Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 6 (2015), pp. 4315-4323

http://www.ripublication.com

K-Medoid and Fuzzy C-Means for Cluster of District/Cities in Indonesia (Case: 497 District/Cities in Indonesia 2012)

Mia Syafrina[1], I Made Sumertajaya[2], Indahwati[3]

Department of Statistics Bogor Agricultural University
Jalan Pajajaran, Kampus IPB Baranangsiang, Bogor 1615, Indonesia
mia_syafrina@yahoo.com, 2.imsjaya@yahoo.com, 3.indahwati_43@yahoo.co.id

Abstract

Cluster analysis is one of multivariate analysis techniques for classifying object into groups based on the characteristics observation. It's also to obtain the highest similarity between objects in the same cluster with a certain criteria. Fuzzy C-Means clustering allows an element placed in more than one cluster. This method could give the best result and also increasing homogeneity in each cluster. The algorithm that used in K-Medoid clustering based on the research of *k* representative objects among the objects of the data set as centroid. So, could give the best result and robust to outlier. Both of the methods applied for grouping district/cities in Indonesia by using the variable forming the Human Development Index (HDI). HDI forming variables consist of the life expectancy, literacy rate, mean year's school, and purchasing power parity. In order to evaluate the best method the average distance between object and center in the cluster and variance within and between clusters was applied. The methods have similar result for each member in cluster and appropriate for grouping district/cities in Indonesia by using the variable forming the HDI.

Keywords: Cluster Analysis, Fuzzy C-Means, K-Medoid, Variance

1 Introduction

In general, cluster analysis divided into two methods: hierarchical clustering method and nonhierarchical clustering method [5]. A hierarchical clustering is the method which has not known yet how many clusters which is formed. Meanwhile a nonhierarchical clustering is the method that the number of the cluster had been determined. It is exactly put each object in one cluster which is known as the basic of

conventional clustering method but also possible that the object placed in other clusters. So that it needs to be done clustering by using fuzzy clustering.
Fuzzy clustering considers a level of membership of fuzzier. So, it allows clustering of the unclear distribution data. They are some methods from fuzzy clustering, one of them Fuzzy C-Means clustering (FCM). Bezdek (1981) introduced Fuzzy C-Means clustering method in 1981. FCM is an improved form of K-Means. FCM allows an element placed in more than one cluster [10, 11]. FCM produces best result and increasing homogeneity in each cluster. In this research will be used K-Medoid method as a comparison of the FCM. K-Medoid selected representative objects on the cluster as central.
Fuzzy C-Means algorithm is more robust to outliers than K-Means algorithm. Both of the methods applied to grouping district/cities in Indonesia with highest similarity between objects of same cluster and lowest similarity between objects in the different clusters. Grouping district/cities in Indonesia by using the variable forming the HDI, it is necessary to be done as a planning and evaluation of the governments programs, particularly relating to the development of the quality of human life.

2 Materials and Methods

2. 1 Fuzzy C-Means Clustering Method

Bezdek introduced Fuzzy C-Means (FCM) clustering method in 1981, extend from Hard C-Means clustering method. FCM is an unsupervised clustering algorithm. In general Fuzzy C-Means clustering method is minimize the objective function. With the main parameters is membership in fuzzy (membership function) called fuzzier [6]. FCM is an improved form of C-Means algorithm which allows the degree of membership. It does mean that an object can belong to more than one cluster in some degree. Generally the points or objects which are on the edge of cluster might have less degree of belonging while the objects in the center might have higher belongingness [4, 8, 11].
The FCM algorithm is used for analysis based on distance between various data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster. The degree of membership of each data item to the cluster is calculated which decides the cluster to which that data item is supposed to belong [9].
For each item, we have a coefficient that specifies the membership degree of being in the cluster as follows [2, 7]

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

Where,

$(0 \le u_{ik} < 1)$, and$\sum_{i=1}^{c} = 1$.

