Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2426~2433
e-ISSN: 2548-6861

Enhancing Clustering Accuracy Using K-Means with Seeds Optimization

Adiyah Mahiruna[1]*, Ngatimin[2]*, Rachmat Destriana[3]*, Eko Hari Rachmawanto[4]*, Herman Yuliansyah[5]*, Muhammad Taufiq Hidayat[6]*
[1,2,5]Faculty of Science and Technology, Institut Teknologi Statistika dan Bisnis Muhammadiyah, Indonesia
[3]Faculty of Informatic, Universitas Muhammadiyah Tangerang, Indonesia
[4]Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[6]Faculty of Informatic, Universitas Ahmad Dahlan, Indonesia
*Faculty of Science and Technology, Institut Teknologi Statistika dan Bisnis Muhammadiyah, Indonesia
adiyah.mahiruna@itesa.ac.id[1], ngatimin@itesa.ac.id[2], rahmat.destriana@ft-umt.ac.id[3], eko.hari.rachmawanto@dsn.dinus.ac.id[4], herman.yuliansyah@itesa.ac.id[5], tahufiq182@gmail.com[6]

Article Info

Article history:

Received 2025-07-24
Revised 2025-08-26
Accepted 2025-09-10

Keyword:

Clustering,
Data Mining,
Machine Learning,
Health, Heredity.

ABSTRACT

In this study, the development of the Mean-based method proposed by Goyal and Kumar will be carried out by changing the initial cluster center determination step, which was originally based on the origin point O (0,0), to be replaced with the arithmetic mean. To assess the performance of the proposed method, it will be compared with the Global K-means method and the Mean-based K-means method. In this study, the performance of these methods will be measured using the Davies-Bouldin Index, and the significance of the proposed method will be measured using the Friedman Test. This study proposes a method of Improving K-Means Performance through Initial Center Optimization based on Second Global Average for Clustering Osteoporosis Diagnosis of lifestyle factors. Evaluation of K-Means performance through Initial Center Optimization based on Second Global Average with DBI measurements. The targeted experimental results of this study include improving the performance of K-means optimized through the initial center based on Second Global Average. From the results of nine experiments with the number of clusters [2,3,4,5,6], it can be seen that the method proposed in this study has the same superior performance compared to the Mean Based method and compared to the Global K-means method.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Data mining is defined as the process of finding models from large datasets, the models formed can be used to help decision making. Data mining methods have been successful in various research fields, one of which is the field of clustering. Clustering is one of the most widely used techniques in unsupervised learning, aiming to group objects based on their similarity so that objects in the same cluster are more similar to each other than to those in other clusters [1]. Data mining is defined as the process of finding models from large datasets, the models formed can be used to help decision making [2]. Data mining is also known as knowledge discovery, data mining can also be interpreted as a scientific discipline that studies methods for extracting new information from large amounts of data. Data mining is used to extract new information from a collection of objects that do not yet have meaning, so that data mining has become one of the research fields that is of interest to researchers. Research topics in the field of data mining are association, estimation, prediction, classification, clustering. Data mining methods have been successful in various research fields, one of which is the field of clustering [3]. Clustering is a method of learning without classes [4]. Clustering means forming subgroups, the subgroups formed are the result of partitioning a dataset [5], the subgroups that are formed are called clusters. One of the popular types of clustering methods is the partition method [6]. The partitioning method involves simple steps: dividing a large collection of objects into several groups or clusters. Each object becomes part of a cluster according to the

predetermined cluster specifications [7]. The K-means method is a popular method [8] and has a simple iteration of steps in clustering the data set [9]. The K-means method is among the top ten most popular data mining methods [10]. In study [11] the centroids are obtained randomly, the K-means method was introduced in 1967 in research conducted by MacQueen. In study [12] Likas et al. proposed a method for determining the initial cluster center based on the average. In study [13] Wang and Bai proposed the Global MinMax K-means method which is a development of the Global K-means method. In study [14] Goyal and Kumar determination of the initial cluster center in the Mean-based method is based on the arithmetic average of the data in the dataset whose data distance from the origin point O (0,0) has been calculated. In this study is proposed

Based on these problems, this study focuses on how to determine the optimal initial centre to improve the performance of the K-Means algorithm in clustering, especially in the application of osteoporosis diagnosis based on family history factors. The method used to determine the initial centre of the K-means algorithm is the second global average.

