Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

The Influence of Optimization of the k-Means Algorithm with Genetic Algorithm on the Results of High Dimension Data Clustering

Yulinda Ramadhana[1], Muhammad Ihsan Jambak[2*]
yulinda.ramadhana98@gmail.com[1], jambak@unsri.ac.id[2]
[1]Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya
[2]Departemt of Informatics Management, Faculty of Computer Science, Universitas Sriwijaya

Article Information

Submitted : 25 Dec 2023
Reviewed: 19 Jan 2024
Accepted : 15 Feb 2024

Keywords

Dimensional Reduction,
Feature Selection,
Singular Value Decomposition,
k-Means,
Genetics Algorithm

Abstract

Clustering k-means begin with the random initial determination of the cen-troid. Initially generated random centroids often cause k-means to be trapped in the optimum local solution, which results in poor clustering quality. Therefore, this study examined the effect of genetic algorithms in determin-ing initial centroids in k-means. Clustering k-means with random initial cen-troids and with initial centroids from genetic algorithm calculations are each tested on the data with dimension reduction and without dimension reduc-tion. Based on the results of the initial centroid testing obtained from genet-ic algorithms, the quality of cluster results increased by 54.9% in the high dimensional data and 52.4% in the data that had been carried out for the di-mensional reduction. This result shows that the k-means clustering with ini-tial centroids obtained from genetic algorithm calculations has the best clus-ter/solution results with significant results.

A. Introduction

Clustering is the process of grouping objects that have similarities into one cluster or group and different objects into other groups [1, 2]. K-means is one of the most popular clustering algorithms because it is a simple one, making it easier to understand [3-5]. However, there is also a weakness of the k-means algorithm that is often trapped in the curse of dimensionality on high-dimensional data, so to overcome it, dimension reduction can be done [6].

The K-Means clustering results are significantly dependent on the initial centroid selection, indicating the K-Means algorithm's sensitivity to the centroid's beginning value, which can have a significant impact on the cluster's ultimate outcome. The initial centroid of the k-means generated randomly is likely to have an impact on the position of the adjacent initial centroid, this can cause k-means to get often stuck on the optimum local solution [4, 7]. Different initial centroid selections can produce different cluster results [2, 4, 8, 9]. However, this does not imply that the initial centroid cannot be established. In terms of determining the original centroid, such as the CCIA method (Cluster Center Initialization Algorithm), coefficient variation and correlation, or other methods [5, 10]. Calculations for determining initial centroids using genetic algorithms are used to overcome the problems that have been mentioned.

Genetic algorithm is an optimization algorithm that can do a global search to find solutions to optimization problems by getting the optimal solution to a problem that has many possibilities [11-14]. Each k-means clustering method with both random initial centroids and initial centroids obtained from genetic algorithms was tested on high dimensional data which was done in dimension reduction and without dimen-sion reduction. Therefore, this study focuses on determining the effect of initial centroid determination using genetic algorithms in grouping high-dimensional data.

B. Research Method

The research test data used was the Indonesian language journal which was down-loaded through the garuda.ristekdikti.go.id site with 100 text documents. Then the documents were copied into a file with the extension of .txt and each file stored in the same folder. Text data then converted to numeric data through the preprocessing stage. The preprocessing stage is case folding, tokenizing, stop words removal, and stemming.

The next step is the weighting process using tf-idf. After the weighting process was carried out, the weighted data were used for the k-Means clustering process. This k-Means clustering process uses data from dimension reduction and without dimension reduction. K-Means clustering with dimension reduction data results using weighted data followed by the dimension reduction process using the SVD (Singular Value Decomposition) method, followed by the k-means clustering pro-cess while the k-means clustering without the dimension reduction process uses weighted data which is directly carried out by the clustering process both with ran-dom initial centroids and initial centroids obtained from genetic algorithms.

The process of determining the value of k (number of clusters) done before the process of clustering k-means. The number of clusters (k) used for the k-means

clus-tering process with random initial centroids and initial centroids is determined using genetic algorithms namely clustering k-means with random initial centroids that have the best DBI values based on the test results of 10 trials with the value of k=2 to k=10.

