www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-3, May-June 2017

International Journal of Engineering and Management Research

Page Number: 387-390

Comparing K-Means and Fuzzy C-Means Clustering (Case: Clustering of Provinces in Indonesia Based on the Indicator of the Health Service in 2015)

Ade Suryani Hamur[1], Budi Susetyo[2], Indahwati[3]
[1,2,3]Department of Statistics Bogor Agricultural University, INDONESIA

ABSTRACT

Health is one of the main factors to develop and found of human resources. The health status of a country is greatly influenced by health service indicators such as health facilities, health workers and health financing. However, these indicators have not spreaded evenly throughout Indonesia. Therefore, grouping of provinces in Indonesia needs to be done in order to see the provincial groups that do not have adequate health services. One of the analysis methods is K-Means, which keeps an objects into a specific cluster. This method is known as hard clustering. Another approach in clustering is based on a fuzzy sets theory which is known as fuzzy clustering. Each element has the probability to become a member of each group like Fuzzy C-Means (FCM). The purpose of this research was to compare the K-Means and FCM in clustering case of provinces in Indonesia based on the health service indicators in 2015. Data used were secondary data taken from health profil of Indonesia in 2015 sourced from the result of the National Sosio-Economic Survey (Susenas). The best clustering was chosen by minimizing ratio of standard deviations in groups and between groups. The results showed that provinces of Indonesia were divided into five cluster with different characteristics. The best clustering was given by the FCM method which had the smallest ratio that was 1.385.

Keywords-- K-Means, Fuzzy C-Means

I. INTRODUCTION

Health is one of the main factors to develop and found of human resources. The health status of a country is greatly influenced by health service indicators such as health facilities, health workers and health financing. However, these indicators have not spreaded evenly throughout Indonesia, especially in remote region [1]. Therefore, grouping of provinces in Indonesia needs to be done in order to see the provincial groups that do not have adequate health services. This is done by dividing the provinces into several groups so that the provinces in the same group has a high similarity compared with the provinces in different groups.

In multivariate analysis one of the grouping method is often done by cluster analysis. In general, cluster analysis can be divided into two methods: hierarchical clustering and nonhierarchical clustering methods. Nonhierarchical methods often used is the K-Means clustering [3]. K-means keeps an objects into a specific cluster. This method is known as hard clustering. Another approach in clustering is based on a fuzzy sets theory is known as fuzzy clustering. Fuzzy sets theory has a value of fuzziness between right or wrong, to know the value between zero and one [6]. So, Each element has the probability to become a member of each cluster. One of the fuzzy clustering methods is Fuzzy C-Means (FCM).

The purpose of this research was to compare the K-Means and FCM in clustering case of provinces in Indonesia based on the health service indicators in 2015.

II. METHODOLOGY

Data used were secondary data taken from health profil of Indonesia in 2015 sourced from the result of the National Sosio-Economic Survey (Susenas). Observation units used in this research was 34 provinces in ndonesia. The variable to be studied was a follows:

- The rasio of health center per 30.000 population (X_1)
- The rasio of hospital per 100.000 population (X_2)
- The rasio doctor per health center (X_3)
- The rasio of nurses per health center (X_4)
- The rasio of midwife per health center (X_5)
- The rasio of five types of health workers promotive and preventive per health center (X_6)
- The rasio of doctor per 100.000 population (X_7)
- The rasio of dentist per 100.000 population (X_8)

- The rasio of public health personnel per 100.000 population (X_9)
- The rasio of environmental health worker per 100.000 population (X_{10})
- Percentage of realization of health decontration funds to ceiling of RKA-KL DIPA (X_{11})
- Percentage of realization of health assistance task funds to ceiling of RKA-KL DIPA (X_{12})
- Percentage of realization of these funds to APBD (X_{13})

Procedur DataAnalysis

The data used were processed using R 3.3.2 and SAS 9.1.3. The step of data analysis conducted in this research were:

1. Saw the number of clustering that possibly to formed using FCM and *K-Means* methods through dendogram the hirarchical method.
2. Doing clustering with FCM and K-Means methods.

Algorithm of FCM [2]:

a. Determine of number c of cluster, $2 \le c < n$, n number of objects.

b. Determine of *weighting exponent* (m). In this research m = 2.

c. Initialize $\mathbf{U}^0$ membership matrix randomly (u_{ik}). Calculate membership values using:

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

where

u_{ik} : membership value

d_{ik} : distances between object and cluster centroids.

m : *weighting exponent*, $1 \le m < \infty$

c : number of cluster

d. Calculate the cluster centroids using:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \times x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad (2)$$

e. Calculate the objective function at t^{th} iteration.

$$J_{FCM}(\mathbf{U}, \mathbf{v}) = \sum_{k=1}^{n} \sum_{i=1}^{c} d_{ik}^2 (u_{ik})^m \quad (3)$$

with the constraints

$$u_{ik} \in [0,1]; \ \sum_{i=1}^{c} u_{ik} = 1 ; \ d_{ik} = d_{ki} \ ; \ m \ge 1$$

f. Update element of matrix $\boldsymbol{U}$.

g. Compare $\boldsymbol{U^t}$ with $\boldsymbol{U^{t-1}}$.

