UNP JOURNAL MANAGEMENT CENTER

math.ppj.unp.ac.id
p-ISSN 2716-0726
e–ISSN 2716-0734

Indexed

Rangkiang Mathematics Journal
Volume 2, No. 1, 2023

Subject Areas:
Mathematics

Keywords:
K-Medoids Clustering, Data Mining, Euclidean Distance, Poverty

Author for correspondence:
Fara Zametha Elsa Reski
e-mail: farazamethaelsa01@gmail.com

RANGKIANG MATHEMATICS JOURNAL
Journal of Mathematics and Mathematics Education

Implementation of the Partitioning Around Medoids (PAM) Clustering Method on Poor Population Data in West Sumatera

Fara Zametha Elsa Reski[1] and Yusmet Rizal[2]

[1,2] Mathematics Department, Faculty of Mathematics and Natural Science, Universitas Negeri Padang, Padang, Indonesia

Abstract- The problem of poverty is complex so it becomes a priority in the development of a country. Various efforts have been made such as social assistance programs, but the assistance provided is not always evenly distributed and not on target. Therefore, it is necessary to determine priority handling for the community by grouping data on the poor, especially in the Province of West Sumatra, Indonesia. In this study, an analysis was carried out using all districts/cities in West Sumatra and poverty indicators, namely the percentage of poor people, poverty depth index, poverty severity index, literacy rate, the average length of schooling, expected length of schooling, and open unemployment rate. Grouping was carried out using the Partitioning Around Medoids (PAM) method, better known as the K-Medoids method. The research results obtained 2 poverty level clusters, namely cluster 1 is a high level with 11 districts/city members, and cluster 2 is a low level with 8 districts/city members.

1. Introduction

Poverty is one of the many problems that are the focus of the government's attention. Poverty is a condition when you are unable to meet basic needs such as food, clothing, education, housing, and health (Faisal & Utami, 2022). The Indonesian government already has many programs to eradicate poverty, but the number of poor people has not decreased significantly. Particularly in West Sumatra Province, data for September 2021 recorded that the number of poor people reached 134.53 thousand people and data for March 2022 recorded poor people in West Sumatra as many as 137.61 thousand people. This shows that there is an increase in the number of poor people in West Sumatra, so that the

West Sumatra regional government must work even harder in eradicating poverty problems.

Various efforts have been made by the government in overcoming the problem of poverty. The difficulty faced by the government in the process of alleviating poverty is that the distribution of social assistance is uneven and not on target. This is because data validation is often neglected, giving rise to inaccurate data. So it should be for the sake of fairness, this aid distribution activity must prepare really valid data about who is entitled to receive this assistance. Therefore, to determine priorities for providing assistance to the community, it is necessary to group or segment data on the poor or what can be called clusters.

The problem in the data taken from the Badan Pusat Statistik (BPS) is that the poverty rate in each region in West Sumatra is different, so it can be concluded that there are high and low poverty rates. According (Simhachalam & Ganesan, 2016), data mining is defined as an analysis process to find valid and unexpected relationships between data sets and convert data into data structures so that they are easy to understand and useful for users. Data mining aims to find the patterns and rules that is found in the data from the pattern and the rule can be done decision-making and predict the effect the decision (Chakraborty et al., 2022). To find out the relationship of the data base, data analysis techniques are needed.

Grouping is one of the descrption techniques of data mining analysis. One of the popular non-hierarchical clustering methods used is the kmeans method. K-means is also known as hard clustering which can group objects with clear boundaries, meaning that they can group objects into certain groups and not members of other groups (Sivarathri & A, 2014). The k-means method is a partition-based method that attempts to partition data into two or more groups using the mean value as the center of the cluster. In addition to the k-means method there is also the k-medoids method which is a partition-based method that uses medoids as the center of the cluster. Medoids is the most centralized cluster data object (Arora et al., 2016), so this method is more robust to outliers than the k-means method (Soni & Patel, 2017).

In this research, segmentation or grouping of data on the number of poor people in West Sumatra by district/city was carried out using the *K-Medoids Clustering method* , where this method is suitable to be applied so that it can be seen how the grouping of districts/cities is based on existing data available at the Badan Pusat Statistik (BPS).