The optimum k value is obtained from the value of k which has decreased significantly based on its DBI value, to find out the value of k which has decreased significantly using the elbow method illustrated in Figure 1. Determination of the optimum k uses the elbow method, which is a method to determine the most optimum number of clusters by looking at the most significant change in value in comparison with the number of other clusters. Based on this curve, it can be concluded that the optimum k value is at k = 6, with an average DBI value of 7.6225949, which is indicated by the sharpest angle on the curve. So the value of k = 6 is the essential number of clusters of the data used.

After obtaining the number of clusters, then calculate each distance between the data to the initial centroid that has been determined in advance using the Euclidean distance formula [15]. Each data grouped into its cluster based on the closest distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \qquad (1)$$

Then determine the new centroid value by using the average formula from the data in the same cluster. Next, calculate the minimum distance between new centroid data that has been obtained using the Euclidean distance formula in equation (1). Determine the new centroid and recalculate the minimum distance until the converg-ing conditions are met. Convergence is a condition where the data obtained in each cluster in the next iteration is the same as the previous iteration [16].

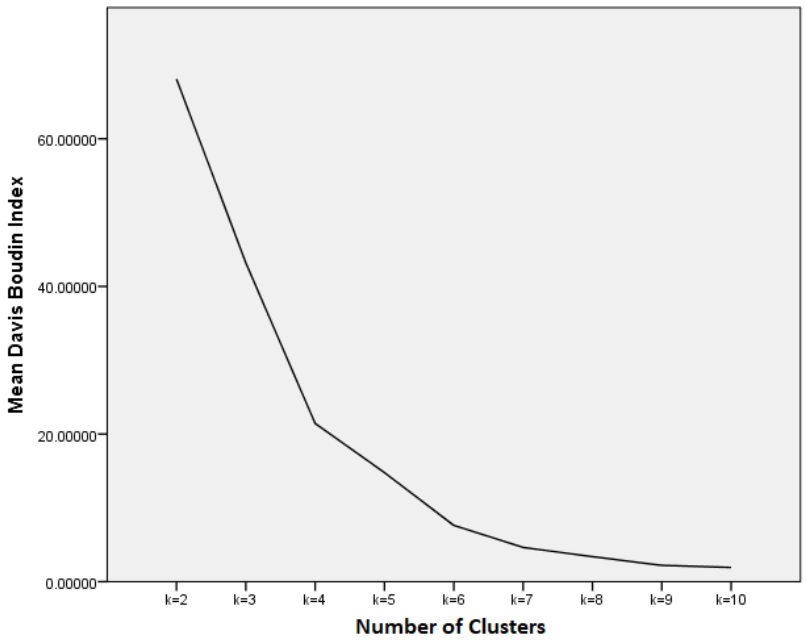

Figure 1. The Curve of Optimal k Value Determination.

The k-means clustering process is carried out using initial centroids obtained from genetic algorithms which are then compared to the results of k-means clustering with random initial centroids. The input values needed for this genetic algorithm are the number of clusters, pop sizes, the maximum number of iterations, cr (crossover rate), mr (mutation rate) [13, 17].

The first step is to generate as many chromosomes or populations as the popula-tion. For each chromosome then generate a centroid value at random as many as the number of clusters, in this case, the number of clusters used is 6 clusters. Then cal-culate the fitness value on each chromosome, the higher the fitness value, the more likely it is to be selected as a parent [18]. For each chromosome do the clustering data process based on minimum distance using the formula from equation (1). Then, do the update centroids process using the average formula from the data in the same cluster. After that, calculate the fitness value using the equation:

