Initial Centroid Determination Using Simulated Annealing Algorithm

Osvari Arsalan, Rizki Kurniati, Elin Darnela
Informatics Engineering, Universitas Sriwijaya
Palembang, Indonesia
Email: osvariarsalan@ilkom.unsri.ac.id, rizkikurniati@ilkom.unsri.ac.id, elindarnela@gmail.com

***Abstract*—Initial randomly generated centroids are commonly used in k-Means clustering method. Random initial centroids k-Means to be trapped in optimum local solution which results in sub-optimal cluster quality. This study examines Simulated Annealing algorithm in determining initial centroids on k-Means. Each k-Means clustering will be tested on result of reduction and without dimension reduction. Based on the results evaluation of k-Means clustering results with initial centroid Simulated Annealing algorithm improve quality cluster with percentage change value 21.2% in the high dimensional data and 25.1% in the dimension reduction data, this shows that initial centroid calculated Simulated Annealing algorithm is able to obtain the best cluster with significant results.**

Kata Kunci*— *Clustering, K-Means, Simulated Annealing Algorithm

I. INTRODUCTION

One of the most popular clustering algorithms is k-Means. Initial centroid generation of k-Means is generally random. Initial random generation of centroids often results in sub-optimal cluster performance, because random centroid selection techniques will sometimes produce initial centroid positions that are close together, causing K-Means to often get stuck in local optimum. The results are inconsistent [3],[4]. Calculations using Simulated Annealing algorithm can be used to obtain k-Means cluster center point. Simulated Annealing is an algorithm that is able to do iterative enhancements to correct problems resulting from clustering techniques such as random, iterated iteratively until desired number of iterations has been reached and until algorithm is stopped because it has reached a stable (convergent) [5]. Each k-means clustering method with both initial random centroids and initial centroids obtained from Simulated Annealing algorithm was tested on high-dimensional data (without dimension reduction) and without dimension reduction. Therefore, this study focuses to determine effect of determining initial centroids using Simulated Annealing algorithm in grouping high dimensional data.

II. METHODOLOGY

2.1. Dataset

The research data test used was Indonesian language journal which downloaded through garuda.ristekdikti.go.id site with 100 text documents. Then contents of document are copied into a file with the extension. Txt and each file is saved in same folder. Text data is then converted to numeric data. The preprocessing stage is case folding, tokenizing, stop word removal, and stemming. Then weighting is done using *tf-idf*. After weighting process is carried out, then weighted data is used for the k-means clustering process. This k-means clustering process uses data from dimension reduction and without dimension reduction. K-means clustering uses dimension reduction data which is the result of weighting followed by process of dimension reduction.

2.2. Clustering

K-means clustering is an extensively used technique for data cluster. Process determining number of clusters or k (numbers of cluster) values is first performed. K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible and refers to averaging of the data finding centroid.

K-value test uses the average DBI value obtained from k-Means clustering with random initial centroids. The best DBI (Davis Boudin Index) evaluation value is used to determine optimum k-value. The most optimum k value is k-value which has decreased most significantly. It can be seen using elbow method which is a method to determine optimum number of clusters looking at the greatest change in value in comparison with number of other clusters [6]. Curves resulting from determining optimum k-value are illustrated in Figure 1.

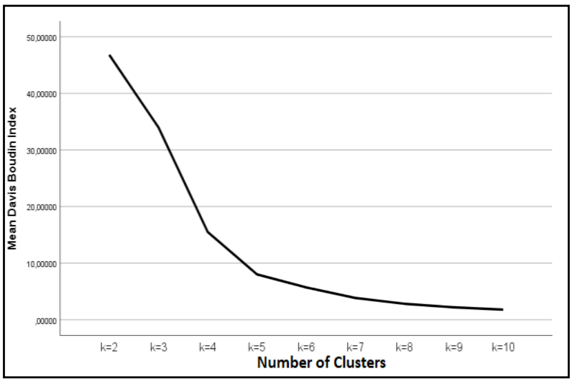

Figure 1. Curve Result of Optimal k Value Determination

Based on the results of testing optimum k value on curve it can be concluded that the most optimum k value lies at k = 5 with an average DBI value of 8.01612996, which is indicated by sharpest angle of the curve results. So, value of k = 5 is optimum and most essential number of clusters of the data used.

After obtaining the optimum number of Cluster (k), in the process of clustering k-Means on high dimensional data and dimensional reduction data. The initial random centroid generator is performed, which will then be carried out the process of calculating the distance between the data to the initial centroid, using the euclidean distance formula [7]. Where each data will be grouped based on the minimum distance.

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

Next step generate new centroid values by calculating average values of different data in the same cluster. Calculate minimum distance between data and new centroid that has been obtained using euclidean distance formula in equation (1). Repeating process of determining new centroid and recalculate minimum distance until convergent conditions are met, ie conditions where the data obtained in each cluster in next iteration is the same as previous iteration (unchanged) [8].