II. METHOD

In research by Hyunjoong Kim [23] it was shown that the sparsity of the centroid with the Public data set showed that the K-means method was accurate in clustering. In research by Junwen Chen [15] shows that the QALO-K method with the Public data set shows that the K-means method is accurate in clustering. In research by Junyan Liu [16] it was shown at the clustering results on ten datasets to assess the accuracy of the suggested method. KCM-K, MKM-K, VMKM-KG, and VMKM-KL initialize the hyper-parameters in the Gaussian kernel using this method. Kernel-based MinMax clustering techniques with metric kernelization and auto-tuning hyper-parameters are presented in this study. The suggested approaches have an advantage over the traditional weighting type k-means approaches and MinMax k-means algorithm in that they can avoid the issues of stochastic initialization and noisy variables. Consequently, the suggested algorithms perform better overall than the KM, KCM-K, W-KMeans, E-WKmeans, MKM, and MKM-K algorithms.

In research by Yating Li [17] The conventional K − means clustering technique relies heavily on the choice of the initial cluster centre and the number of clusters K. If these choices are not made, the clustering results may become less stable. In this paper, a new cluster centre determination method is provided and a method for determining the number of clusters K is introduced. This technique successfully improves the clustering process by determining the initial cluster centre using MNN, density, and distance. Through experiments, the algorithm's excellent robustness is confirmed.

In research by Hailun Xie [7] to address the issues with initialization sensitivity and local optimum pitfalls of the conventional KM clustering algorithm, we have developed two FA variations in this study: IIEFA and CIEFA. In IIEFA and CIEFA, two novel approaches have been put forth to improve search efficiency and diversification. As a result, the assurance of adequate variation among fireflies in comparison at the early convergence stage enhances the search efficiency.

In research by Nan Han [18] Analysis of Effectiveness Test the enhanced DP method against other clustering strategies on the TCM datasets and the UCI benchmark from the perspectives of clustering accuracy and quality. According to the results IABC-DP performs significantly better than other algorithms in the SC, entropy, purity, precision, recall, and F1-measure metrics on both the TCM and UCI benchmark datasets. On the UCI and TCM datasets, it is intriguing to discover that, while having a larger computing complexity than other clustering methods, the IABC-DP approach performs similarly in terms of runtime.

In research by Ahmad Ilham [19] The findings demonstrated that the suggested approach produces great SSE values, particularly for k=4, which has the lowest SSE value as opposed to k=3.It has been demonstrated that applying DT to enhance Goyal and Kumar's approach [6] to the initial centroid improves k-means performance. Thus, it can be said that DT can enhance the initial centroid's k-means performance. There are datasets with numerous properties in this study. The attribute selection approach is responsible for more study. Some researchers, including Tsai et al. and Breaban et al. [19], claim that.[20] Since not all attributes are valuable, attribute selection techniques are crucial to employ in further study since they can enhance the effectiveness of clustering approaches by eliminating irrelevant attributes.

In research by Srividya [20] Enhancing the quality of outlier detection by clustering is the aim of the paper's comparison of the different techniques. Because LOF places greater emphasis on identifying local outliers than the other techniques, it is determined in this research to be the best algorithm for outlier detection. The fact that LOF implements the method using a single parameter is the primary factor contributing to its superior performance. The best outlier identification performance is obtained by analyzing every single point in the dataset, whereas the global outlier results in a lesser outlier detection performance. Thus, the researcher may effectively use LOF to identify the outlier.

In research by Xiaohui Huang [21] present a new clustering framework of the k-means type that regularizes feature weights using the l 2 norm. According to experimental results, the new algorithms outperform the state-of-the-art algorithms in the majority of cases for clustering both numerical and categorical data sets based on four evaluation metrics: accuracy, RandIndex, Fscore, and NMI. This implies that both numerical and categorical data sets can benefit from better clustering outcomes when feature weights are regularized using the l 2 norm.

