cls.1	Cls.2	Cls.3			
7	8	29			
Amount member of each cluster for 4 cluster					
Cls.1	Cls.2	Cls.3	Cls.4		
6	8	10	20		
Amount member of each cluster for 5 cluster					
Cls.1	Cls.2	Cls.3	Cls.4	Cls.5	
5	6	10	7	16	
Amount member of each cluster for 6 cluster					
Cls.1	Cls.2	Cls.3	Cls.4	Cls.5	Cls.6
5	5	6	8	8	12

Where cls.1 is cluster1, cls.2 is cluster2, and so on. The table 3 describes the distribution of amount member of each cluster where it can be stated that the number of cluster increases then the distribution or spread of cluster member more relative homogen.

4.2. The SFCM result based on the FCM

The clustering of the SFCM method is tested on several radius values to get the same group such as the clustering of the FCM method. The following is the membership degree of 3 clusters where the radius value is 0.9 calculated by using R Software

Table 4. the degree of membership of each data point in the clustering into 3 groups with a radius of 0.9

Data point	Cluster1	Cluster2	Cluster3
1	0.00036	0.05728	0.00003
2	0.31326	0.00246	0.03969
3	0.0004	0.05867	0.00004
4	0.39311	0.00184	0.00895
5	0.04858	0.00036	0.25962
.	.	.	.
.	.	.	.
.	.	.	.
40	0.41196	0.0014	0.04661
41	0.00115	0.41905	0.00001
42	0.45129	0.00072	0.07629
43	0.86839	0.00488	0.0173
44	0.95871	0.00287	0.0217

Based on the table 4, the point will be a member of a cluster if its membership degree is the largest in the cluster. For example, the data point 40-th has the largest membership degree in thecluster 1 that means the data point 40-th is member of cluster 1.

4.3. . The comparation of validity index

The values of PC, MPC, and CE validity index are calculated based on the membership degree of the data points in a cluster. The validity index values in the FCM method can be seen in Table 5 as follows:

Table 5. The validity index results of the FCM clustering

Cluster number	PC	MPC	CE
3	0.74749	0.62123	0.15055
4	0.71515	0.6202	0.15903
5	0.69229	0.61537	0.16622
6	0.65674	0.58809	0.1643

Based on table 5, it can be known that the cluster number of 3 has the largest value of the PCs index that it becomes the best cluster number. The largest MPC index value of 0.62123 is on the cluster number of 3, while the best value of the CE index is the smallest value is in the cluster number of 3 with an index value of 0.15055. The results of the three indices can be concluded that the FCM method best cluster number is 3 clusters. In the SFCM method, the calculating validity index value does not only consider the cluster numbers but also involves several radius values. The validity index values of the SFCM method are given in Table 6 as follows:

Table 6. The validity index values of the SFCM clustering

R	C	PC	MPC	CE
0,9	3	0.25631	0.11554	0.09075
0,85	4	0.26856	0.02474	0.08784
0,8	4	0.26258	0.01677	0.08862
0,78	5	0.29886	0.12358	0.09597
0,75	6	0.31109	0.13886	0.09434
0,7	6	0.33407	0.16759	0.09401

Where c is a cluster number and R is radius values. Based on table 6 it can be known that the cluster number is equal to 4 and the radius value of 0.85 has the smallest CE index of 0.08784, while the MPC and PC I ndex have a better value for the cluster number of 6 and the radius of 0.7 where the MPC and PC index values of 0.33407 and 0.6759 respectively. In clustering with many clusters 3 and 6, the largest index validity values of the PC and MPC are in the FCM method, so it can be concluded that the FCM method yields better clustering results than the SFCM method based on the index validity values of the PC and MPC. Another hand, the index validity value of CE showed the SFCM method gives better results than the FCM method because it has a smaller index value in both clusters. In the clustering with many clusters is 4 and 5, the PC and CE index values such as described in table 6 that the SFCM method has a relatively stable index validity value in each cluster when compared to the validity index value in the FCM..