The distance $d_{ik}^2(\mathbf{x}_k, \mathbf{v}_i)$ is represented as:

$$d_{ik}^2(\mathbf{x}_k, \mathbf{v}_i) = \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 = (\mathbf{x}_k - \mathbf{v}_i)^T(\mathbf{x}_k - \mathbf{v}_i) \quad (2)$$

With cluster center $\mathbf{v}_i$ is calculated by form:

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} {u_{ik}}^m \mathbf{x}_k}{\sum_{k=1}^{N} {u_{ik}}^m} \quad (3)$$

The function FCM takes a data set and a desired number of cluster and returns optimal cluster centers and membership grades for each data point. It starts with an initial guess mark the mean location of each cluster. The initial guess for this cluster center is most likely in correct. Next, FCM assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point. FCM iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represent the distance from any given data point to a cluster center. Objective function that was used in the FCM are [2]:

$$J_{FKR}(\mathbf{U}, \mathbf{X}, \mathbf{V}) = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m d_{ik}^2(\mathbf{x}_k, \mathbf{v}_i) \quad (4)$$

Assuming constraint

$\sum_{i=1}^{c} u_{ik} = 1$, for $\forall\, k \,\epsilon\, \{1, \dots, N\}$

With:

- c is the number of cluster $(2 \leq c < N)$
- $m \geq 1$ is degree of fuzziness in the cluster or weight from membership value.
- u_{ik} is degree of membership for object k belonging to i cluster (elements of matrix **U**)
- N is the number of observation
- ${d_{ik}}^2$ is Euclid distance

Algorithm of Fuzzy C-Means [2]

a) Determine the number of cluster (c)
b) Determine the degree of membership (m)
c) Initialization matrix $\mathbf{U}^{(0)}$ with
$$u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1, 0 < \sum_{k=1}^{n} u_{ik} < n \text{ for } \forall\, k \in \{1, 2, \dots, n\}$$
d) Calculate the center of cluster by using equation (2)
e) Update element of matrix **U** by using equation (1)
f) Compare value of membership in matrix **U**. If $\Delta < \varepsilon$algorithms already converging and iterating terminated otherwise we have to return to step 3, with:

g) $\Delta = abs\ (\boldsymbol{U}^{r+1} - \boldsymbol{U}^{r})$

ε = *threshold*. *Threshold* is a positive numbers that small to near zero, 0.00001 (10^{-5})
r = iterating process 1, 2,...

2.2 K-Medoid Clustering Method

K-Medoid is one of clustering method, which is similar to K-Means but in K-Means using means as centroid while K-Medoid using median as center of each cluster [1]. One of the algorithms that often used in k-medoid is Partitioning Around Medoid (PAM). This method use the data are being in middle the cluster, so that this method more robust than K-Means method [6].

One approach in the methods of PAM is by using an optimization mode presented by Vinod (1969) in Kaufman and Rousseeuw (1990). Let ${}_nX_p$ is the set of objects with n objects and p variables. The dissimilarity between objects x_i and x_j is denoted by d (i, j). The selection of objects as representative objects in clusters: y_i is defined as a 0-1 variable, equal to 1 if only if object i (i = 1, 2, …, n) is selected as a representative object.

The assignment of each object j to one of the selected representative objects: z_{ij} is a 0-1 variable, equal to 1 if and only if object j is assigned to the cluster of which i is the representative object (and also the medoid).

The corresponding optimization mode, which was first proposed by Vinod (1969), can then be written as:

$$Minimize \sum_{i=1}^{n} \sum_{j}^{n} d(i,j) z_{ij} \quad (5)$$

Subject to

$$\sum_{i=1}^{n} z_{ij} = 1, j\ = 1, 2, \dots, n \quad (i)$$

$$z_{ij} \leq y_i, i, j = 1, 2, \dots, n \quad (ii)$$

$$\sum_{i=1}^{n} y_i = k, k = number\ of\ clusters \quad (iii)$$

$$y_i, z_{ij}\ \in \{0, 1\}, i, j = 1, 2, \dots, n \quad (iv)$$

Constraints (i) express that each object j must be assigned to a single representative object. They imply together with constraints (iv) that for a given j, one of the z_{ij} is equal to 1 and all others are 0. Equation (iii) expresses that exactly k objects are to be chosen as representative objects. As the clusters are formed by assigning each object to the most similar representative object there will be exactly k nonempty clusters. Equation (i) implies that the dissimilarity between an object j and its representative object is given by

$$\sum_{i=1}^{n} d(i,j) z_{ij} \quad (6)$$

As all objects must be assigned, the total dissimilarity is given by

$$\sum_{j=1}^{n}\sum_{i=1}^{n} d(i,j)z_{ij}$$

This is the function to be minimized in the model.
The steps of the PAM are as follows [12]:

1) Select k objects as medoid arbitrarily from all the objects as the initial k clusters.
2) Distribute the remaining objects to their most similar cluster with the shortest distance.
3) Randomly select non-medoid object O'.
4) Compute the distance of O' and all the other objects
5) Set O' as new medoid if the total distance is decreased
6) Repeat the step 2 to 5 above until all medoid don't change anymore

2.3 Compare the clustering results

The best method is a method that produces minimum value of the average distance between objects and center in their own cluster, the higher ratio of the variance and the minimum objective functions.

3 Results and Conclusion

3.1 K-Medoid Clustering Method for Cluster District/Cities in Indonesia

There are several well-known methods for *k*-medoid clustering in the literature. One of the most popular is partitioning around medoid (PAM) from Kaufman and Rousseeuw (1990). This algorithm requires the number of clusters, *k*, be to known *a priori*. To find the *k* medoid, PAM begins with an arbitrary selection of *k* medoid. Then, in each step, a swap between a selected object and a non-selected object is made, as long as such a swap would result in improvement of the quality of the solution.
The experimental results with K-Medoid algorithm is presented in Table 1.

Table 1 Result of K-Medoid

Cluster	Number of District/Cities	Center of Cluster
1	99	Kota Palembang (Prov. Sumatera Selatan)
2	221	Kab. Bengkulu Utara (Prov.Bengkulu)
3	158	Kab. Landak (Prov.KalBar)
4	19	Kab. BovenDigoel (Prov. Papua)

As can be seen in Table 1, the most district/cities in Indonesia grouped in second cluster. It means most of district/cities in Indonesia could obtain high human development. The result of K-Medoid clustering is presented as a figure 1. The most

of district/cities in Sumatra, Java and Borneo grouped into high human development category in HDI achievement. The district/cities in Sulawesi grouped into high human development and medium human development. In Papua, most of the district/cities grouped into low human development. Meanwhile, Riau, North Sumatera, and Jakarta are the district/cities that could achieve very high human development.