Clustering is a data mining method that performs separation/splitting/segmentation of data into a number of groups (*clusters*) according to certain desired characteristic. One of the characteristics of clustering is good or optimal performance is if the produce cluster that contains data with the level of similarity (similarity) is high on the cluster and the same level of low similarity on different clusters (Jain et al., 1999). They partition the objects into groups or clusters, so that objects within a cluster are "similar" to one another an "dissimilar" to objects in other clusters. (Kamber & Han, 2018). Cluster analysis can be applied in the field of science,planners of marketing, social and industry (Berkhin, 2006).

Therefore this research was conducted to group or cluster data on the poor in West Sumatra using data mining with the *K-Medoids Clustering method* or also called *Partitioning Around Medoids* (PAM) *Clustering*. K-Medoids algorithm is the coupling method to retrieve the value of the average of the objects in a cluster as a point of reference, medoid screened is the object in a cluster is the most concentrated (Balabantaray et al., 2015). This method is a group of partitional clustering methods that use objects in a collection of objects to represent a cluster. The advantage of this method is that it is able to overcome the weaknesses of the *K-Means method* which is sensitive to outliers and the results of the clustering process do not depend on the order in which the dataset is entered (Atmaja, 2019). According to (Qona'ah et al., 2020) the steps of the *K-Medoids Clustering method* are as follows:

1. Determine k (number of groups)
2. Choose randomly the initial medoids *as much as* k from n data.
3. Computes each data (object) to the initial *medoids* using the *Euclidean Distance formula*. The *Euclidean Distance* formula is as follows:

$$d(x,y) = \sqrt{\sum\nolimits_{i=1}^{n} (x_i - y_i)^2} \,; i = 1,2,3,\dots,n$$

with :
$d(x,y)$ = data distance to x to the center of group y
n = the number of attributes in a data
i = data index
x_i = the value of object i in the variable x

y_i = the value of object *i* in variable *y*

4. Randomly selects an object from each group as a candidate for new medoids.
5. Calculates the distance between each object from each group with new candidate medoids.
6. Calculating the total deviation (S)

$$S = \text{Total } cost \text{ baru} - \text{Total } cost \text{ lama}$$

If $S < 0$, then swap objects with group data to form a new set of k objects as *medoids.*

7. Repeating steps 4-6 until there is no change in the medoids so that the groups and their respective group members are obtained.

To get the optimal number of K, one of them is by using the sillhoutte coefficient method, which is a combined method of cohesion that functions to measure the closeness of relationships with objects in a group, and separation, which aims to measure how far a group is separated from other groups (Xu et al., 2016) . Below is the formula for the sillhoutte coefficient :

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

To calculate the average *sillhoutte coefficient* :

$$SC = \frac{1}{n}\sum\nolimits_{i=1}^{n} s(i)$$

The data grouping gets better if the sillhoutte value is close to 1 and conversely, the data grouping gets worse if it gets closer to -1 (Struyf et al., 1997).

2. Methods

This type of research is applied research and the type of data used is secondary data. Data were obtained from the official website of the Central Statistics Agency (BPS). The main data source used in this study is the poverty factor dataset in Indonesia in the form of data on the percentage of poor people, poverty depth index, poverty severity index, literacy rate, the average length of schooling, expected length of schooling, and open unemployment rate in 2021 which are sourced from the Central Bureau of Statistics. Data received in Excel file form.

The steps of data analysis in this study are:

a. Retrieve the data to be used
b. Normalize data
c. Determine optimal K using the Silhouette Coefficient method with the help of RStudio software
d. Perform cluster analysis with the K-Medoids Clustering algorithm, by determining the size of the distance using the Euclidean method
e. Validation of the calculation results for each cluster with the Rstudio software system

3. Results and Discussion

The data in this research are accumulated data on poverty factors in the form of data on the percentage of poor people, poverty depth index, poverty severity index, literacy rate, the average length of schooling, expected length of schooling, and open unemployment rate in 19 districts/cities in the province. West Sumatra in 2021.