2.3. Determination Initial Centroid

In the previous process, k-Means are processed in the initial centroid randomly, with the same two data that is the result of reduction and without reduction data, then k-Means is processed using the initial centroid which is determined first using the Simulated Annealing algorithm. Then the two results are compared. Simulated Annealing was developed by Kirkpatric (1983) which is used for combinatorial optimization which is a variant of Metropolis algorithm. Simulated Annealing is a numerical optimization technique with principle of thermo-dynamic [9]. Simulated Annealing is an algorithm that is able to carry out iterative improvements to correct problems resulting from clustering techniques such as random, iterated iteratively until thedesired number of iterations has been reached and until the algorithm is stopped because it has reached a stable (convergent) [10].

To be able to do Simulated Annealing process, obtained 3 parameters that will be a benchmark for calculation of Simulated Annealing, namely the number of iterations (i), initial temperature (T0), and final temperature (Tf / Final). The three input parameters that will be selected are parameters that have the lowest energy reduction value, where the energy in this method is the average result of calculating the distance between data to each of the smallest centroids, and also from the results of faster time computing. Amount of iteration will affect the length of time the process is performed on a calculated object. To be able to continue Simulated Annealing process, an optimal initial temperature (T0) input is also needed. The initial temperature (T0) in this method is the initial condition of algorithm with the initial centroid state still random. Simulated Annealing algorithm simulates the cooling process which gradually decreases the system temperature to converge in a frozen and stable state. After obtaining an optimal number of iterations and initial temperatures (T0) Simulated Annealing algorithm also requires an optimal final temperature (Low) as the location of temperature drop which can be said that the algorithm has reached convergence [5].

Initial stages of this algorithm are after entering the algorithm parameters, generating the initial State (S). The initial state is an initial centroid that is generated randomly. Followed by the initial energy calculation (Efirst) as the energy calculated from the initial state, which is calculated using the calculation of the minimum distance of data as in equation (1), then do a check during Ti> Low continue on the iteration process i = 0 to reach Imax before checking the temperature (Evaluation of Ti = new Temperature). Furthermore, the new state update, the new state is a combination of one of the new centroids that have been updated and the old centroids (new centroids will be used one by one). After the new state is updated, calculate the final energy (Efinal). Next update the energy to produce new energy using the formula ΔE = Efirst - Efinal. To be able to check the state calculate p Probability Boltzman with the formula (2).

$$p = e^{\left(\frac{-\Delta E}{kT}\right)} \qquad (2)$$

and generate random numbers P where 0 <p <1. If ΔE <0, new state is accepted, instead use comparison if P (RandNum) < p (Boltzmann probability) new state is accepted instead, use initial state. Next do temperature decrease (Ti).

$$= T_0 x \left(\frac{T_A}{T_0}\right)^{\frac{i}{n}} \qquad (3)$$

Repeat the stages new state update, and new energy update, then if i0 ≤ i max do the temperature check until it is stable and if Ti ≤ T0 then, do k-Means clustering with S as initial centroid obtained from last state on this algorithm [10].

2.4. Clustering Testing

K-means clustering testing with random initial centroids and initial centroids obtained from the Simulated Annealing algorithm each tested 30 times using both dimension reduction and dimension reduction data, which then recorded DBI values, the number of iterations and computational time for each test. in order to know comparison of results of each clustering method. The steps taken are the normality test and data homogeneity test, in this case using Kolmogorov-Smirnov test for data normal test. Based on the results, the data used did not meet norm of normality, so that parametric test was performed using Kruskal-Wallis H and the Mann-Whitney test to determine whether there were 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, it is found that the optimum k value is k = 5. Before testing the clustering method, the parameter of the Simulated Annealing algorithm is tested, namely number of iterations is performed 5 times by entering 10, 20, 30, 40, 50for iteration. Test the initial temperature value (ti) 5 times by entering 50, 100, 150, 200, 250. Testing the final temperature value (t final) 5 times by entering 0.1, 0.2, 0.3, 0.4, 0.5. Based on the results of testing the Simulated Annealing algorithm parameters that have been done, the most optimum Simulated Annealing algorithm parameters obtained are the number of iterations = 20 iterations, an optimal initial temperature value T0 = 150 is obtained for data without dimensional reduction, and T0 = 100 for the dimension reduction data.

