Initial Centroid Determination Using Genetic Algorithm in Data Clustering

Rizki Kurniati[1], Osvari Arsalan[2], Yulinda Ramadhana[3]
Informatics Engineering, Universitas Sriwijaya
Palembang, Indonesia
Email: rizkikurniati@ilkom.unsri.ac.id, osvariarsalan@ilkom.unsri.ac.id, yulinda.ramadhana98@gmail.com

***Abstract*—Clustering K-Means using random initial determination centroid. Generated random centroids using K-Means trapped in optimum local which results in poor clustering quality. Initial centroids in k-means will examine effect of genetic algorithms are each tested on data with dimension reduction and without dimension reduction. Based on the results of initial centroid testing obtained from genetic algorithms, quality of cluster results increase 54.9% in high dimensional data and 52.4% in data had been carried out for dimensional reduction. This shows that K-Means clustering with initial centroids obtained from genetic algorithm calculations has best cluster with significant results.**

Keywords— Clustering, Genetic Algorithm, K-Means, Principal Component Analysis, Particle Swarm Optimization

I. INTRODUCTION

Clustering is process grouping objects that have similarities into one cluster or group and different objects into other groups [1], [2]. K-Means algorithm in data mining of randomly selected centroids, which are used beginning points for every cluster, calculations to optimize positions of the centroids. Define a target number k, which refers to number of centroids in dataset. Initial centroid of K-means generated randomly have an impact on position initial centroid, this can cause K-Means get stuck on optimum local solution [3], [4]. Calculations for determining initial centroids using genetic algorithms are used to overcome problems that have been mentioned. Genetic algorithm is an optimization algorithm that can do a global search to find solutions optimization problems by getting optimal solution to a problem that has many possibilities [5], [6]. Each k-means clustering method with both random initial centroids and initial centroids obtained from genetic algorithms tested on high dimensional data which was done in dimension reduction and without dimension reduction. Therefore, this research focuses to determine effect initial centroid determination using genetic algorithms in grouping high-dimensional data.

II. METHODOLOGY

A. Description Dataset

Data used Indonesian language journal which was downloaded through garuda.ristekdikti.go.id with 100 text documents. Then contents of document are copied into a file with extension *.txt* and each file is stored in same folder. Text data converted to numeric data through preprocessing stage. Preprocessing stage is case folding, tokenizing, stopwords removal, and stemming. Next step is weighting process using *tf-idf*. After weighting process is carried out, weighted data is used for 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 dimension reduction process using the SVD (Singular Value Decomposition) method while k-means clustering without the dimension reduction process uses weighted data which is directly carried out by clustering process both with random initial centroids and initial centroids obtained from genetic algorithms.

B. Clustering K-Means

Process of determining value k (number of clusters) is first done before process of clustering k-means. The number of clusters (k) used for k-means clustering process with random initial centroids and initial centroids is determined using genetic algorithms namely clustering k-means with random initial centroids that have best DBI values based on test results of 10 trials with the value of k=2 to k=10. The optimum k value is obtained from value of k which has decreased significantly based on dbi value, to find out value of k which has decreased significantly using elbow method illustrated in Figure 1.
Determination of optimum k uses elbow method, which is method to determine most optimum number of clusters looking at the greatest change in value in comparison with number of other clusters.

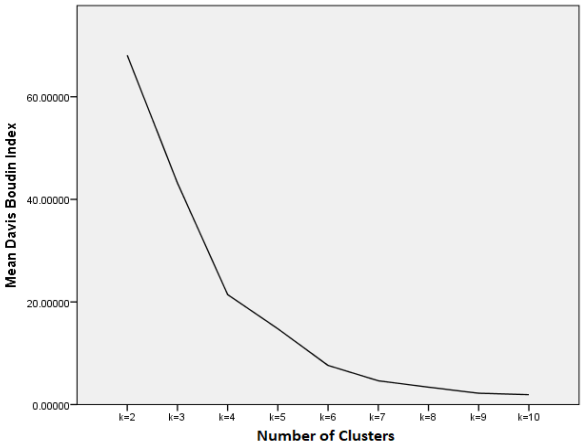

Figure 1. Curve Result of Optimal k Value Determination

Based on this curve, it can be concluded that optimum k value is at k = 6, with an average DBI value of 7.6225949,

which is indicated sharpest angle on the curve. So, value of k = 6 is the most essential number of clusters data used.

After obtaining number of clusters, then calculate each distance between data to initial centroid that has been determined in advance using euclidean distance formula [7]. Each data will be grouped into cluster based on closest distance.

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

Then determine 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 euclidean distance formula in equation (1). Determine new centroid and recalculate minimum distance until converging conditions are met. Convergence is a condition where data obtained in each cluster in next iteration is same as previous iteration [8].

C. *Determination of the Initial Centroid Value of k-means*

The k-means clustering process is carried out using initial centroids obtained from genetic algorithms which are then compared to results of k-means clustering with random initial centroids. Input values needed for this genetic algorithm are number of clusters, pop sizes, maximum number of iterations, cr (crossover rate), mr (mutation rate) [6], [9]. First step is to generate as many chromosomes or populations as population. For each chromosome then generate a centroid value at random as many as number of clusters, in this case, number of clusters used is 6 clusters. Then calculate fitness value on each chromosome, the greater the fitness value,more likely it is to be selected as a parent [10]. For each chromosome do clustering data process based on minimum distance using formula from equation (1). Then, do update centroids process using average formula from data in same cluster. After that, calculate fitness value.

$$F = \frac{1}{J} \quad (2)$$

F is the fitness value and *J* is minimum distance as described in equation (1). Then do selection process using the rank selection method, chromosomes are sorted based on the largest fitness value. In each chromosome, random value is compared then this random value is compared with the value of cr (crossover rate). If random value is less than cr, then the chromosome is selected as the parent. After the parent is obtained, a crossover process will be carried out using the Single Point Crossover method to generate new individuals (offspring). Then do mutation process using the Uniform mutation method. Then the new individual obtained 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 is done, elitism method is used, i.e 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 clustering k-means process [12].