(a) Normalize each Data with the Min-Max method

Data is normalized using *min-max normalization*. It aims to change the data range to produce the same *range of values*. Normalized weights can be calculated using equation (1).

$$X' = \frac{X - min_X}{max_X - min_X} \qquad (1)$$

Information :
X' = normalized data
X = data to be normalized
To obtain the minimum and maximum values from the poverty data in West Sumatra Province which are listed in Table 1.

Table 1. An Example of a Table

Min/Max	X1	X2	X3	X4	X5	X6	X7
Min	0.1397	1.76	0.36	1	11.92	16.54	0.1169
Max	0.0228	0.34	0.05	0.9846	7.84	12.51	0.0139

The description for finding the X1 weight value with equation (1) is as follows:

$$X' = \frac{X - min_X}{max_X - min_X}$$

By knowing that the value of X(1,1) in the data in Agam Kabupaten is worth 0.0622, then we get:

$$X' = \frac{X - min_X}{max_X - min_X}$$

$$v' = \frac{0{,}0622 - 0{,}1397}{0{,}0228 - 0{,}1397}$$

min $v' = 0{,}33704$
Next to the variable X_2 until X_7 so on up to n data.

(b) Optimization of K C clusters with the Sillhoutte Coefficient Method

K-Medoids cluster analysis , it is necessary to determine the optimal number of K to be formed. Determination of the amount of K is carried out using the *Sillhoutte Coefficient method* with the help of the RStudio application. The silhouette can reflect the data is grouped that objects are organized into groups that match it (Thinsungnoen et al., 2015). The results of the *Sillhoutte Coefficient method* can be seen in Figure 1.

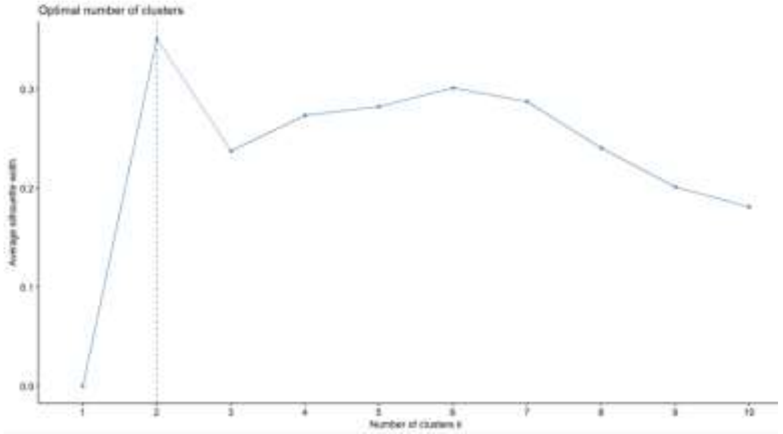


Figure 1. Determining the Number of Clusters Using the Sillhoutte Coefficient Method

Sillhoutte width value is when the number of clusters is 2 clusters, this can be seen by the dotted vertical line on the x-axis when k = 2 with a sillhoutte width value of 0.376166. Therefore, based on the results obtained from the Sillhoutte Coefficient method, it can be concluded that the optimum number of clusters using the K-Medoids method is 2 clusters.

(c) Clustering Analysis with the K-Medoids Algorithm

Data on poverty factors in the form of data on the percentage of poor people, poverty depth index, poverty severity index, literacy rate, the average length of schooling, expected length of schooling, and open unemployment rate in 19 districts/cities in West Sumatra Province in 2021 grouped using the *K-Medoids Clustering method* with the number of *K* = 2.

With the K = 2 group in the analysis, the clustering results were obtained with a total of 3 iterations. In this 3rd iteration, objects are selected as the new medoids shown in Table 2 below.