In research by Wang Gaochao [22] this work proposed CMDC-IA, a novel density-based clustering technique based on affinity propagation and CFSFDP. The suggested approach offers a fresh approach to recommending the crucial parameter dc, which complements this method's clustering

approach nicely. A significant advantage of the suggested approach is that, because of the specification of cluster centre candidates and quantity independence, cluster centre can be selected automatically rather than by hand from the decision graph. The suggested approach performs better overall on the eight well-known synthetic benchmark datasets with low dimensional features when compared to the state-of-the-art techniques. Among the kernel-based clustering methods, CDMC-IA performs competitively when compared to other methods on high-dimensional datasets. To make CDMC-IA a dependable tool in real-world applications, more advancements could be made to reduce the computation time.

In research by Kumar Majhi [23] K-Means has gained popularity for cluster analysis due to its ease of use and effectiveness. However, this clustering method's drawback is that centroid positions are initialized randomly. Ant Lion Optimization, a nature-inspired optimization technique, has been integrated into this work attempt to enhance the clustering quality of the K-means clustering algorithm. The performance of K-Means, KMeans-PSO, KMeans-ALO, DBSCAN, and Revised DBSCAN is contrasted with that of KMeans-ALO. The outcomes of the simulation confirm that KMeans-ALO outperforms K-Means and the other two hybrid techniques.Significant differences exist between Kmeans, Kmeans-PSO, Kmeans-FA, Kmeans-ALO, DBSCAN, and Revised DBSCAN, according to the Friedman test. Additionally, Kmeans-ALO outperforms K-Means, KMeans-PSO, Kmeans-FA, DBSCAN, and Revised DBSCAN according to the Holm test. The statistical analysis's level of confidence is 0.10, indicating that the suggested Kmeans-ALO produced findings with 90% accuracy.

In research by S.A. Sajidha [24] Finding initial seed artifacts for the K-modes technique is the primary goal of the algorithm the researchers present in their paper. In order to select the seed artifacts from distinct clusters and dense places. The researchers suggested approach performs better in four of the six data sets they examined, demonstrating the effectiveness of the proposed algorithm in detecting initial seed artifacts and guaranteeing the repeatability of the clustering solution at the best possible cost. The researchers suggested methodology takes longer than previous methods since it takes into account the density estimation of all the data artifacts, which necessitates calculating the distances of each data artifact from all the data facts. However, for most data sets, the proposed method produces superior clustering results when compared to the past.

In research by Donghua Yu [25] The INCK algorithm, a novel K-medoids cluster algorithm, was suggested. It is based on a subset of candidate medoids and gradually increases the number of clusters. The INCK algorithm maintains the simplicity and speed of the FastK algorithm by utilizing its updated medoids approach and distance matrix.

In research by Md Abdul Masud [26] The suggested algorithms considerably beat two popular approaches, Elbow and Silhouette, in identifying the correct 565 number of clusters in the data, according to the experimental findings reported on both synthetic and real-world datasets. The clustering findings using the k-means++ approach, MMCA, and randomly chosen beginning cluster centre were worse to those from the suggested method, and demonstrated that the accuracy of clustering using initial cluster centre discovered by I-niceSO was comparable to that of clustering using genuine class centre. Furthermore, the clustering accuracy in the k-means clustering process was enhanced by the I-niceMO technique. It is possible to enhance the existing study on choosing initial cluster cente54 from unbalanced datasets. In the future, use distance distributions of several effective observation points to extend the I-nice technique to a semi-supervised clustering model.

In research by Erzhou Zhu [27] In order to discover cluster centers fast, increase the clustering algorithm's stability and accuracy, and decrease the number of iterations, this work first suggested an enhanced K-means algorithm based on density parameters for the initial centre selection. Compared to other previous cluster validity indices, VCVI is more accurate at determining the ideal clustering number. In particular, for spatial distribution datasets with "within-cluster compactness, between-cluster separation," the experimental findings from testing various datasets showed that our new VCVI and algorithm can efficiently obtain the optimal clustering number and the optimal clustering partition.

In research by Huanqian Yan [28] Based on statistical automatic centroid identification from the decision graph, a novel clustering technique is presented in this study. It is demonstrated that datasets with different distributions and dimensionalities may be handled by the suggested ADPC approach, and that the suggested statistical-based centroid identification outperforms the straightforward threshold-based centroid identification. Furthermore, picture segmentation can also be accomplished successfully with the ADPC approach. In subsequent research, we intend to enhance the ADPC approach for handling challenging decision graphs in order to increase its accuracy in estimating the number of the cluster.

