Research Article

Comparison Approaches of the Fuzzy C-Means and Gaussian Mixture Model in Clustering the Welfare of the Indonesian People

Erlin Windia Ambarsari*, Nurfidah Dwitiyanti, Noni Selvia, Wahyu Nur Cholifah, Putri Dina Mardika

Informatics Department, Universitas Indraprasta PGRI, Jakarta, Indonesia

Abstract.
We compared the previous study about clustering the welfare of the Indonesian people using the Fuzzy C-Means (FCM) approach to a recent study, the Gaussian mixture model (GMM). Both of which were soft clustering. The case analyzes by classifying 34 provincial data in Indonesia, based on eight welfare of people indicator variables in 2017, which the Central Statistics Agency had issued. We compared the FCM and the GMM approaches to determine a better level of accuracy in clustering data using the Silhouette index, the Davies-Bouldin index, and the Calinski-Harabasz index values as a validity test method. The FCM and GMM methods found that the optimal clusters were 2 and 6. When we observed the consistency of the three tests' validity results, the GMM method was preferable to the FCM clustering method.

Keywords: fuzzy, Gaussian mixture model, clustering

Corresponding Author: Erlin Windia Ambarsari; email: erlinunindra@gmail.com

Published 26 May 2023

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICASI Conference Committee.

1. Introduction

Datasets commonly interpreted depend on what is analysis needs. One method was to analyze data with clustering. Several cases had handling differences, such as grouping data that had a label to observe, which correlated—forming a Venn diagram to determine the data relationship between datasets [1], [2]. However, some data have characteristics that need observation to discover patterns and behavior. Therefore, clustering datasets had two types, specifically hard clustering and soft clustering. Hard clustering had happened when the datasets had explicit clusters separated from each other; soft clustering was when some data had an intersection of sets and overlap [3].

The cluster had overlap needed to identify datasets that did not become obvious to the tendency towards particular groups using unsupervised clustering. A previous

How to cite this article: Erlin Windia Ambarsari*, Nurfidah Dwitiyanti, Noni Selvia, Wahyu Nur Cholifah, Putri Dina Mardika, (2023), "Comparison Approaches of the Fuzzy C-Means and Gaussian Mixture Model in Clustering the Welfare of the Indonesian People" in *International Conference on Advance & Scientific Innovation*, KnE Life Sciences, pages 16–22. DOI 10.18502/kss.v8i9.13315

study [4] used the Fuzzy C-Means method as soft clustering to identify the Welfare of the Indonesian People. The concept of the FCM algorithm is dividing a finite collection of points into an aggregate of clusters based on predetermined criteria. As a result, points on the cluster's periphery possibly in the cluster to a lower extent than points in the cluster's center [5].

The prior study observed that the Welfare of the Indonesian People divided clusters became two: the welfare cluster (16 provinces) and not the welfare cluster (18 provinces). Density of population, poverty rate, growth rate, life expectancy, school participation rate, labor force participation rate, open unemployment rate, and average spending per capita are all factors to consider as an indicator to build clusters.

Then we try to figure out how many clusters the Mixture Modeling can create. There are two methods to apply the Mixture Modeling: classification and clustering. Mixture model clustering applies two distributions, such as gaussian and multinomial. Gaussian Mixture Modelling that we chose to solve clustering for Welfare of the Indonesian People. Several cases in [6]–[8] had Gaussian distribution and used data training with Expectation-Maximization (EM). However, certain cases use different approaches depending on the handling of the dataset.

2. Methods

The data obtained from the previous case [4] included 34 provinces on a sample of Indonesian people—the variable based on people's welfare gained from the Central Statistics Agency in 2017 as an indicator. Therefore, the algorithm in the study using Gaussian Mixture Modeling (GMM) to observe how many clusters developed, then compared to clusters of FCM as solve the problem in an article [4], as shown below:

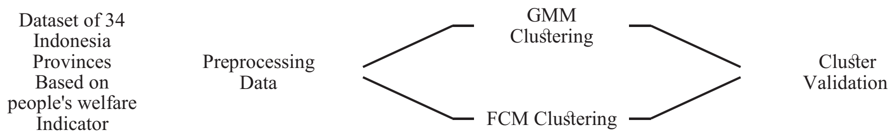
Dataset of 34 Indonesia Provinces Based on people's welfare Indicator
Preprocessing Data
GMM Clustering
FCM Clustering
Cluster Validation

Figure 1: The Data Clustering Workflow.

Preprocessing data performs preparation and transformation according to mining procedures [9]. The first step carried out used two feature scaling methods, including normalization and standardization. The formula of both features is as follows:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} (1)$$

$$x_{std} = \frac{x - \mu_x}{\sigma_x} (2)$$

Where x is the data feature, x_{min} and x_{max} are the minimal and maximum value of x, μ_x is the average, and the standard deviation σ_x [10]. After normalizing and standardizing, then in the next step, do data clustering. The earlier study used FCM clustering based on cluster center determination to mark locations based on the average for each cluster. The technique is to continually correct the cluster of a central point and the membership degree of each data point until the cluster midpoint moves to the precise location. The iteration minimizes an objective function that describes the range between a particular data point and the cluster's center, weighted by the data point's degree of membership.