Table 2. Medoids Alternative

Medoid	X1	X2	X3	X4	X5	X6	X7
Lima Puluh Kota	0.3686912	0.4084507	0.6129032	0.863636	0.135135	0.220844	0.2262136
Pariaman	0.1582549	0.084507	0.1290323	0.798701	0.74343	0.521092	0.368932

Next, determine the *medoids distance* on each object with the *Euclidean Distance formula*. The following describes the calculation of the distance to one of the objects with the first *medoid* :

$$d(\text{Agam}, \text{C1}) = \sqrt{\begin{matrix}(0.337040 - 0.3686912)^2 + (0.253521 - 0.408451)^2 + \\ (0.225806 - 0.612903)^2 + (0.987013 - 0.863636)^2 + \\ (0.337838 - 0.135135)^2 + (0.339950 - 0.220844)^2 + \\ (0.343689 - 0.226214)^2\end{matrix}}$$

$$= 0.5090638$$

And so on calculating the distance with the first medoids to all objects. Next is the calculation of the distance to the second *medoid*. Here is the distance calculation on one of the objects with the second *medoid* :

$$d(\text{Agam}, \text{C2}) = \sqrt{\begin{matrix}(0.337040 - 0.1582549)^2 + (0.253521 - 0.084507)^2 + \\ (0.225806 - 0.129032)^2 + (0.987013 - 0.798701)^2 + \\ (0.337838 - 0.743243)^2 + (0.339950 - 0.521092)^2 + \\ (0.343689 - 0.368932)^2\end{matrix}}$$

$$= 0.550599$$

And so on calculating the distance with the second *medoid* to all objects. Then mark the closest distance to the value of each distance that was searched before, then calculate the total shortest distance, which is 26.932667

Next, calculate the difference between the total closeness of the 3rd iteration distance and the total closeness of the 2nd iteration of 25.049288, namely:

$S = 26.932667 - 25.049288$

$S = 1.883380$

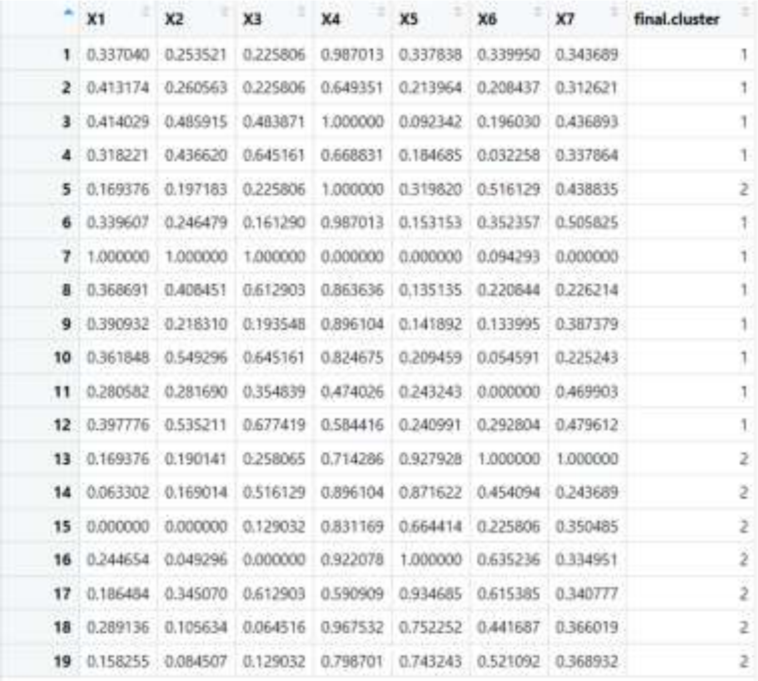
It can be seen that the result of the difference between the total closeness of the 3rd iteration distance and the total closeness of the 2nd iteration distance> 0, namely 1.883380. Therefore the grouping process was stopped, to obtain cluster results for each district/city in West Sumatra in Table 3 below :

Table 3. Clustering Kabupaten/Kota of Sumatera Barat

Clusters	Kabupaten/city
Clusters 1	Kabupaten Agam, Kabupaten Pesisir Selatan, Kabupaten Solok, Kabupaten Sijunjung, Kabupaten Padang Pariaman, Kepulauan Mentawai, Kabupaten Lima Puluh Kota, Kabupaten Pasaman, Kabupaten Solok Selatan, Kabupaten Dharmasraya and Kabupaten Pasaman Barat.
Clusters 2	Kota Padang, Kota Solok, Kabupaten Pesisir Selatan, Kota Sawahlunto, Kota Padang Panjang, Kota Bukittinggi, Kota Payakumbuh and Kota Pariaman.