In research by Weiling Cai [29] A dimension reduction approach that preserves both local and global clustering structure information is presented in this study. An unsupervised linear dimension reduction technique that works well with cloud-distributed data is what we have. Their technique first generates cluster labels using the clustering method, after which the global and local structures can be defined.

In research by Wei Xue [30] proposed a new K-means algorithm that updates the cluster centre model and uses the distance metric of spatial density similarity. The approach makes the new k-Means more appropriate for nonlinear manifold data by utilizing global information on the data geometric distribution. When compared to the conventional k-Means algorithm, they obtain highly outstanding results

when evaluating its performance on both synthetic and real-world data sets.

In this study focuses on how to determine the optimal initial centre to improve the performance of the K-Means algorithm in clustering. Figure 1 was shown flow of the proposed method in this study.

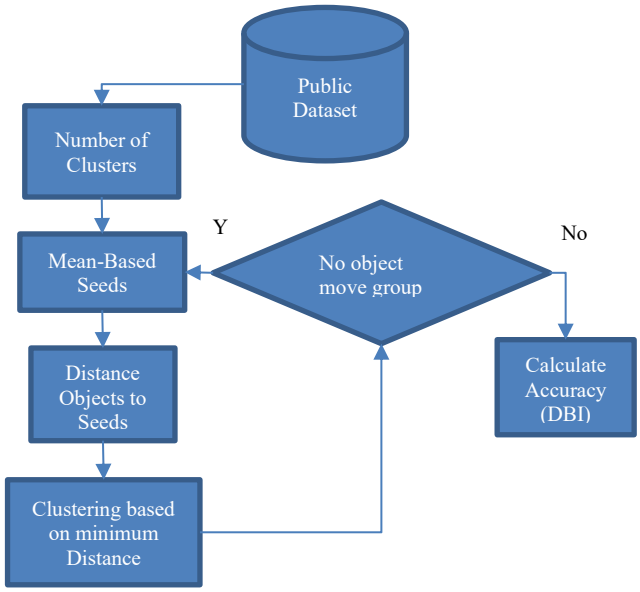

Figure 1. Global K-means

Step by step the Mean-Based method:

1. Prepare the dataset
2. Determine the number of clusters k
3. Random Seeds
4. Calculate the distance of each data with the seeds
5. No object move group then Calculate the Davies-Bouldin Index (DBI)

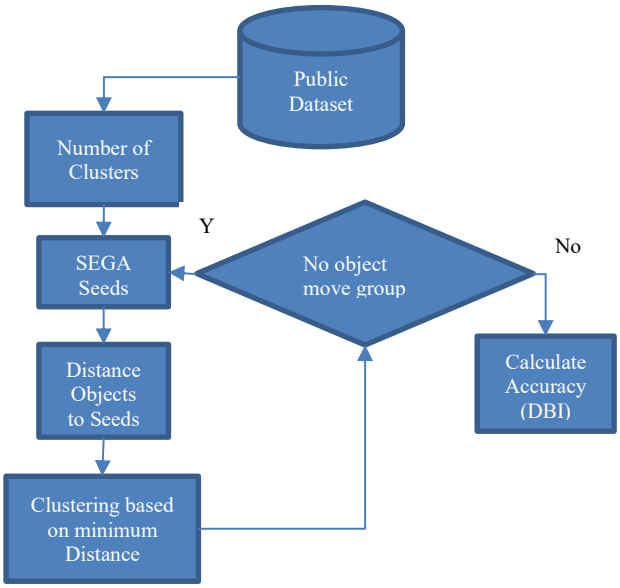

Figure 2. flow of proposed method

Step by step the proposed method:

1. Prepare the dataset
2. Determine the number of clusters k
3. Calculate the Scond lobal Average (SEGA) of the dataset to determine the Seeds.

$$\text{Average} = \frac{1}{N}\sum_{i=1}^{N} xi, xi \in 1, 2, \ldots, N \quad (1)$$

Where:

N = number of data
X = attribute in dataset
Xi = i-th data from X-th attribute

4. Calculate the distance of each data with the seeds $\bar{x}$ obtained

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

Where xi is the i-th data of the X-th attribute in the dataset and n is the number of the dataset in the dataset, $\bar{x}$ is the global average.