$$F = \frac{1}{J} \tag{2}$$

F is the fitness value, and J is the minimum distance, as described in equation (1). Then do the selection process using the rank selection method, i.e., chromosomes are sorted based on the most substantial fitness value. In each chromosome, the random value is compared then this random value is compared with the value of cr (crossover rate). If the random value is less than cr, then the chromosome is selected as the parent. After the parent is obtained, a crossover process was carried out using the Single Point Crossover method to generate new individuals (offspring). Then do the mutation process using the uniform mutation method. Then the new individual ob-tained from the mutation process is then re-inserted into the population. Perform the same process again, starting from calculating the fitness value of each chromosome until the mutation process until the iteration reaches the maximum number of iterations [11]. Furthermore, to choose the best centroid that was done, the elitism meth-od is used, i.e., the new individuals obtained are sorted according to the best fitness value. This individual with the best fitness value is used as the initial centroid in the clustering k-means process [19].

Random initial centroids and initial centroids obtained from genetic algorithms were tested 30 times each using dimension reduction data or without dimension reduction which then recorded its DBI value, the number of iterations and its computational time in each test. Based on the test results, a data processing test is then performed. The steps taken are the normality test and data homogeneity test, in this case, the Kolmogorov-Smirnov test is used for the data normality test. Based on the results of the normality test data, the results obtained that the data used does not meet the norms of normality, therefore conducted parametric tests using Kruskal-Wallis H test and Mann-Whitney test to determine whether there are significant differences from the methods tested.

C. Result and Discussion

It found that the optimum k value is k=6. Therefore, the number of clusters used in the k-means clustering method in this study is 6 clusters. Before testing the

clustering method, the genetic algorithm parameter is tested first, namely testing 5 times iteration by entering 100, 150, 200, 250, 300. Testing the value of cr (crossover rate) 5 times by entering 0.9, 0.8, 0.7 , 0.6, 0.5. Test mr (mutation rate) 5 times by entering 0.1, 0.2, 0.3, 0.4. Based on the results of genetic algorithm testing that has been done, the most optimal genetic algorithm parameters are obtained based on the larg-est fitness value, namely the number of iterations = 250 iterations, pop size = 15, mutation rate (mr) = 0.1 if tested on data reduction or without dimension reduction, and crossover rate (cr) = 0.9 in dimension reduction data and 0.8 in data without dimension reduction.

After obtaining the optimum number of clusters and genetic algorithm parameters, further testing of the clustering method with random initial centroids and initial centroids obtained from genetic algorithms is each tested on the result of dimension reduction data and data without dimension reduction. The testing results of the clus-tering method can be seen in table 1.

Based on the results of the test that have been done it can be seen that the k-means clustering with initial centroids obtained from genetic algorithms has decreased DBI values and the number of iterations compared to k-means with random initial centroids both on high and low dimensional data, as seen in Figure 2. On the other hand, clustering k-means with the initial centroid of the genetic algorithm, has a longer overall computational time, and this is due to an increase in the computa-tional time of the genetic algorithm. When compared between k-means with initial centroids using genetic algorithms with dimension reduction data and without di-mension reduction, k-Means clustering tested on dimensional reduction data experienced a decrease in cluster quality and the number of iterations compared to k-means clustering using data without dimension reduction. The decrease also occurs when computing the clustering k-means using dimension reduction data with the initial centroid of genetic algorithms.

Table 1. Comparison Result of Clustering Method

Evaluation Items	Random Centroid		Genetic Algorithm Centroid	
	Without Dimension Reduction	With Dimension Reduction	Without Dimension Reduction	With Dimension Reduction
	k-Means	SVD + k-Means	k-Means	SVD + k-Means
DBI	7.68528	7.33021	3.46538	3.4833
Iteration	13	10	2	1
Time (sec)	65.241	286	27682.785	1486.837

Analysis of the results of k-means clustering with random centroids on data performed dimension reduction using SVD and data without dimensional reduction, based on the result of the data processing DBI, an insignificant decrease of 7.68528 to 7.33021. Then, based on time computing, there was a significant increase in the total time from 65241 nanoseconds to 286 seconds. The number of iterations decreased significantly, from 13 to 10. Thus obtained findings that k-means clustering with random centroids early on data performed dimensional reduction and without dimension reduction had no significant differences in cluster quality.