(d) Clustering Results Using RStudio Software

At this stage, it displays the final results and the last step in using the RStudio software. The final result that will be displayed is in the form of grouping where the results of testing the data will show *clusters* with each member. Shown in Figure 2 below:



	X1	X2	X3	X4	X5	X6	X7	final.cluster
1	0.337040	0.253521	0.225806	0.987013	0.337838	0.339950	0.343689	1
2	0.413174	0.260563	0.225806	0.649351	0.213964	0.208437	0.312621	1
3	0.414029	0.485915	0.483871	1.000000	0.092342	0.196030	0.436893	1
4	0.318221	0.436620	0.645161	0.668831	0.184685	0.032258	0.337864	1
5	0.169376	0.197183	0.225806	1.000000	0.319820	0.516129	0.438835	2
6	0.339607	0.246479	0.161290	0.987013	0.153153	0.352357	0.505825	1
7	1.000000	1.000000	1.000000	0.000000	0.000000	0.094293	0.000000	1
8	0.368691	0.408451	0.612903	0.863636	0.135135	0.220844	0.226214	1
9	0.390932	0.218310	0.193548	0.896104	0.141892	0.133995	0.387379	1
10	0.361848	0.549296	0.645161	0.824675	0.209459	0.054591	0.225243	1
11	0.280582	0.281690	0.354839	0.474026	0.243243	0.000000	0.469903	1
12	0.397776	0.535211	0.677419	0.584416	0.240991	0.292804	0.479612	1
13	0.169376	0.190141	0.258065	0.714286	0.927928	1.000000	1.000000	2
14	0.063302	0.169014	0.516129	0.896104	0.871622	0.454094	0.243689	2
15	0.000000	0.000000	0.129032	0.831169	0.664414	0.225806	0.350485	2
16	0.244654	0.049296	0.000000	0.922078	1.000000	0.635236	0.334951	2
17	0.186484	0.345070	0.612903	0.590909	0.934685	0.615385	0.340777	2
18	0.289136	0.105634	0.064516	0.967532	0.752252	0.441687	0.366019	2
19	0.158255	0.084507	0.129032	0.798701	0.743243	0.521092	0.368932	2

Figure 2. Clustering Results with RStudio

To see the results of clustering in graphical or visual form can be seen as follows:

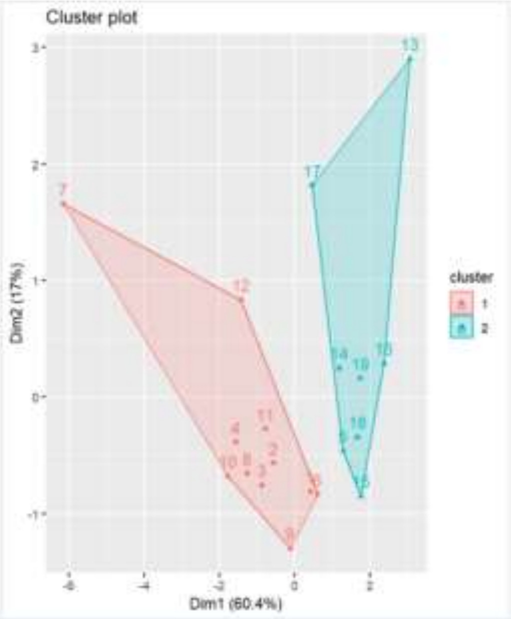


Figure 3. Cluster Plots

Based on Figure 3 it can be seen that *Cluster 1* has 11 districts/cities and *Cluster 2* has 8 districts/cities

4. Conclusion

Based on the results of the study it can be concluded that classifying poverty levels based on districts/cities in West Sumatra can be applied using the K-Medoids method which is divided into 2 clusters that are obtained by the process of determining the optimal number of K using the Sillhouette Coefficient method. The cluster results from the state that 11 districts/cities are in Cluster 1, and 8 districts/cities are in Cluster 2.