The result of Fuzzy C-Means is a series of cluster centers and several degrees of membership for every data. Meanwhile, in grouping the data with GMM clustering based on the similarity between individuals using a probabilistic model approach. The first step in GMM clustering is to identify the number of clusters determined from the data set. The data are assumed to originate from a mixture of two or more probability distributions with specific proportions. Data clustered using Gaussian Mixture Models (GMM), a mixture of the Gaussian probability distribution. Each distribution represents a group with specific parameters. The parameter is assumed to use the Expectation-Maximization (EM) algorithm with the parameter's initial value obtained from agglomerative hierarchical clustering. The effectiveness of GMM on the data be able known by calculating the average level of misclassification. Another condition considered in generating data is the distance among the middle of the cluster and the diversity of each cluster to observe the method's effectiveness if the groups are far from each other, close together, or overlap.

To overcome the difficulty of the clustering algorithm with determining the correct amount of clusters based on the data used, then using a validity index to assess the output of the clustering algorithm to get the best number of groups [11]. Therefore, three validity indices used relative criteria in the study: the Silhouette value, the Davies-Bouldin index, and the Calinski-Harabasz index.

3. Result and Discussion

Grouping with two predetermined algorithms for observed which one to appropriate model: GCM or FCM Clustering. After preprocessing the data, then use the optimum amount of clusters to search for the grouped data. The K value is determined by looking at the silhouette value. The silhouette value's function for the exposition and verify method of sensitive cluster data. This process provides a graphical rendition of how well each object is located within the cluster [12]. The Silhouette uses in length among (-1) to 1. The higher the value, the better the quality [13]. The results of the silhouette values as shown in table 1 with the number of K, which is 2 to 7.

Table 1: Silhouette Scores on FCM and GMM.

k	FCM	GMM
2	**0.4894812**	0.261252
3	0.1380583	0.188988
4	0.2461348	0.221910
5	0.2038619	0.235964
6	0.219264	**0.262161**
7	0.1515257	0.221327

Table 2: Davies-Bouldin Index value on FCM and GMM.

k	FCM	GMM
2	1.584991	1.490767
3	1.547854	1.471133
4	1.697148	1.412524
5	1.47431	1.244925
6	1.451747	**1.143779**
7	**1.436499**	1.183180

Table 3: Calinski-Harabasz Index values on FCM and GMM.

k	FCM	GMM
2	**10.13987**	6.824538
3	6.833079	10.038688
4	6.46217	9.272092
5	6.145671	9.563079
6	5.783985	**10.279471**
7	5.479927	9.776508

Table 1 indicates the results of GMM clustering. The number of optimal clusters in six groups with a silhouette validity index of 0.262161, where cluster 0 has four provinces;

cluster 1 has six; cluster 2 has seven; cluster 3 has six; cluster 4 has six; cluster 5 has five. Previous FCM investigations had two groups, each with a silhouette validity index of 0.4894812, with cluster 1 having 18 provinces and cluster 2 having 16 areas. In addition, we noticed the DBI value in table 2 as well. The ideal number of groups in the FCM technique is in cluster 7, where the DB value is 1.436499, which is less than the other DB values. Better clustering performance is indicated by a lower DB value [14]. We find that the FCM approach has a different number of clusters than the initial Silhouette index value. The reasons include whether the distance between the two cluster centers calculated during the clustering process has attained convergence or not because the distance between the two cluster centers has altered throughout the clustering process. As for the GMM method, the optimal number of clusters is found in cluster 6, the same as the number of groups shown in the previous Silhouette index value.

Table 3 shows the value of the Calinski-Harabasz index. The substantial weight of the Calinski-Harabasz index indicates the best number of clusters [15]. In the FCM method, an enormous Calinski-Harabasz value is 10.13987, located in cluster 2. Meanwhile, in the GMM method that an immense Calinski-Harabasz value is 10.279471, located in cluster 6. The Calinski-Harabasz index value corresponds to the optimal number of groups at the silhouette index value.