However, there is a significant decrease in the number of iterations produced, indicating that the k-means clustering with the initial random centroid and using dimensional reduction data has a faster computing time, if only viewed in terms of the computational time of the k-means clustering process and faster convergence conditions. However, in terms of computing time, the entire process begins with dimensional reduction, followed by clustering k-means with the initial random centroid and employing data reduction, which takes longer overall.

Analysis of k-means clustering results with early centroids of calculations of genetic algorithms if tested on data performed dimension reduction using SVD and data not performed dimensional reduction, based on data behavior of value DBI versus k-means Clustering with the initial centroid of the calculation of the genetical algorytm using the data carried out the process of dimensional reduktion reduction and data without dimensional diminution resulted in an insignificant decrease, i.e., an average value of DBI of 3.46538 to 3.48330. In addition, the number of iterations has lowered dramatically, from two to one. However, the data processing of time computing has resulted in a considerable decrease of 27682.785 seconds to 1486.837 seconds.

As a result, k-means clustering with early centroids derived using a genetic algorithm on data from a previous dimensional reduction procedure requires less overall processing time than a non-dimensional processing technique. Based on this, it is possible to deduce that the data acquired for the dimensional reduction process can help speed up the early centroid calculation process using genetic algorithms while maintaining the same cluster quality results.

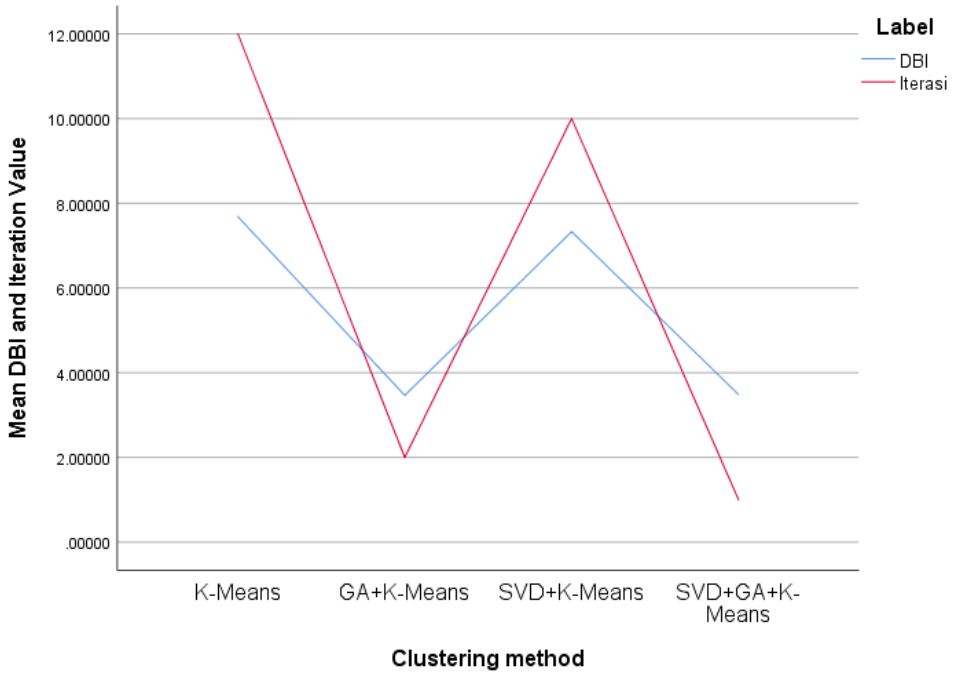

Figure 2. Comparison of DBI Value and Iteration Between K-means Clustering Method

