Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering

Syawal Gultom[1], S. Sriadhi[2], M. Martiano[2], Janner Simarmata[2]

[1]Department of Mathematics Education, Universitas Negeri Medan, Jalan Willem Iskandar Pasar V Medan 20221, Medan - Indonesia
[2]Department of ICT Education, Universitas Negeri Medan, Jalan Willem Iskandar Pasar V Medan 20221 Medan - Indonesia

Email: sriadhi@unimed.ac.id

Abstract. This study aims to analyze the comparison of object clustering results in big data using K-Means method and K-Medoid method. Combination testing using three algorithms namely Ecludiance Distance, Canberra Distance, and Chebyshev Distance respectively for both methods. The sample of the research is the data set of students of three classes with six variables with a total of 147,679 data. The test results found that K-Means method is more optimal in data clustering than K-Medoid method, both in Ecluid Distance, Chanberra Distance and Chebyshev Distance algorithms which in overall comparison of clustering process with 1: 110.7 rationality. The results also suggested not to useChanberra Distance algorithms for both K-Means and K-Medoid methods because the cluster quality index in the Davies Bouldin test was undefined (∞). The best level of accuracy and quality of cloning is to use K-Means with Chebyshev Distance method, which is to produce five clusters with 0.1 second processing time

1. Introduction

Datamining is a set of processes for knowing and extracting information from an unknown data set, as a solution to problems including data classification [1]. The process of data classification generally uses two methods, namely K-Means and K-Medoids. Arora has conducted clustering testing using algorithms to discuss the time complexity with two variables, Item Id and Quantity, with noisy databases [2]. To calculate the distance of data in both methods are many formulas used, among others, using the algorithm Ecludional Distance, Chanberra Distance, and Chebyshev Distance.

This study aims to find the best results in the process of clustering datamining, with K-Means and K-Medoids method. K-Means and K-Medoids clustering process combined with three algorithms namely Ecludional Distance, Chanberra Distance, and Chebyshev Distance. Thus, the research will find the best pairing method and algorithm model in big data clustering for cluster quality criteria and time needed in clustering process in big data.

2. Research Methods

This study was conducted by taking sample data-set of Medan State University students for three classes with a total of 147,679 data. This research is a continuation of previous researchusing only two variables of credits and semesters, whereas in this study developed four variables namely home-status, income-parent, transportation-tool, and form-home so that there are six overall variables. The experiment uses two methods, K-Means and K-Medoids combined with the Ecludional Distance, Chanberra Distance, and Chebyshev Distance algorithms. Cluster quality tests use the Davies Bouldin criterion which takes into account the number of caster and the time required during clustering.

There are two methods and three algorithms used in the data clustering in this study, as described below.

2.1 K-Means Algorithm

K-Means Algorithm is based on partitioning methods for clustering tasks, especially in low-dimensional datasets[3]. In the process of clustering K-means known unsupervised learning because the label group is not known. K-Means can separate similar abstract objects into groups and form other groups for unlike data [4]. The process of clashing with the K-Means algorithm [5] is as follows:

a. Determine the number of clusters formed into the data set and the dataset as the value to be established.
b. Determine the first K value by specifying the value of k in the data set or specifying the value of k at random.
c. Calculate the distance of each object to the center of the centroid by using the Euclidience distance formula until it finds the closest distance from centroid center.

$$d = \left(\sum_{i=1}^{n}(x_i - \mu_i)^2\right)^{\frac{1}{2}} \quad (1)$$

Where :
: distance between cluster x and center cluster μ to object to x_i
i : cluster size to search distance,
μ :the size of the i-cluster center

d. Repeat from step two to step five so that cluster members are not changed.

2.2 K-Medoid

K-medoid is a classical partitioning technique of clustering data by determining the number of objects and clusters. This algorithm minimizes characteristic differences[6]. The selected object may represent a cluster called the medoid, and the cluster is formed from the proximity calculations between the medoid and the object. The process of K-Medoids method [7] is as follows:

a. Calculate the distance of each pair of objects based on the unequal selection
b. Calculating v_j to object *j* with equation:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} j = 1....n \quad (2)$$

c. Sort v_j from largest to smallest and select the object *k* that has the smallest k value as the initial medoid.
d. Produce the result of the initial cluster by determining to the nearest object of the medoid
e. Find a new medoid from each cluster by minimizing the total distance in its cluster and update the medoid of each cluster by replacing it with a new medoid.
f. Set each object to the nearest medoid and get cluster results.
g. Calculate the sum distance of all objects to their medoids.