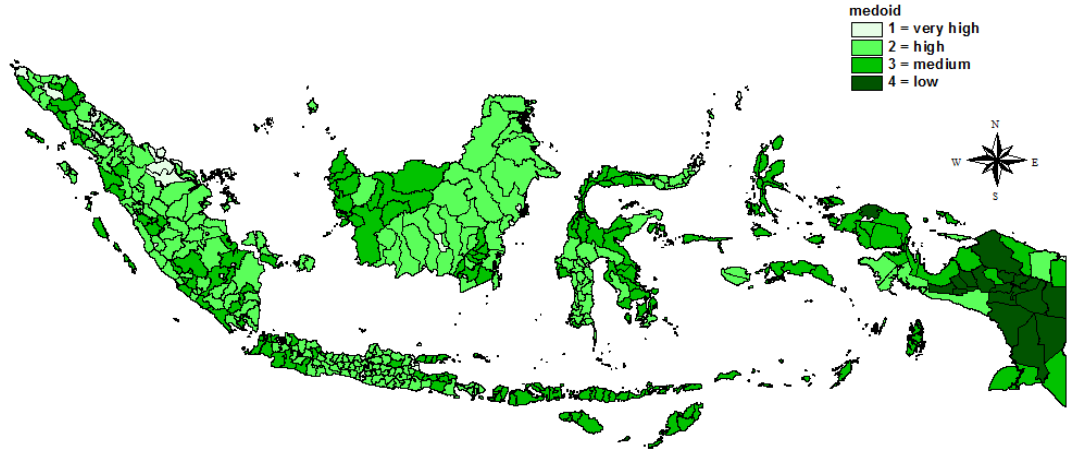

Figure 1.Map of the area with the K-Medoid clustering method

3.2 Fuzzy K-Means Clustering Method for Group district/cities in Indonesia

The existences of district/cities in each cluster depend on the value of membership. Suppose Simeulue, the value of membership is 0.14818 for first cluster, 0.2591 for second cluster, 0.55828 for third cluster, and 0.03444 for forth cluster. The highest membership value is in cluster 3. So, Simeulue is a member of cluster 3.

Fuzzy K-Means clustering method grouped district/cities in Indonesia into 121 districts/cities in cluster 1, 192 districts/cities in cluster 2, 165 districts/cities in cluster 3 and 19 districts/cities in cluster 4. The result of Fuzzy K-Means clustering can be seen in Figure2.

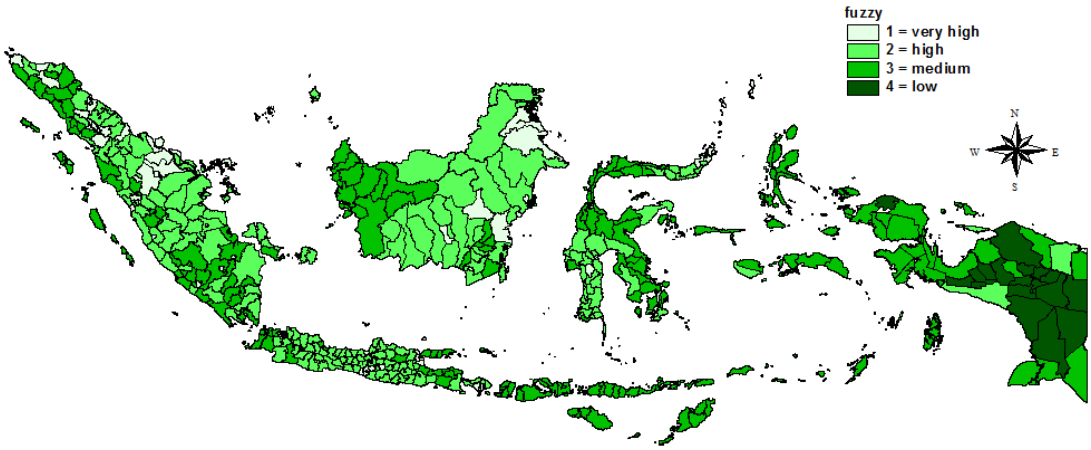

Figure 2.Map of cluster areas with Fuzzy K-Means

Based on Figure 2, most of district/cities in Sumatra, Kalimantan and Sulawesi grouped into the high human development and medium human development. The most district/cities of Java Island grouped into the second cluster (high human development). The district/cities in east Indonesia grouped into medium human development. The district/cities in West Papua grouped into low human development categories. The district/cities can achieve a very high human development is North Sumatra, West Sumatra and Jakarta.

3.3 Goodness Clustering Method

Measure the quality of clustering by dissimilarity/similarity metric. There is a separate quality function that measures the goodness of a cluster. It used to measure the distance between object and center. Table 2 and Table 3 show the average distance between object and center in their own cluster and the average distance between object to center in other cluster. The average distance between objects and center in their own cluster can be seen in the diagonal elements.

Table 2 The Average distance between object and center in cluster by using K-Medoid Clustering Method

Object	Center			
	1	2	3	4
Cluster 1	0.84	1.92	3.07	7.62
Cluster 2	2.06	0.99	1.95	6.28
Cluster 3	3.22	1.98	1.22	5.80
Cluster 4	7.61	6.22	5.76	1.64

Table 3 The Average distance between object and center in cluster by using Fuzzy C-Means Clustering Method

Object	Center			
	1	2	3	4
Cluster 1	0.94	1.96	2.78	7.49
Cluster 2	1.84	0.91	1.64	6.24
Cluster 3	2.96	1.83	1.23	5.74
Cluster 4	7.38	6.16	5.70	1.64

Table 2 and Table 3 show the smallest value placed in diagonal element. So, each object in cluster placed in the right cluster. The other method to find the best cluster is measuring the variance. The best method of clustering is determined by the smallest variance value within cluster and the highest variance value between clusters.

