PERFORMANCE ANALYSIS OF COMBINED METHODS OF GENETIC ALGORITHM AND K-MEANS CLUSTERING IN DETERMINING THE VALUE OF CENTROID

1Adya Zizwan Putra, 2Muhammad Zarlis, 3Erna Budhiarti Nababan

123Master of Informatics Engineering - Fasilkom TI - University of North Sumatera
1E-mail : putrazizwanadya@gmail.com

Abstract. The determination of Centroid on K-Means Algorithm directly affects the quality of the clustering results. Determination of centroid by using random numbers has many weaknesses. The GenClust algorithm that combines the use of Genetic Algorithms and K-Means uses a genetic algorithm to determine the centroid of each cluster. The use of the GenClust algorithm uses 50% chromosomes obtained through deterministic calculations and 50% is obtained from the generation of random numbers. This study will modify the use of the GenClust algorithm in which the chromosomes used are 100% obtained through deterministic calculations. The results of this study resulted in performance comparisons expressed in Mean Square Error influenced by centroid determination on K-Means method by using GenClust method, modified GenClust method and also classic K-Means.

Keywords: Centroid, K-Means, GenClust, Chromosome

1. Introduction
Clustering is a process in searching and classifying data that has similarity characteristics between one data with other data. Clustering is widely used in areas such as social network analysis, software engineering, and crime detection. There are several clustering algorithms that can be used, but the K-Means and Fuzzy C-Means algorithms are commonly used algorithms because they are quite simple (Bai et al., 2011). Clustering is one of a group of data mining (Larose, 2006). In the K-Means algorithm, determining the number of clusters and centroid determination is quite difficult to do. Determination of the number of clusters and centroid determination directly affects the quality of the clustering process (Maitra, et al., 2010).

The GenClust method is considered good enough to determine the number of clusters and also the centroid of each cluster as it allows for increased clustering performance for each generation. However, what needs to be taken into consideration is the experiments performed by Rahman and Islam (2014) using 50% of the chromosomes obtained through deterministic calculations and 50% of chromosomes obtained through random numbers. Rahman and Islam (2014) at the end of the study suggest that the GenClust method will experience computational constraints when applied to large datasets. Therefore, Rahman and Islam (2014) at the end of his research suggested taking samples of data on the dataset and implementing GenClust to obtain the best chromosome and then implement it as an initial centroid on K-Means. However, there is no further research that discusses the performance comparison of the GenClust method and the modified GenClust method in accordance with the advice of Rahman and Islam (2014).

This study will discuss the comparison between met This study will discuss the comparison between the GenClust method, the modified GenClust method and the classical K-Means in centroid
determination especially in the comparison to measure the measured performance value of Mean Square Error occurring for a dataset.

1.1. Formulation of The Problem
The placement of a data within a dataset in the K-Means algorithm is based on the proximity of the data with the centroid of each cluster, so that the centroid value is very influential on the clustering results by using the K-Means algorithm. The problem is in the K-Means algorithm centroid determination is done by generating random numbers, so the clustering results can show unfavorable results. This study will use the modified GenClust method, in which the chromosomes used are all obtained through deterministic calculations. Through this research we will get the performance comparison expressed in Mean Square Error (MSE) value between classic K-Means algorithm, GenClust method, and also modified GenClust method. The smaller Mean Square Error (MSE) means the better the performance of the obtained K-Means algorithm.

1.2. Scope of Problem
Due to the extent of the problem and the limited time and knowledge of the researchers, the researchers limit the problems to be discussed in this study as follows.
1. This research will discuss the effect of determining the number of clusters and centroid on a large dataset by comparing GenClust method with classic K-Means algorithm.
2. Comparison of performance in this study is based on the Mean Square Error value obtained in each generation (iteration).

1.3. Benefits of Research
The benefits of research are as follows.
1. Through this research the researcher will get the result of analysis about the relation between cluster number and centroid determination on clustering result.
2. Knowing the achievement of performance obtained by using GenClust method when compared with using classic K-Means algorithm.

2. Literature Review

2.1. Genetic Algorithm
The genetic algorithm is a stochastic algorithm that modeled the evolutionary process of biological species through natural selection (Konar, 2005). In general, this population is generated randomly and the solution that is generated after the consecutive stage of the crossover and mutation process. Each individual of the population has an associated value into a fitness value, in relation to solving a problem (Rabunal, 2006). The block diagram of the classical genetic algorithm can be seen in Figure 1.

![Figure 1. Block diagram of the Genetic Algorithm (Negnevitsky, 2005)](image-url)
2.2. GenClust

The GenClust algorithm is a type of Hybrid Clustering that combines genetic algorithms with K-Means. The study of GenClust was first performed by Rahman and Islam (2014). Collaboration between Genetic Algorithm and K-Means can prevent K-Means from being trapped under Local Minima and can achieve Global Optima conditions. The GenClust method is considered good enough to determine the number of clusters and also the centroid of each cluster as it allows for increased clustering performance for each generation. The process stages of the GenClust Algorithm can be seen in Figure 2.