5. CONCLUSION

The application of the SFCM method to cluster the public health facilities in Batu City and Malang Regency produces information that overall the classification of the SFCM method provides better results in the stability of the validity index value than the FCM method. However, the FCM method has a better validity index value than the FCM method because it excels in two validity indices, namely PC and MPC in all groups

REFERENCES

1. J..F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, , & R.L. Tatham. Multivariate data analysis . Upper Saddle. 2000.
2. M. Swamynathan. Mastering machine learning with python in six steps: A practical implementation guide to

predictive data analytics using python. Apress; 2019. https://doi.org/10.1007/978-1-4842-4947-5
3. S. Madhavan. Mastering Python for Data Science. Packt Publishing Ltd; 2015.
4. B. Balasko, J. Abonyi, B. Feil. Fuzzy clustering and data analysis toolbox for use with matlab. Veszprem, Hungary. 2005.
5. S.L. Chiu. Fuzzy model identification based on cluster estimation. Journal of Intelligent & fuzzy systems. 1994 Jan 1;2(3):267-278. https://doi.org/10.3233/IFS-1994-2306
6. S. Tafazoli, L. Mathieu, and X. Sun, "Hysteresis modeling using fuzzy subtractive clustering", International journal of computational cognition, 4(3), 2006, pp 15-27
7. L.A. Lisangan, A. Musdholifah, and S. Hartati, "Two Level Clustering for Quality Improvement using Fuzzy Subtractive Clustering and Self-Organizing Map", Indonesian Journal of Electrical Engineering and Computer Science, 15(2), 2015, pp. 373-380.
8. S.H. Rouhani, A. Sheikholeslami, H. Hosseini, H. Kazemi, "Application of fuzzy subtractive clustering for optimal transient performance of automatic generation control", IETE Journal of Research, 59(6), 2013, pp. 753-760. https://doi.org/10.4103/0377-2063.126967
9. Marji, S. Handoyo, I.N. Purwanto, M.Y. Anizar. The Effect Of Attribute Diversity In The Covariance Matrix On The Magnitude of The Radius Parameter In Fuzzy Subtractive Clustering. Journal Of Theoretical & Applied Information Technology. 2018. 96(12): 3717-3728.
10. R.L. Cannon, V.D. Jitendra, and J.C. Bezdek. "Efficient implementation of the fuzzy c-means clustering algorithms." *IEEE transactions on pattern analysis and machine intelligence* 2 (1986): 248-255 https://doi.org/10.1109/TPAMI.1986.4767778
11. J.C. Bezdek, R. Ehrlich, and W. Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences* 10.2-3 (1984): 191-203. https://doi.org/10.1016/0098-3004(84)90020-7
12. S. Handoyo, A. Efendi. Generating of Fuzzy Rule Bases with Gaussian Parameters Optimized via Fuzzy C-Mean and Ordinary Least Square. *International Journal of Recent Technology and Engineering.* 2019. 8(4): 5787-5794. https://doi.org/10.35940/ijrte.D8561.118419
13. S. .Handoyo, Marji, I.N. Purwanto, F. Jie The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia. International J. of Opers. and Quant. Management (IJOQM). 2018;24(4):277-92.
14. S. Handoyo, H. Kusdarwati. Implementation of Fuzzy Inference System for Classification of Dengue Fever on the villages in Malang. InIOP Conference Series: Materials Science and Engineering 2019 Jun (Vol. 546, No. 5, p. 052026). IOP Publishing. https://doi.org/10.1088/1757-899X/546/5/052026
15. A. Efendi, S. Handoyo, A.P.S. Prasojo, and Marji. "The Implementation Of The Optimal Rule Bases Generated By Hybrid Fuzzy C-Mean And Particle Swarm Optimization", " *Journal of Theoretical & Applied Information Technology*". 2019. 97(16): 4453-4453.
16. W. Wang, Y. Zhang. On fuzzy cluster validity indices. Fuzzy sets and systems. 2007 Oct 1;158(19):2095-2117. https://doi.org/10.1016/j.fss.2007.03.004
17. J. Li, X.B. Gao, L.C. Jiao. A new feature weighted fuzzy clustering algorithm. Acta electronica sinica. 2006 pp. :89.-102.