3.1. Test Result of Clustering

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

Table 1. Comparison Result of Clustering

	Random Centroid		**Simulated Annealing 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	11.63715	10.05319	9.16683	8.71066
Iteration	5	6	7	5
Time (sec)	865.5394	842.9769	1999.10	1473.63

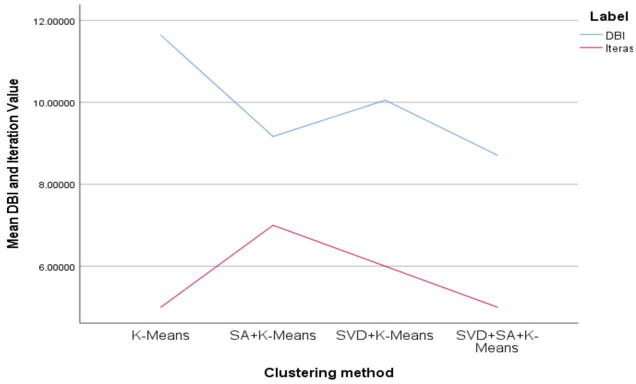

Figure 2. Comparison Graph of DBI Value and Iteration Between K-means Clustering Method

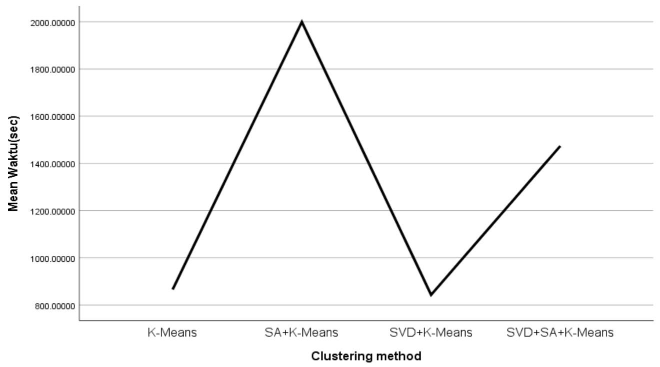

Figure 3. Comparison Graph of Computational Time Between K-Means Clustering Method

Based on the results of tests that have been done it can be concluded that the k-means clustering with the initial centroid of the calculation of the Simulated Annealing algorithm has better cluster quality. However, k-means clustering with initial centroids obtained from the calculation of the Simulated Annealing algorithm on data carried out in the previous dimension reduction process has a faster overall processing time when compared to the k-means clustering process with the initial centroids obtained from the Simulated Annealing algorithm and use data that is not done in process dimension reduction.

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

Seen from section 3.1 After processing data on the k-Means clustering with centroids from the calculation of the Simulated Annealing algorithm on high dimensional data or dimensionally reduced data, when viewed from the results the DBI value results in a significant decrease. So in this test in terms of formatting, the DBI k-Means value which is calculated by centroids first using the Simulated Annealing algorithm has a better cluster quality with a significant difference. When viewed based on the complexity of the algorithm in terms of time and iteration there is a significant change that is equal to 0,000, but the average value generated both time and iteration of the k-Means algorithm with random centroids is faster and smaller than the k-Means algorithm which calculated the centroid first using the Simulated Annealing algorithm.

IV. Conclusion

In research that focuses on determining the initial centroid k-means using the Simulated Annealing algorithm in addition to the effect of dimensional reduction on high-dimensional data, it can be concluded that it is proven necessary to do an initial centroid determination on k-Means algorithm using Simulated Annealing algorithm. Calculating initial centroids on k-Means can improve quality of clustering results, with an increase in cluster quality results by 21.2% of data without dimension reduction with an average DBI value of 9.16683586. and an increase in cluster quality results of 25.1% of the data with first dimension reduction process with an average DBI value of 8.71066892. Accelerate process of achieving a convergent condition on k-Means but has long overall computational time due to addition of the process time of initial centroid calculation using Simulated Annealing algorithm, both if tested on result of dimension reduction or without dimension reduction. In future work proposed method can be produced faster computational time with better quality of cluster results.

REFERENCES

[1] J. Yadav and M. Sharma, “A Review of K-mean Algorithm,” vol. 4, no. 7, pp. 2972–2976, 2013.
[2] R. K. Mishra, “*Text Document Clustering on the basis of Inter passage approach by using K-Means*,” pp. 110–113, 2015.
[3] “Optimizing K-Means by Fixing Initial Cluster Centers,” vol. 4, no. 3,

pp. 2101–2107, 2014.

[4] 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.

[5] D. M. Bennett, “[No Title],” *Br. J. Psychiatry*, vol. 205, no. 01, pp. 76–77, 2014.

[6] A. Syakur, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” 2018.

[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] A. Kaushik and S. Ghosh, “*A Survey on Optimization Approaches to K-Means Clustering using Simulated Annealing*,” vol. 847, no. 3, pp. 845–847, 2014.

[10] A. R. Barakbah, A. Fariza, and Y. Setiowati, “*Optimization of Initial Centroids for K-Means using Simulated Annealing*,” 2005.