If $\left| U^t - U^{t-1} \right| < \varepsilon; t = 1,2,\ldots$ where t is the iteration process and ε is value of threshold, then stop else return to the step *c*.

h. Calculate the value of cluster validity index.

Algorithm of *K-Means* [3]:

a. Determine of number of cluster and centroid (mean) on each cluster. Cluster centroids were calculated using:

$$v_{ij} = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{kj} \quad (4)$$

where

x_{ik} : the observed value of k^{th} object to j^{th} variable

n_{ik} : number of data in i^{th} cluster

b. Assigning an object to the cluster whose centroid (mean) is nearest.

c. Recalculate the centroid for the cluster receiving the new object and for the cluster losing the item

d. Repeat the step *b* until all object don't change anymore.

3. Comparing the result of clustering formed from K-means and FCM methods by looking at the ratio of standard deviations in groups and between groups (S_w/S_B). The mean of standard deviation in groups were calculated using [4]:

$$S_w = \frac{1}{c} \sum_{i=1}^{c} S_i \quad (5)$$

where

S_i : standard deviation of i^{th} cluster

while standard deviation in between groups were calcuted using:

$$S_B = \sqrt{\frac{1}{c-1} \sum_{i=1}^{c} \left(\bar{X}_i - \bar{X} \right)^2} \quad (6)$$

where

$\bar{X}_i$: mean of objects in i^{th} cluster

$\bar{X}$: mean of objects in all cluster

III. OUR APPROACH

Clustering of Provinces in Indonesia using FCM method

The first step on nonhierarchical clustering method is finding number of cluster. Looking at the initial description of the number of clustering that possibly to formed then hierarchical clustering method was firstly perfomed which was shown by the dendogram (Figure 1). Based on the dendogram the objects could be grouped into 3 until 6 clusters. Therefore, number of cluster determination to get many optimum cluster by FCM method started from 3 until 6 clusters with 5 repetitions on each cluster. Repetition was performed because the initial matrix of membership cluster was randomly initialized.

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

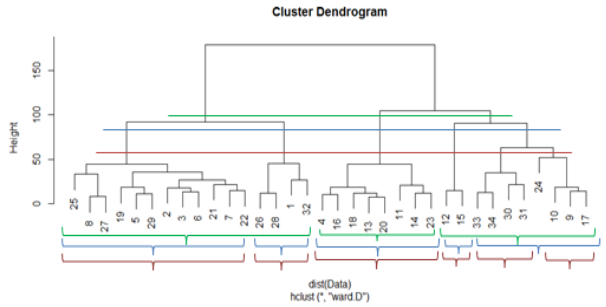

Figure 1 Dendogram of Hirarchical Method

The value of the objective function in the FCM method for each replication can be seen in figure 2(a).

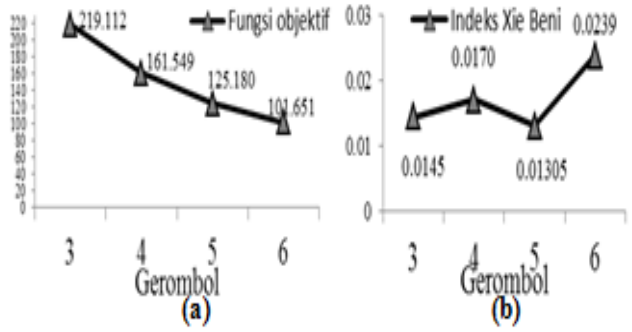

Figure 2 (a) The value of the objective function, (b) XB Index

It gave relatively the same result for each repetition of each cluster. It decreased inline with the increasing of number of cluster. The identification of many optimum cluster based on the XB index could be seen in Figure 2 (b). The minimum of *XB* when clusters were five then rose up back on the six clusters, so that the optimum number of cluster was five clusters.

Membership determination in each cluster was depend on the greatest probabiility of the object in clusters. Suppose that Aceh Province had the value of membership was 0.095 for being first cluster, 0.345 for second cluster, 0.134 for third cluster, 0.183 for forth cluster and 0.244 for fifth cluster. The highest membership value was in cluster 2. So, Aceh Provinces was a member of second cluster. In the same way for other provinces were obtained in Table 1.

Table 1 Members of FCM cluster

Cluster	Name of Provinces
1	DKI Jakarta, West Java, East Java
2	Aceh, Maluku, North Maluku, West Papua, Papua
3	NTB, West Kalimantan, East Kalimantan, North Sumatera, West Sumatera, Riau, South Sumatera, Lampung, Central Java, DI Yogyakara, Banten
4	Bali, Kep.Bangka Belitung, Kep. Riau, North Kalimantan, West Sulawesi
5	Jambi, Bengkulu, NTT, Central Kalimantan, South Kalimatan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo

Seeing the difference in mean of each cluster, Manova test was done and Wilk's Lambda value was obtained as 0.002893 with small enough p-value that was less than 0.0001, so Ho was rejected with a 95% confidence level which meant there was a difference of mean between the formed clusters.