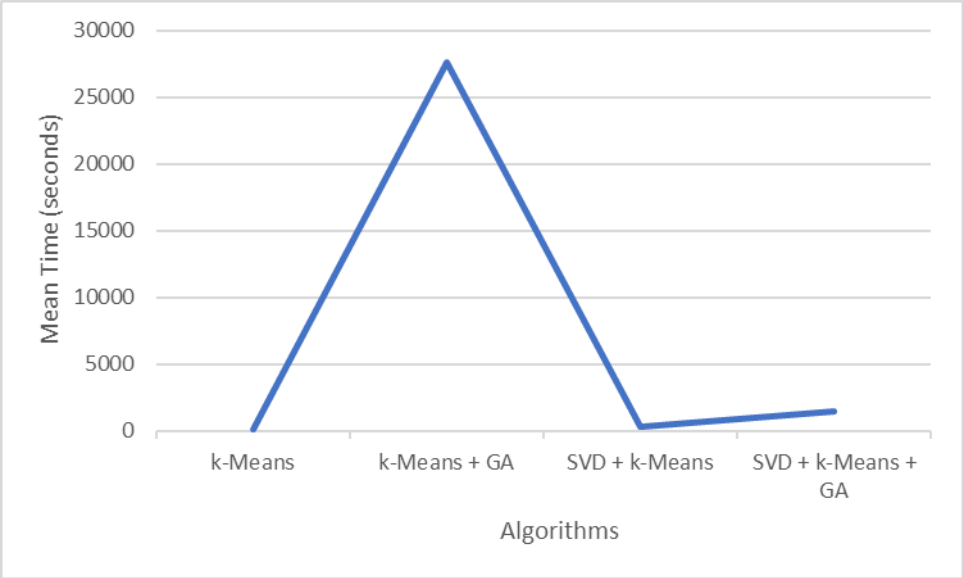

Figure 3. Comparison of Computational Time between K-Means Clustering Method

It can be concluded that clustering k-means with initial centroids from genetic algorithm calculations have better cluster quality with a significant difference in comparison. The initial centroid obtained from genetic algorithm calculation can also accelerate k-means in achieving the convergence condition, this can be seen in the smaller number of iterations based on the results of the k-means clustering test with the initial centroid of the genetic algorithm that has been done.

In addition, k-means clustering with initial centroids obtained from genetic algorithm calculations on data carried out in the previous dimension reduction process has a faster overall processing time when compared to k-means clustering processes with initial centroids obtained from genetic algorithms and using data which does not do the process of dimension reduction, this means that the process of dimension reduction can help speed up the process of clustering k-means with initial centroids obtained from genetic algorithms.

D. Conclusion

Based on the results of testing and analysis of k-means clustering methods with random initial centroids and initial centroids obtained from calculations of genetic algorithms tested in Indonesian-language journal documents, it can be concluded that it is necessary to do the determination of the initial centroid at k-means using genetic algorithms. This can be seen in the test results, both using the data performed for dimensional reduction as well as without the reduction of the initially determined centroid. Using the genetic algorithm can improve the quality of the cluster based on the internal evaluation results of DBI and is able to speed up the process of clustering k-means in achieving convergence conditions, i.e., a smaller number of iterations compared to the clusters of k-means with random centroid beginnings. However, k-means clustering with early centroids determined using genetic algorithms has a longer computation time due to the addition of computational time to the methods of genetic algorithms.

In research that focuses on the effect of determining the initial centroid of k-means using genetic algorithms in addition to the effect of dimensional reduction on high-dimensional data, it can be concluded that the determination of initial centroids using genetic algorithms can improve the quality of cluster results and speed up the process of achieving convergent conditions on k-means if compared to k-Means with random initial centroids. This better result can be seen from the percentage change of 54.9% in the data without dimension reduction and 52.4% in the data done in dimension reduction. However, k-means with initial centroids obtained from genetic algorithms have a long overall computational time due to an increase in the time of the initial centroid calculation process using genetic algorithms, both when tested on dimensional reduction data or without dimensional reduction. In the future study, we expect that the proposed method can be produced faster computational time with a better quality of cluster results.

E. References

[1] R. K. Mishra, K. Saini, and S. Bagri, "Text document clustering on the basis of inter passage approach by using K-means," in *International Conference on Computing, Communication & Automation*, 2015: IEEE, pp. 110-113.