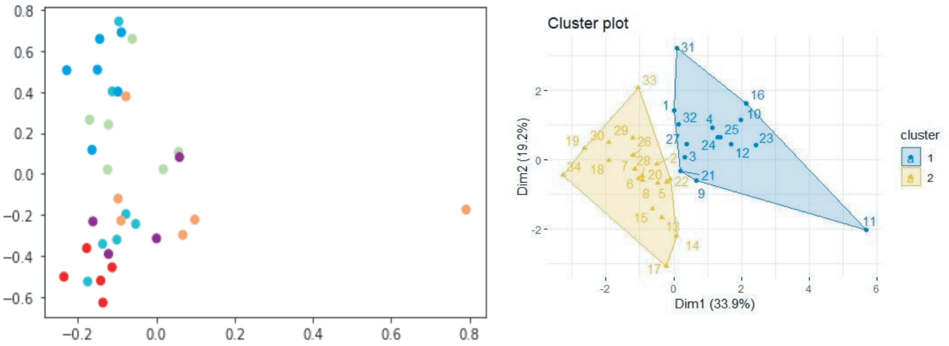

Figure 2: Result of Data Clustering on GMM dan FCM [4].

The GMM technique identifies the existence of data overlaps, as seen in Figure 2. The Gaussian distribution and the average of each cluster organized data in GMM. Meanwhile, FCM groups data based on the degree of membership and data allocation in each set determined by the cluster's center and target function (object function) [16]. The two clustering algorithms, GMM and FCM, work differently, resulting in various groups. Furthermore, the data type specifies the number of clusters generated and the amount of data features available. Then, to separate the overlapping data provided by GMM, more observations are required in a future study.

4. Conclusions

The validity index values were tested, including the Silhouette, Davies-Bouldin, and Calinski-Harabasz indices. The results of grouping 34 province data with people's welfare indicators in 2017 with two soft clustering methods, such as the FCM and GMM methods, found the number of clusters in the optimal results was 2 and 6. When observed from the consistency of the validity results of the three tests, We said that the test on 34 province data shows that the quality of clusters obtained from the GMM method is better than the FCM method. The quality of groups obtained from data type influenced the clustering method and the number of available data features.

Acknowledgments

References

[1] Kustian N, Julaeha S, Parulian D, Selvia N, Ambarsari EW. Venn versus relation diagram models for database relation in SQL command line. J Phys Conf Ser. 2021 Feb;1783(1):012050.

[2] Yang XD, Tan HW, Zhu WM. SpinachDB: A well-characterized genomic database for gene family classification and SNP information of spinach. PLoS One. 2016 May;11(5):e0152706.

[3] Dimitriadis SI, Messaritaki E, Jones DK. The impact of graph construction scheme and community detection algorithm on the repeatability of community and hub identification in structural brain networks. Hum Brain Mapp. 2021 Sep;42(13):4261–4280.

[4] Dwitiyanti N, Selvia N, Andrari FR. Penerapan Fuzzy C-Means Cluster dalam Pengelompokkan Provinsi Indonesia Menurut Indikator Kesejahteraan Rakyat. Fakt Exacta. 2019;12(3):201–209.

[5] Rajkumar KV, Yesubabu A, Subrahmanyam K. Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset. Iran J Electr Comput Eng. 2019;9(4):2760–2770.

[6] Lin X, Yang X, Li Y. A Deep clustering algorithm based on Gaussian Mixture Model. J Phys Conf Ser. 2019;1302(3):032012.

[7] Baid U, Talbar S, Talbar S. “Comparative study of K-means, Gaussian Mixture Model, Fuzzy C-means algorithms for Brain Tumor Segmentation.” In Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016), 2017;137:592–597.

[8] Rashid S, Ahmed A, Al Barazanchi I, Jaaz ZA. Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set. Period Eng Nat Sci. 2019;7(2):448–457.

[9] Alasadi SA, Bhaya WS. Review of data preprocessing techniques.pdf. J Eng Applie Sci. 2017;12(16):4102–4107.

[10] Patel E, Kushwaha DS. “Clustering cloud workloads: K-Means vs Gaussian Mixture Model.” Procedia Comput Sci. 2019;171:158–167. https://doi.org/10.1016/j.procs.2020.04.017

[11] Khairati AF, Adlina AA, Hertono GF, Handari BD. “Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA,” in Prosiding Seminar Nasional Matematika, 2019, vol. 2, pp. 161–170.

[12] Subbalakshmi C, Krishna GR, Rao SK, Rao PV. A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set. Procedia Comput Sci. 2015;46:346–53.

[13] T. HENDRAWAN NATA UTAMA. Analisis Performa Algoritma FUZZY C-MEANS dan K-MEANS Clustering untuk Pengelompokan Pelanggan pada PT. PART STATION JEMBER. Universitas Muhammadiyah Jember; 2019.

[14] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979 Feb;1(2):224–227.

[15] Baarsch J, Celebi ME. Investigation of internal validity measures for K-means clustering. Proc Int Multiconf Comp Sci Inf Technol. 2012;1:471–476.

[16] Sari HL, Suranti D. “Perbandingan Algoritma Fuzzy C-Means (FCM) dan Algoritma Mixture Dalam Penclusteran Data Curah Hujan Kota Bengkulu.” In Proceeding SNATI 2016. 2016. pp. 7–15.