2.3 Ecludiance Distance

Ecludiance distance is a calculation (*p*) 2 or 3 points based on the distance and angle associated with Phytagoras theory. The calculation of distance in this study using 3 variables [8]with the formula:

$$\mathrm{d} = \sqrt{\sum_{i=1}^{v}(p_{1i} - p_{2i})^2} \tag{3}$$

The difference of the 3 variables that counted the proximity of the object is p_1-p_2, p_3-$p_1$and p_2-p_3.

2.4 Chanberra Distance

Chanberra distance is a numerical calculation based on the distance between pairs of points in a vector space by formula:

$$\mathrm{d}(x,y)=\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i|+|y_i|} \tag{4}$$

Where:d is the distance between the vectors x, y and$x$ =x_i,..,x_n; y =y_i,..,y_n

2.5 Chebyshev Distance

Chebyshev Distance is known as the maximum matrix known as the chessboard spacing. Ponnmoli and Selvamuthukumaran [9] determine the vectors x and y with the coordinate standards x_i and y_i calculated using the formula:

$$d_{(x,y)} = max_{i=1,2,..n}|x_i - y_i| \tag{5}$$

2.6 Performance Measurement

For cluster measurement, the method of Davies Bouldin was used with criteria based on the value of data attachment to the centroid of the cluster followed (intra-cluster) and the distance between the centroid of the cluster (inter-cluster). In determining the intra cluster used the formula:

$$intra\ cluster_j = \frac{1}{N}\sum_{p_i=j} d(c_j, x_i)^2 \tag{6}$$

The inter cluster distance of cluster *j* and *k* is calculated as the distance between the centroids c_j and x_i. A good cluster should have a low intra-cluster value and an inter-cluster value as large as possible. The next step is to calculate the ratio value to the formula:

$$R_{j,k} = \frac{intra\ Cluster_j + intra\ Cluster_k}{|c_j - c_k|} \tag{7}$$

Once the ratio value is found, the Davies Bouldin value can be determined with the following formula:

$$\mathrm{DB}=\frac{1}{M}\sum_{I=1}^{M} R_I \tag{8}$$

3. Results and Discussion

The result of the research as presented below is obtained through experiment using K-Means method and K-Medoids method, with combination of three algorithms namely Ecluid Distance, Chanberra Distance, and Chelbyshev Distance.

3.1. Algoritma Ecluid Distance

In testing with the Ecluid Distance algorithm, before the experiment was first selected the cluster's central point was randomly assigned and the group determined 5 clusters, with a maximum of 10 iterations. Test results using the two methods are presented in the following figure.

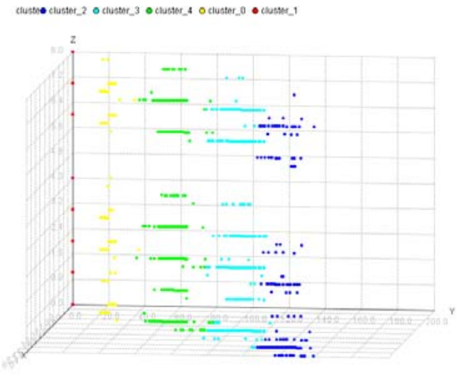

Figure 1.K-Means Ecluid Distance

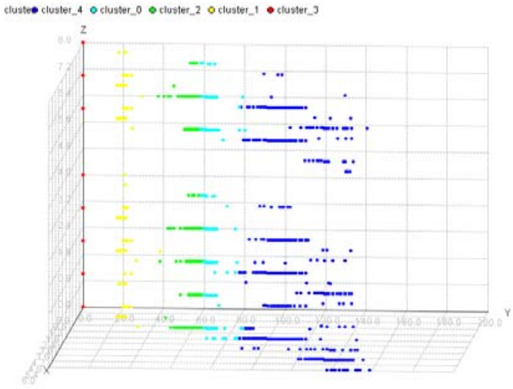

Figure 2.K- MedoidEcluid Distance

Furthermore, the result of clustering data objects using K-Means and K-Medoids method with Ecluid distance algorithm is presented in table 1.