Table 4 The value of within and between cluster variance in FCM and K-Medoid Clustering.

Method	Tr W	Tr B
K-Medoid	125620.57	157121.11
Fuzzy K-Rataan	110433.78	172307.98

As can be seen in Table 4 Fuzzy C-Means shows the smallest value for within cluster (Tr W = 110433.78) and the highest value for between cluster variance (Tr B = 172307.98). It means FCM more homogeneity in cluster than K-Medoid. The experimental result with both K-Medoid and FCM algorithm are presented in Table 5.

Table 5 Goodness Clustering Methods

Method	The Average Distance Between Object And Center In The Own Cluster	The Average Distance Between Object And Center In The Other Cluster	The Average Distance Between Cluster Centroids	Approx F
K-Medoid	1.17481	3.26435	4.3568	360.54
Fuzzy C-Means	1.18306	3.12162	4.1147	386.71

Table 5 reported a similar value of the average distance between object and center in the own cluster, the average distance between object and center in the other cluster, the average distance between cluster centroids. On the other side, the approxF shows different value. The approx F of FCM is higher than K-Medoid, which means FCM produced higherhomogeneity in cluster.

4. Conclusion and Remarks

Fuzzy C-Means and K-Medoid produced the similar result for each member in cluster. Based on the average distance and variance, both of the methods have similar value. It means K-Medoid and Fuzzy C-Means are good enough for clustering district/cities in Indonesia by using variables of HDI, but FCM more homogeneity than K-Medoid.

Acknowledgment

This study is support by Statistics Indonesia Agency (BPS) of Indonesia for providing Human Development Index data.

References

[1] Alsulaiman ET (2013), "Classifying Technical Indicators Using K-Medoid Clustering, " Journal of Trading 8(2): 29-39.
[2] Bezdek J (1981), "Pattern Recognition with Fuzzy Objective Function Algorithm, " New York: Plenum Press.
[3] [BPS]. Badan Pusat Statistik (2014), "Indeks Pembangunan Manusia 2013, " Jakarta (ID): BPS.
[4] Gosh S, Dubey SK (2013), "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, " Journal of Advanced Computer Science and Application 4(4): 35-39.
[5] Johnson R, Wichern D (2007), "Applied Multivariate Statistical Analysis.Sixth Edition, " New Jersey: Pearson Education.
[6] Kaufman L, Rousseau JP (1990), "Finding Groups in Data: An Introduction to Cluster Analysis, " New Jersey: John Wiley & Sons.
[7] Li MJ, Ng MK, Cheung Y, Huang JZ (2008), "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Cluster", Journal of IEEE Transactions on Knowledge and Data Engineering 20(11): 1519-1534.
[8] Simbolon, Cary L (2013), "Clustering LulusanMahasiswaMatematika FMIPA UNTAN Pontianak MenggunakanAlgoritma Fuzzy C-Means", BuletinilmiahMatematikadanTerapannya(Bimaster) 2(1): 21-26.
[9] Singaravelu S, Sherin A, Savitha S (2013), "Agglomerative Fuzzy K-Means Clustering Algorithm", Journal of Nehru Arts and Science College 1(1): 16-20.
[10] Singh T, Mahajan M (2014), "Performance Comparison of Fuzzy C Means with Respect To other Clustering Algorithm", Journal of Advanced Research in Computer Science and Software Engineering 4(5): 89-93.
[11] Sivarathri S, Govardhan A (2014), "Experiments On Hypothesis Fuzzy K-Means is Better Than K-Means for Clustering", Journal of Data Mining&Knowledge Management Process 4(5): 21-34.
[12] Zhao Y, Liu X, Zhang H (2013), "The K-Medoids Clustering Algorithm with Membrane Computing", Telkomnika 11(4): 2050-2057.