5. No object move group then Calculate the Davies-Bouldin Index (DBI).

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max (\text{Ri}, \text{j}) \quad (4)$$

Where k = number of clusters

$R_{i,j}$ = Ratio Comparation cluster-i, cluster-j

In this study, a public dataset sourced from the University of California Irvine (UCI) will be used. In this study, experiments will be conducted using several public datasets, there are datasets containing outlier data. The dataset used in this study can be seen in Table 1.

TABLE 1.
DATASET USED TO TEST THE METHOD

No	Dataset Name	Data Amount	Number of Attributes	Number of Classes
1	Immunotherapy	90	7	2
2	Breast Tissue	106	9	4
3	Glass	214	9	6

III. RESULTS AND DISCUSSION

The results obtained by measuring the performance of the Mean-based method, the Global K-means method, and the method proposed in this study. The performance of all methods to be tested will be measured using the Davies-Bouldin Index (DBI). The datasets used are publicly available and include glass, immunotherapy, and breast tissue.

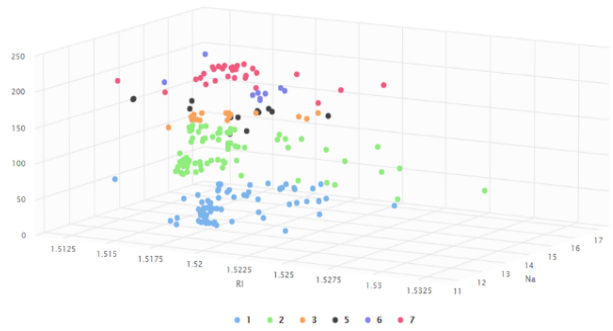

Figure 2. Scatter Diagram for Glass Dataset

Figure 2 shows the visualization of the Glass dataset, the number of clusters formed is 6 clusters. This Glass dataset has a total of 214 data. The number of attributes of the Glass dataset is 9 attributes, namely Refractive Index (RI), Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), Iron (Fe). Figure 3 shows the visualization results of the Immunotherapy dataset, with the number of clusters equal to two clusters. The number of attributes of the Immunotherapy dataset is 7 attributes, namely Sex, Age, Time, Number of Warts, Type, Area, Induration Diameter. The Iris dataset downloaded from UCI has a total of 90 data.

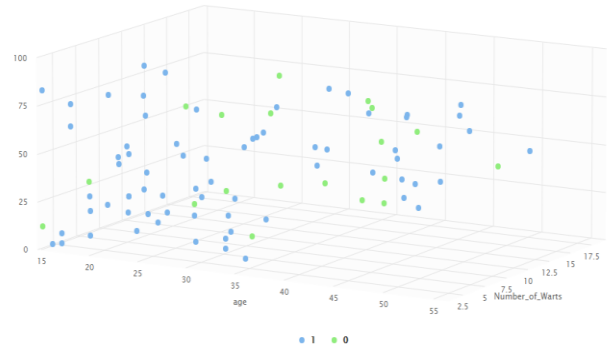

Figure 3. Scatter Diagram for Immunotherapy Dataset

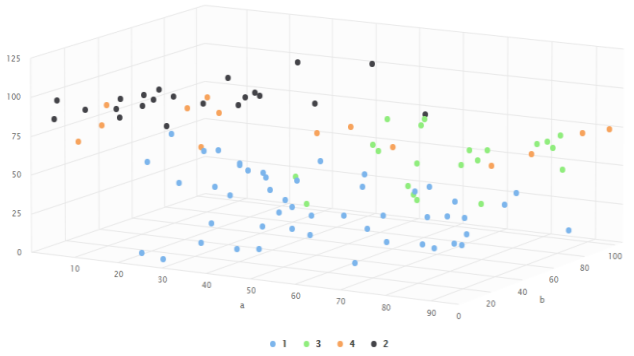

Figure 4. Scatter Diagram for Breast Tissue Dataset

Figure 4 shows the visualization of the Breast Tissue dataset, the number of clusters owned by this Breast Tissue dataset is four clusters. The number of samples from this Breast Tissue dataset is 106 data. This Breast Tissue dataset has 9 attributes, namely Age, Body Mass Index (BMI), Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1.