Table 1.K-Means clustering and K-Medoids with Ecluid distance on iteration 10

No	Method	Clusters					Time
		(1)	(2)	(3)	(4)	(5)	
1	K-MeanswithEcluid Distance	1025	2982	450	2314	2988	0.00.01
2	K-MedoidwithEcluid Distance	1660	1024	1323	2982	2770	0.12.39

Table 1 shows that the time for clustering with an Ecluid distance algorithm, the K-Means method (0.1 seconds) is much shorter than the K-Medoids method (12 minutes and 39 seconds). Experimental results prove that clustering with Ecluid distance algorithm is more effective using K-Means method than K-Medoids method with 1: 126.5 time comparison ratio. This is in line with previous research results [10][11].

3.2. Algoritma Chanberra Distance

The second test was performed by comparing K-Means method and K-Medoids method using Chanberra Distance algorithm. The experimental results show that the K-Means method of distance between point pairs and the center of the cluster produces data objects into three groups, i.e. clusters 1, 2, and 4 with full distribution to the clusters it occupies. This is different from the method of K-Medoids which has a smaller cluster that is 2 just cluster. Test results with Chanberra distance algorithm are presented in the following figure.

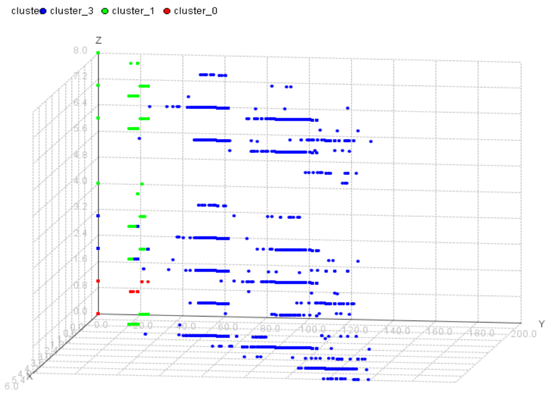

Figure 3. K-Means Chanberra Distance

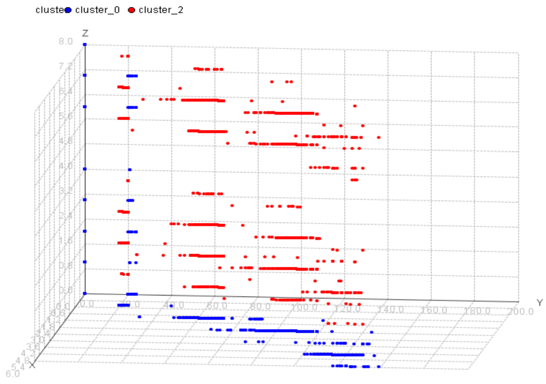

Figure 4. K-Medoid Chanberra Distance

For clustering using the Ecluid Distance algorithm, the results of clustering of data objects in the K-Means and K-Medoid methods are presented in Table 2.

Table 2.K-Means and K-Medoid Clustering with Chanberra Distance on Iterations 10

No	Method	Clusters					Time
		(1)	(2)	(3)	(4)	(5)	
1	K-Means with Chanberra Distance	1392	1866	0	6501	0	0.00.02
2	K- Medoidwith Chanberra Distance	5917	0	3842	0	0	0.15.14

Test results with Chanberra Distance algorithm resulted in three clusters with 0.2 second time on K-Means method, while K-Medoids method produced two clusters with time of 15 minutes and 14 seconds. The comparison of clustering time between the K-Means method and K-Medoid is 1: 76.10. The results of this study reinforce previous research findings[4],[7].

3.3. Algoritma Chebyshev Distance

The last test with the Chebyshev Distance algorithm uses K-Mean method and K-Medoids method. The experimental results in both methods have the ideal difference of the coordinate pairs, so that the distribution of the object meets the specified cluster.

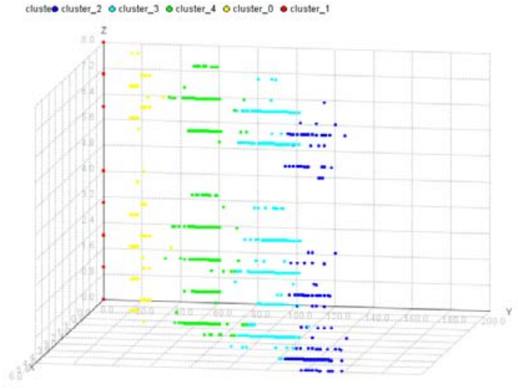

Figure 5.K-Means with Chebyshev Distance

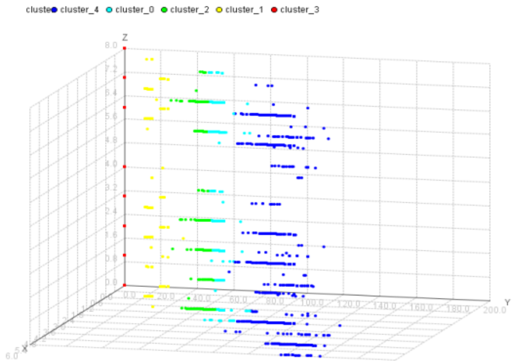

Figure 6.K-Medoid with Chebyshev Distance

The test results with Chebysheva distance algorithm for clustering of data objects using K-Means method and K-Medoid method are presented in table 3.