![Figure 2. Process Stages of the GenClust Algorithm](image)

The basic idea of determining the number of clusters is to use random numbers generated with a range of cluster numbers of 2 to \( n \), where \( n \) is the sum of the objects in the dataset. The position of centroid generated in chromosome form is 50% of chromosomes obtained through deterministic calculations and 50% of chromosomes are obtained through random numbers, eg the number of chromosomes is 60, then 30 chromosomes are obtained from deterministic calculations and 30 chromosomes are obtained randomly. The steps in obtaining chromosomes through deterministic calculations are as follows (Rahmad and Islam, 2014).

1. Find the random number coordinates of \( r_x \) which express the coordinates of the centroid used as references.
2. Take the coordinates of the position of each object in the dataset.
3. Calculate the distance from each object to random number \( r_x \) using euclidean distance.
4. Take 30 objects with the smallest distance to the random number \( r_x \) (suppose it takes 30 chromosomes).
5. Coordinate positions of 30 selected objects fill chromosomes obtained through deterministic calculations.

As for 50% of other chromosomes to be obtained at random, the coordinates that will fill the gene on the chromosome are obtained by generating random numbers that are between the smallest position coordinates and the largest position coordinates. The genetic algorithm is a stochastic algorithm that modeled the evolutionary process of biological species through natural selection (Konar, 2005).

2.3. Performance Measurement on K-Means Algorithm

This study will examine the performance of the K-Means algorithm in relation to centroid determination. Centroid is the center of each cluster. This Centroid will greatly affect the performance of the K-Means algorithm because the placement of a data in a dataset in the K-Means algorithm is based on the proximity of the data with the centroid of each cluster, so that the value of centroid greatly influences the clustering results by using K-Means algorithm.

The researchers in this study will use the modified GenClust method where the overall generated chromosome (100%) is obtained through a deterministic calculation, then the centroid for each cluster
will be obtained based on the best chromosome for use in the clustering process using the K-Means algorithm.

3. Research Methodology
In the K-Means algorithm, determining the number of clusters and centroid determination is quite difficult to do. Determining the number of clusters and centroid determination directly affects the quality of the clustering process. The Hybrid Clustering method known as GenClust combines the use of the K-Means algorithm with the Genetic Algorithm. Genetic Algorithm is used to determine the number of clusters and also the centroid of each cluster. The use of the GenClust method may prevent the K-Means algorithm from being trapped under local optima.

It is necessary to test the determination of the number of clusters and also the centroid determination using GenClust and compare it with the classical K-Means on a large dataset. The existing dataset uses UCI Machine Learning Repository. UCI Machine Learning Repository is a collection of database, domain theory, and data generator used by community learning machine learning, for the purpose of empirical analysis of machine learning algorithm. The dataset available at UCI Machine Learning Repository is used by students, educators, and researchers around the world as the primary source of data sets in machine learning. The amount of data sets available at UCI Machine Learning Repository at this time has amounted to 320 data sets that can be used in accordance with the needs of machine learning learning. This study will discuss the comparison between the GenClust method with the classic K-Means algorithm especially in the comparison to measure the measured performance value of Mean Square Error occurring for a large dataset.

3.1. Data Used
The data used are Iris Data Set benchmark data taken from UCI Machine Learning Repository. Iris Data set is a data set that is widely used in pattern recognition problems. The attributes of information contained in the Iris Data Set are composed of: Sepal Length, Sepal Width, Petal Length, and Petal Width. Iris Data Set has 3 classes: Iris Setosa, Iris Versicolour, and Iris Virginica. Measurement of performance in this research using Mean Square Error (MSE) method.

4. Analysis Results
In this research will be shown the results of performance assessment in connection with centroid determination on K-Means algorithm. This study will compare the performance between centroid determination using the K-Means algorithm using random centroid determination, centroid determination using GenClust algorithm, and centroid determination using the modified K-Means algorithm. Measurement accuracy based on Mean Square Error. The accuracy value will be expressed in terms of Mean Square Error value which is the average value for error based on the test by using the number of iterations which are varied, that is: 50, 75, and 100. Small MSE value indicates that the clustering result using K-Means Clustering has been able to recognize existing patterns, whereas large MSE values indicate that clustering results using K-Means Clustering still have not achieved the desired results. The test will be performed using Iris dataset sourced from UCI Machine Learning Repository. The results of tests conducted by researchers will be presented in tabular form.
Table 1. Testing Results Using Classical K-Means Algorithms, GenClust Algorithms, and Modified GenClust Algorithms

| Algorithm   | 50 iterations | 75 iterations | 100 iterations |
|-------------|---------------|---------------|----------------|
| K-Means     | Average MSE: 1.041, Best MSE: 0.6 | Average MSE: 0.893, Best MSE: 0.43 | Average MSE: 0.503, Best MSE: 0.24 |
| GenClust    | Average MSE: 0.9637, Best MSE: 0.36 | Average MSE: 0.818, Best MSE: 0.4 | Average MSE: 0.506, Best MSE: 0.36 |
| Modified GenClust | Average MSE: 0.673, Best MSE: 0.34 | Average MSE: 0.606, Best MSE: 0.35 | Average MSE: 0.468, Best MSE: 0.11 |