D. *Clustering Testing Method*

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. Steps taken are the normality test and data homogeneity test, in this case, Kolmogorov-Smirnov test is used for data normality test. Based on 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 Kruskall-Wallis H and Mann-Whitney test to determine whether there are significant differences from methods tested.

III. Result and Discussion

Based on the results of testing the optimum k value carried out in section *2.2*. its 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 largest 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.

A. *Test Result of Clustering Method*

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 clustering method can be seen in table 1.

Table 1. Comparison Result of Clustering Method

	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

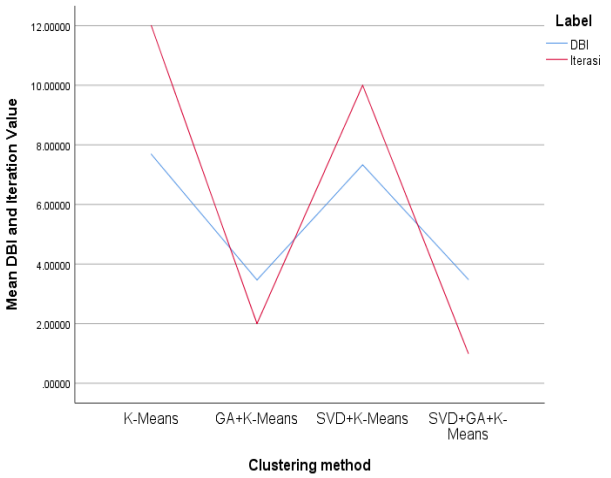

Figure 2. Comparison Graph of DBI Value and Iteration Between K-means Clustering Method

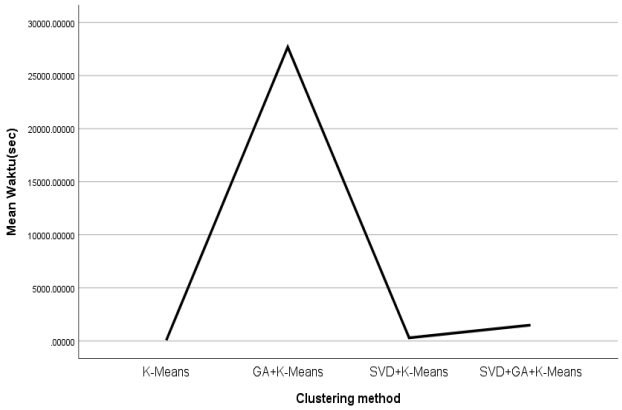

Figure 3. Comparison Graph of Computational Time Between K-Means Clustering Method

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

B. Analysis of K-Means Clustering Test Results with Initial Centroids from Genetic Algorithm Calculations

Based on the test results in Section 3.1, 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 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 process of dimension reduction can help speed up process of clustering k-means with initial centroids obtained from genetic algorithms.

VI. CONCLUSION

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

REFERENCES

[1] R. K. Mishra, “Text Document Clustering on the basis of Inter passage approach by using K-means,” pp. 110–113, 2015.

[2] J. Yadav and M. Sharma, “A Review of K-mean Algorithm,” vol. 4, no. 7, pp. 2972–2976, 2013.

[3] T. Badriyah, “Hybrid Modeling KMeans – Genetic Algorithms in the Health Care Data,” vol. 2, no. 1, 2014.

[4] “Optimizing K-Means by Fixing Initial Cluster Centers,” vol. 4, no. 3, pp. 2101–2107, 2014.

[5] M. Kaushik and B. Mathur, “Comparative Study of K-Means and Hierarchical Clustering Techniques,” vol. 2, no. 6, pp. 93–98, 2014.

[6] B. K. Khotimah, F. Irhamni, and T. Sundarwati, “A GENETIC ALGORITHM FOR OPTIMIZED INITIAL CENTERS K-MEANS CLUSTERING IN SMEs,” *J. Theor. Appl. Inf. Technol.*, vol. 1590, no. 1, 2016.

[7] Z. S. Younus *et al.*, “Content-based image retrieval using PSO and k-means clustering algorithm,” *Arab. J. Geosci.*, no. Salamah 2010, 2014.

[8] M. Kaur and U. Kaur, “Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 7, pp. 2277–128, 2013.

[9] P. M. Dhanya, M. Jathavedan, and A. Sreekumar, “Implementation of Text clustering using Genetic Algorithm,” vol. 5, no. 5, pp. 6138–6142, 2014.

[10] 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, pp. 278–284.

[11] M. Sciences, “COMPARATIVE ANALYSIS OF K-MEANS AND GENETIC,” vol. 3, no. 2, pp. 257–265, 2012.

[12] W. Lesmawati, A. Rahmi, and W. Firdaus Mahmudy, "Optimization of Frozen Food Distribution Using Genetic Algorithms," *J. Enviromental Eng. Sustain. Technol.*, vol. 3, no. 1, pp. 51–58, 2016.