Table 3. K-Means and K-Medoid clustering with Chebysheva distance on iteration 10

No	Method	Clusters					Time
		(1)	(2)	(3)	(4)	(5)	
1	K-Means with Chebyshev Distance	1025	2982	450	2314	2988	0.00.01
2	K- Medoidwith Chebyshev Distance	1398	1024	1591	2982	2764	0.12.57

Based on table 3, the Chebysheva distance algorithm produced five clusters with a time of 0.1 second on the K-Means method, while the K-Medoids method produced five clusters within 12 minutes and 57 seconds. Thus, the K-Means Method is more efficient in clustering large data than the K-Medoids method with the Chebysheva distance algorithm and a time ratio of 1: 129.5. Overall, the results of clustering with the Ecluid distance, Chanberra distance, and Chebyshev distance algorithms prove K-Means method is more optimal than K-Medoid method in data clustering process with a very large time ratio of 1: 110.7. The results of this study enrich the research findings and are the development of previous research [5]–[7].

Furthermore, to determine the quality of the clusters formed by Davies Bouldin test. The near-zero index is identified as a good cluster. The results of the Davies Bouldin test calculation using the two methods are presented in Table 4.

Table 4. Indeks Davies Boldin withK-Means and K-Medoids Method

No	Method	Davies Bouldin	Time (seconds)
1	K-Means Ecluid distance	0.3111	1
2	K-Means Chanberra Distance	∞	2
3	K-Means Chebyshev Distance	0.3111	1
4	K-Medoid Ecluid distance	0.933	759
5	K- Medoid Chanberra Distance	∞	914
6	K- Medoid Chebyshev Distance	0.955	777

Table 4 shows that the K-Means method has a smaller Davies Bouldinindex than the K-Medoids method. However, on the distance calculation of the Chanberra Distance algorithm test results in both methods are undefined (∞) so that clustering large data measurement with Chanberra Distance is not recommended.

4. Conclussion

The results of experimental data clustering using K-Means method is more optimal than K-Medoids method, both from the number of clusters produced and the time required, both in the Ecluid Distance, Chanberra Distance and Chebyshev Distance algorithms. The comparison of the clustering time between K-Means method and K-Medoid is 1: 110.7. However, the Chanberra Distance algorithm is not recommended for bigdata clustering, since the resulting value for the Davies Bouldin test is very large and undefined (∞). The best level of accuracy and quality of cloning is to use the K-Means with Chebyshev Distance method, which is to produce five clusters with a 0.1 second processing time.

5. References

[1] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*. Elsevier Science, 2011.

[2] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016.

[3] A. Bansal, M. Sharma, and S. Goel, "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining," *Int. J. Comput. Appl.*, vol. 157, no. 6, pp. 35–40, Jan. 2017.

[4] K. Kouser and Sunita, "A comparative study of K Means Algorithm by Different Distance Measures," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 1, no. 9, pp. 2443–2447, 2013.

[5] jelili O. oyelade, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *Arxiv Prepr. arXiv …*.

[6] R. Pratap, K. Suvarna, J. Rama, and D. . Nageswara, "An Efficient Density based Improved K- Medoids Clustering algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, 2011.

[7] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009.

[8] Pbarrett.net, "Euclidean Distance Whitepaper," *Tech. Whitepaper Ser. 6*, 2005.

[9] K. M. Ponnmoli, "Analysis of Face Recognition using Manhattan Distance Algorithm with Image Segmentation," vol. 3, no. 7, pp. 18–27, 2014.

[10] C. C. Aggarwal and C. K. Reddy, *Data clustering : algorithms and applications*. .

[11] S. Sriadhi, R. Rahim, and A. S. Ahmar, "RC4 Algorithm Visualization for Cryptography Education," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, p. 012057, Jun. 2018.