Based on Table 4.13 it can be seen that in general centroid determination influences the performance of the K-Means algorithm shown in the Mean Square Error (MSE) value obtained based on the classification result. Determination of centroid effect on the performance of genetic algorithm because the placement of a data into a dataset based on the proximity between the coordinates of the object with the coordinates of the centroid where the distance calculation used is Euclidean Distance. In the classic K-Means algorithm centroid determination is done by using random numbers. In the GenClust algorithm the centroid determination is based on the genetic algorithm, where 6 chromosomes are used, ie 3 chromosomes are obtained using random numbers and 3 chromosomes are obtained by using deterministic calculations, then the chromosomes will undergo selection, crossover, and mutation. The best chromosome that will be used as a centroid on K-Means Algorithm.

Better performance of the modified GenClust algorithm is shown in terms of Average MSE values as well as better Best MSE values, and obtained both on tests using 50, 75, and 100 iterations. The results in this study may vary due to the random numbers generated, especially within the classic K-Means algorithm and the GenClust algorithm. In addition, the generation of random numbers can also define a number of processes within genetic algorithms such as selection, crossover, and mutation.

5. Conclusions

The conclusions that can be drawn from this research are as follows.
1. The performance shown in terms of the average MSE value and also the Best MSE score indicates that centroid determination determines the performance of the K-Means algorithm. Determination of centroid using K-Means algorithm has been modified better than centroid determination using GenClust algorithm and also K-Means Classic algorithm.
2. The results showed that the larger number of iterations will also provide better performance results on the K-Means algorithm. This happens because at each iteration stage of the K-Means algorithm will be adjusted to the value of centroid based on data items placed in a cluster.

References
[1] Ahmad, A. and L. Dey. 2007. A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data. Data & Knowledge Engineering63: pp. 503-527
[2] Bai, L., J. Liang, & C. Dang. 2011. An Initialization Method to Simultaneously Find Initial Cluster Centers and The Number of Clusters for Clustering Categorical Data. Knowledge-Based Systems24: pp. 785–795.
[3] Cao, F., J. Liang, and L. Bai. 2009. A New Initialization Method for Categorical Data Clustering. Expert Systems with Applications36: pp. 10223-10228
[4] Fayyad, U., G.P. Shapiro, and P. Smyth. 1996. From Data Mining to Knowledge Discovery in Database. AI Magazine: pp. 37-53
[5] Konar, Amit. 2005. Computational Intelligence Principles, Techniques, and Applications. Springer: Calcutta, India
[6] Larose, D.T. 2006. Data Mining Methods and Models. John Wiley and Sons, Inc: New Jersey.
[7] Li, R., X. Q. Li and G. Wang. 2015. Global Optimal and Minimal Solutions to K-Means Cluster Analysis, Proceedings of the International Conference IP, Computer Vision, and Pattern Recognition, pp. 274-283
[8] Maitra, R., A.D. Peterson, A.P. Ghosh. 2010. A Systematic Evaluation of Different Methods for Initializing the K-Means Clustering Algorithm. IEEE Trans. Knowl. Data Eng.
[9] Muzid, Syafiul. 2014. Dinamisasi Parameter AlgoritmaGenetikaMenggunakan Population Resizing on Fitness Improvement Fuzzy Evaluationary Algorithm (PROFIFEA). Prosiding SNATIF 2014, pp. 471-478
[10] Negnevitsky, Michael. 2005. Artificial Intteligence-A Guide to Intelligent Systems. Addison Wesley: Edinburg
[11] Rabunal, J.R. & Dorado, J. 2006. Artificial Neural Networks in Real-Life Applications, Ideal Group Publishing: Hershey, United States of America.
[12] Rahman, M.A. and M.Z. Islam. 2012. CRUDAW: A Novel Fuzzy Technique for Clustering Records Following User Defined Attribute Weights, Proceedings of the Tenth Australasian Data Mining Conference (AusDM), pp. 27-41
[13] Rahman, M.A and M.Z. Islam. 2014. A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means. Knowledge-Based Systems71: pp. 345-365
[14] Rokach, L. And O. Maimon. 2005. Data Mining and Knowledge Discovery Handbook. Springer: Tel Aviv.
[15] Salman, M.S., O. Kukrer, and A. Hocanin. 2017. Recursive Inverse Algorithm: Mean-Square-Error Analysis. Digital Signal Processing 66: pp. 10-17
[16] Serapiao, A.B.S., G.S. Correa, F.B. Goncalves, and V.O. Carvalho. 2016. Combining K-Means and K-Harmonic With Fish School Search Algorithm for Data Clustering Task on Graphics Processing Units. Applied Soft Computing41: pp. 290-304
[17] Shapiro, G.P. 1991. Knowledge Discovery in Real Databases: A Report on The International Joint Conference in Artificial Intelligence (IJCAI). AI Magazine: pp. 68-70
[18] Steinbach, Michael, Ertoz, Leventdun Kumar, Vipin. 2014. The Challenges of Clustering High Dimensional Data. Minnesota Supercomputing Institute