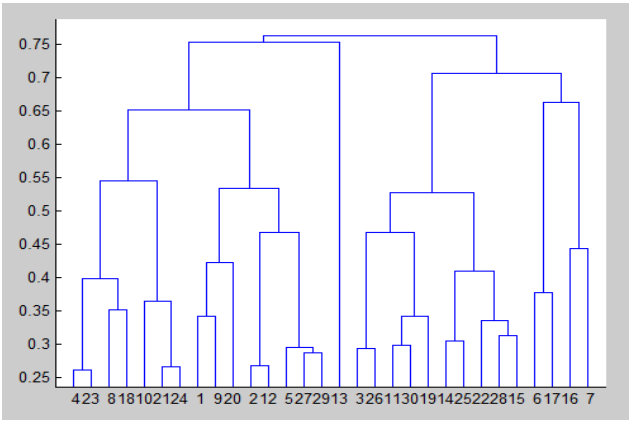

Figure 5. Dendogram visualization of Breast Tissue Dataset

Figure 5. is a Dendogram diagram of the Breast Tissue dataset, the x-axis is the member data of the cluster, the y-axis is the cut point. The number of data samples from this Breast tissue dataset is 106 data. The dendogram diagram can be used to show the existing cluster members if it is to be determined how many clusters should be formed. From Figure 5. it is known that if two clusters are to be formed, then the first cluster has members of data 4 to 13, for the second cluster has members of data 3 to 7.

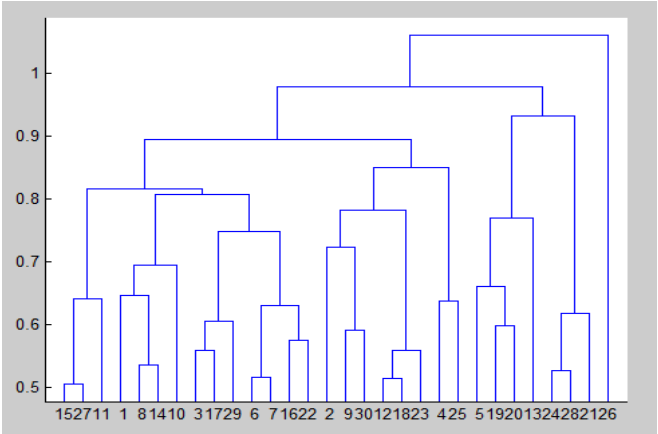

Figure 6. Dendogram visualization of Immunotherapy Dataset

Figure 6. is a Dendogram diagram of the Immunotherapy dataset, the x-axis represents the member data of the cluster, the y-axis is the cut point. The number of data samples from this Immunotherapy dataset is 90 data. From Figure 6. it is known that if two clusters are to be formed, then the first cluster will have data members from 15 to 2 data, for the second cluster only has data members of 126 data.

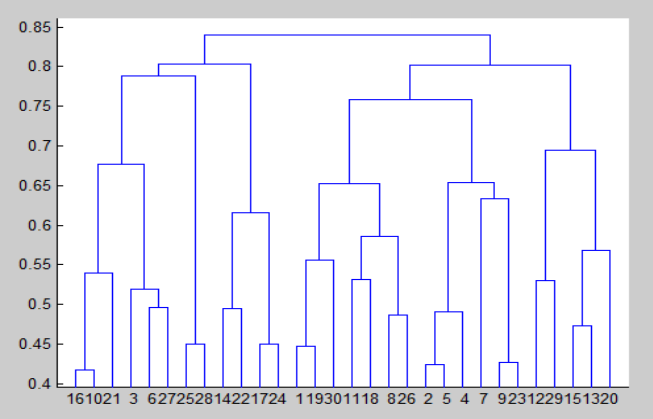

Figure 7. Dendogram visualization of Glass Dataset

Figure 7. is a Dendogram diagram of the Glass dataset, the x-axis is the member data of the cluster, the y-axis is the cut point. The number of data samples from this Breast tissue dataset is 214 data. From Figure 7. it is known that if two clusters will be formed, then the first cluster will have members of data 16 to 24, for the second cluster will have members of data 1 to 20. From Figure 7. it is known that if three clusters will be formed, then the first cluster will have members of data 16 to 24, for the second cluster will have members of data 11 to 23, and for the third cluster will have members of data 12 to 20. Figure 8 is a K-means graph of the Glass dataset with the number of clusters [2,3,4,5,6]. The cut point of the Glass dataset is obtained by dividing the number of sample data from the dataset by the number of clusters.