Cluster 1 was characterized by health center, public health personnel and environmental health worker which was minimum, as well as realization of health decontration funds. Provinces on cluster 2 was characterized by high health facilites of health center, Low health workers including types of health workers promotive and preventive and dentist, as well as realization of health decontration funds. Provinces on cluster 3 had the most dominant characteristic of hight performing midwives and high health financing after the provinces of cluster 5. The characteristic of provinces in cluster 4 were almost all the values of health service indicators were high compared to provinces in other clusters. However, health financing such as the realization of health decontration funds and the realization of health assistance task funding were low. The last, provinces of cluster 5 had similar characteristics based on health financing that was high.

Clustering of Provinces in Indonesia using K-Means Method

In contrast to the FCM method, each object in the K-Means method only had the probability to enter one cluster. Clustering with K-Means allocated data into each cluster based on the comparison of the distance between the data with the center of each existing cluster.

Data were reallocated explicitly to the cluster that had the nearest center to the data. Members of each cluster of the K-means method were presented in Table 2.

Table 2 Members of K-Means cluster

Cluster	Name of Provinces
1	DKI Jakarta, West Java, East Java, West Sulawesi, Maluku
2	Aceh, North Maluku, West Papua, Papua
3	North Sumatera, West Sumatera, Riau, Soulth Sumatera, Lampung, Central Java, DI Yogyakara, Banten, NTB, NTT, West Kalimantan, South Kalimatan, East Kalimantan, North Sulawesi, South Sulawesi
4	Bali, North Kalimantan, Kep.Bangka Belitung, Kep. Riau
5	Jambi, Bengkulu, Central Kalimantan, Central Sulawesi, Southeast Sulawes, Gorontalo

Member of cluster with FCM and K-Means were not much different. However, there were some differences in members of the cluster as in the FCM method, Maluku was in the cluster 2 the same cluster with Aceh, North Maluku, West Papua and Papua Province while in the K-Means method Maluku become a member of the cluster 1 joining the provinces of DKI Jakarta, Java West, East Java, and West Sulawesi. Viewed the value of Maluku membership with FCM method 0f 0.347 which was higher than probability 0.219 into cluster 1. In this case the result of clustering with FCM method was better than result by K-Means method. Futhermore, Manova test was perfomed and Wilk's Lambda (Λ^*) 0.032028 was obtained with small enough p-value which less than 0.0001 or <0.05 so with a 95% confidence level, we had sufficient evidence to reject H_0, that meant there was an average difference between the clusters formed.

Cluster Validity

A better comparison of clustering methods can be seen from the ratio of standard deviations in group and between group (Table 3).

Table 3 ratio of standard deviations in groups and between groups of K-Means and FCM methods

Variable	S_W/S_B	
	K-means	FCM
X_1	0.909	0.523
X_2	1.440	1.717
X_3	0.912	1.443
X_4	1.499	1.437
X_5	3.160	1.767
X_6	1.285	1.258
X_7	1.088	1.789
X_8	1.223	1.888
X_9	0.600	0.623
X_{10}	1.589	0.705
X_{11}	0.541	0.508
X_{12}	0.769	0.631
X_{13}	3.441	3.715
Rata-rata	1.420	**1.385**

The ratio of standard deviations in groups and between group of FCM method gave smaller values than the K-Means method. When viewed in each variable, almost all the variables in the FCM method gave smaller value than the K-Means method. In this case the FCM method was better to be used to grouping provinces in Indonesia based on health service indicators. The ineffectiveness of the K-Means method for the clustering of objects because each element must be a member of a particular cluster and not had probability to be a member of another cluster.

IV. CONCLUSION

Provinces in Indonesia based on health service indicators were grouped into five groups. The number of groups were based on the optimum cluster resulting from FCM method with 0.013 XB value. In the case, clustering method using membership degree gives a better result than the K-Means method with ratio of standard deviations in groups and between groups was 1.385.

REFERENCES

[1] Badan Pusat Statistik. (2016). Potret Awal Tujuan Pembangunan Berkelanjutan (Sustanable Development Goals) di Indonesia. Jakarta: Badan Pusat Statistik.

[2] Hopper F, Klawonn F, Kruse R, Runkler T. (1999). Fuzzy Cluster Analysis (Methods for Classification, Data Analysis and Image Recognition). New York: John Wiley & Son, Ltd.

[3] Johnson R, Wichern D. (2002). Applied Multivariate Statistical Analysis. Sixth Edition. New Jersey: Pearson Education.

[4] Kalkstein LS, Tan G, Skindlov JA. (1987). An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification. Journal of Climate and Applied Meteorolog. Vol. 26, pp. 717-730.

[5] Kementerian Kesehatan Republik Indonesia. (2016). Profil Data Kesehatan Republik Indonesia 2015. Jakarta: Kementerian Kesehatan Republik Indonesia.

[6] Zadeh LA. (1965). Fuzzy Sets. Information and Control. Vol. 8, pp. 338-353.