[2] J. Yadav and M. Sharma, "A Review of K-mean Algorithm," *Int. J. Eng. Trends Technol,* vol. 4, no. 7, pp. 2972-2976, 2013.

[3] J. Zade, D. Bamnote, and P. Agrawal, "Text document clustering using K-Means algorithm with its analysis and implementation," *Imperial Journal of Interdisciplinary Research,* vol. 3, no. 2, pp. 1528-1531, 2017.

[4] T. Badriyah, "Hybrid Modelling KMeans-Genetic Algorithms in the Health Care Data," *EMITTER International Journal of Engineering Technology,* vol. 2, no. 1, pp. 63-74, 2014.

[5] P. Beena, S. Kumar, and K. Balachandran, "K-Means Clustering—Review of various methods for initial selection of centroids," *International Journal of Scientific & Engineering Research,* vol. 4, no. 8, pp. 1844-1847, 2013.

[6] M. I. Jambak, F. Mohammed, N. Hidayati, R. Efendi, and R. Primartha, "The Impacts of Singular Value Decomposition Algorithm Toward Indonesian Language Text Documents Clustering," in *International Conference of Reliable Information and Communication Technology*, 2018: Springer, pp. 173-183.

[7] N. Arora and M. Motwani, "Optimizing K-Means by fixing initial cluster centers," *Int. J. Curr. Eng. Technol,* vol. 4, no. 3, pp. 2101-2107, 2014.

[8] A. M. Baswade and P. S. Nalwade, "Selection of initial centroids for k-means algorithm," *IJCSMC,* vol. 2, no. 7, pp. 161-164, 2013.

[9] G. N. W. Paramartha, D. E. Ratnawati, and A. W. Widodo, "Analisis Perbandingan Metode K-Means Dengan Improved Semi-Supervised K-Means Pada Data Indeks Pembangunan Manusia (IPM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 1, no. 9, pp. 813-824, 2017.

[10] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern recognition letters,* vol. 25, no. 11, pp. 1293-1302, 2004.

[11] R. Dash and R. Dash, "Comparative analysis of k-means and genetic algorithm based data clustering," *International Journal of Advanced Computer and Mathematical Sciences,* vol. 3, no. 2, pp. 257-265, 2012.

[12] M. Kaushik and B. Mathur, "Comparative study of K-means and hierarchical clustering techniques," *Int. J. Softw. Hardw. Res. Eng,* vol. 2, no. 6, pp. 93-98, 2014.

[13] B. K. Khotimah, F. Irhamni, and T. Sundarwati, "A Genetic algorithm for optimized initial centers K-means clustering in SMEs," *Journal of Theoretical and Applied Information Technology,* vol. 90, no. 1, p. 23, 2016.

[14] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and S. Zeebaree, "Combination of K-means clustering with Genetic Algorithm: A review," *International Journal of Applied Engineering Research,* vol. 12, no. 24, pp. 14238-14245, 2017.

[15] Z. S. Younus *et al.*, "Content-based image retrieval using PSO and k-means clustering algorithm," *Arabian Journal of Geosciences,* vol. 8, pp. 6211-6224, 2015.

[16] M. Kaur and U. Kaur, "Comparison between K-mean and hierarchical algorithm using query redirection," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 7, pp. 1454-1459, 2013.

[17] P. Dhanya, M. Jathavedan, and A. Sreekumar, "Implementation of text clustering using genetic algorithm," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 5, pp. 6138-6142, 2014.

[18] N. U. Roiha, Y. K. Suprapto, and A. D. Wibawa, "The optimization of the weblog central cluster using the genetic K-means algorithm," in *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 2016: IEEE, pp. 278-284.

[19] W. Lesmawati, A. Rahmi, and W. F. Mahmudy, "Optimization of frozen food distribution using genetic algorithms," *Journal of Environmental Engineering and Sustainable Technology,* vol. 3, no. 1, pp. 51-58, 2016.