1. Cluster 1 is Kabupaten Agam, Kabupaten Pesisir Selatan, Kabupaten Solo, Kabupaten Sijunjung, Kabupaten Padang Pariaman, Kepualauan Mentawai, Kabupaten Lima Puluh Kota, Kabupaten Pasaman, Kabupaten Solok Selatan, Kabupaten Dharmasraya dan Kabupaten Pasaman Barat with the characteristics of the percentage of poor people, poverty depth index, a high poverty severity index, and the percentage of literacy rates, average years of schooling, and expected years of schooling are low compared to the rest of West Sumatra.
2. Cluster 2 is Kota Padang, Kota Solok, Kabupaten Pesisir Selatan, Kota Sawahlunto, Kota Padang Panjang, Kota Bukittinggi, Kota Payakumbuh dan Kota Pariaman with characteristics that show the percentage of poor people with a much lower poverty depth index and poverty severity index. Meanwhile, the percentage of literacy rate, the average length of schooling, and the expected length of schooling have high scores but have a high percentage of unemployment.

References

1. Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. Physics Procedia, 78(December 2015), 507–512. https://doi.org/10.1016/j.procs.2016.02.095
2. Atmaja, E. H. S. (2019). Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta. International Journal of Applied Sciences and Smart Technologies, 1(1), 33–44. https://doi.org/10.24071/ijasst.v1i1.1859
3. Balabantaray, R. C., Sarma, C., & Jha, M. (2015). Document Clustering using K-Means and K-Medoids. http://arxiv.org/abs/1502.07938
4. Berkhin, P. (2006). A survey of clustering data mining techniques. Grouping Multidimensional Data: Recent Advances in Clustering, c, 25–71. https://doi.org/10.1007/3-540-28349-8_2
5. Chakraborty, S., Islam, S. H., & Samanta, D. (2022). Introduction to Data Mining and Knowledge Discovery. In EAI/Springer Innovations in Communication and Computing. https://doi.org/10.1007/978-3-030-93088-2_1
6. Faisal, M., & Utami, W. S. (2022). Application of Data Mining Using the K-Medoids Algorithm for Poverty Index Clustering. CCIT Journal, 15(2), 272–281. https://doi.org/10.33050/ccit.v15i2.2311
7. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323. https://doi.org/10.1145/331499.331504
8. Kamber, M., & Han, J. (2018). Data Mining: Concepts and Techniques : Concepts and Techniques. In The Fundamentals of Political Science Research.
9. Qona'ah, N., Devi, A. R., & Dana, I. M. G. M. (2020). Laboratory Clustering using K-Means, K-Medoids, and Model-Based Clustering. Indonesian Journal of Applied Statistics, 3(1), 64. https://doi.org/10.13057/ijas.v3i1.40823
10. Simhachalam, B., & Ganesan, G. (2016). Performance comparison of fuzzy and non-fuzzy classification methods. Egyptian Informatics Journal, 17(2), 183–188. https://doi.org/10.1016/j.eij.2015.10.004
11. Sivarathri, S., & A, G. (2014). Experiments on Hypothesis "Fuzzy K-Means is Better than K-Means for Clustering." International Journal of Data Mining & Knowledge Management Process, 4(5), 21–34. https://doi.org/10.5121/ijdkp.2014.4502
12. Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. International Journal of Computational Intelligence Research, 13(5), 899–906. http://www.ripublication.com
13. Struyf, A., Hubert, M., & Rousseeuw, P. J. (1997). Integrating robust clustering techniques in S-PLUS. Computational Statistics and Data Analysis, 26(1), 17–37. https://doi.org/10.1016/S0167-9473(97)00020-0
14. Thinsungnoen, T., Kaoungku, N., Durongdumronchai, P., Kerdprasop, K., & Kerdprasop, N. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. 44–51. https://doi.org/10.12792/iciae2015.012
15. Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., & Li, L. (2016). Reviews on determining the number of clusters. Applied Mathematics and Information Sciences, 10(4), 1493–1512. https://doi.org/10.18576/amis/100428