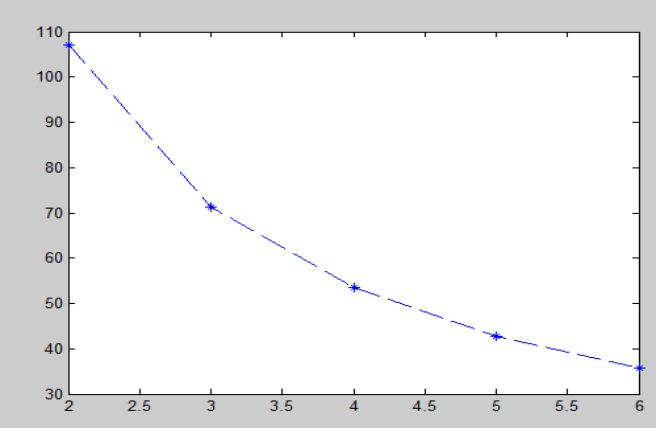

Figure 8. K-means graph of Glass Dataset

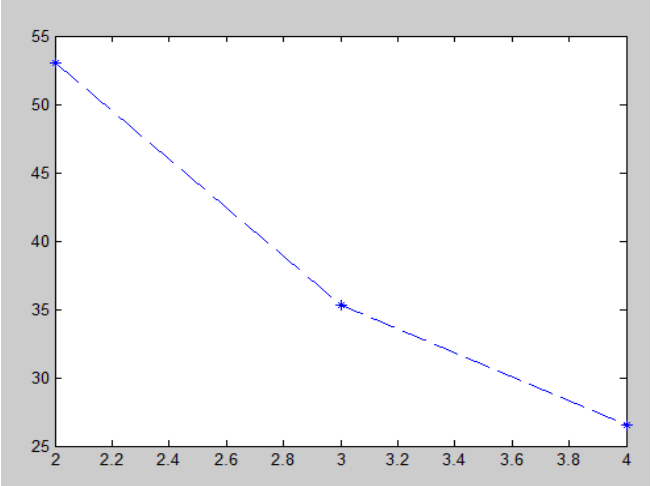

Figure 9. K-means graph of Breast Tissue dataset

Figure 9 is a K-means graph of the Breast Tissue dataset with the number of clusters [2,3,4]. The cut point of the Breast Tissue dataset is obtained by dividing the numbers of data from the dataset by the number of clusters. Figure 10. is a K-means graph of the Immunotherapy dataset with the number of clusters [2:9], the x-axis is the number of clusters, the y-axis is the cut point. The cut point of the Immunotherapy dataset is obtained by dividing the numbers of data by the number of clusters.

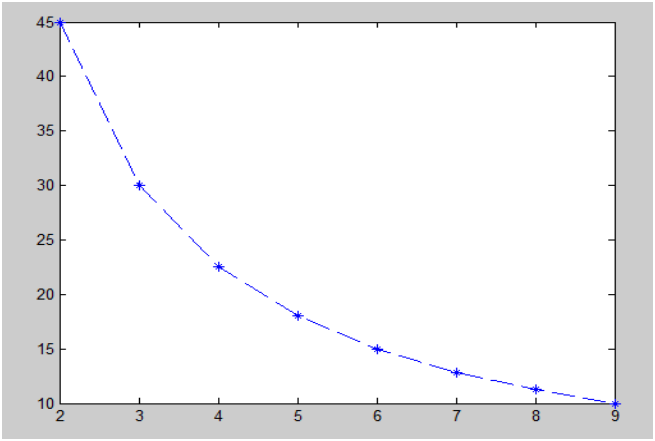

Figure 10. K-means graph of Immunotherapy dataset

TABLE 2
TESTING PHASE GLOBAL K-MEANS METHOD [12]

#	Dataset	Number of Clusters	DBI
1	Breast Tissue	2	2.3231
2	Breast Tissue	3	2.1067
3	Breast Tissue	4	1.8335
4	Glass	2	0.6690
5	Glass	3	0.8293
6	Glass	4	0.8573
7	Glass	5	0.9301
8	Glass	6	1.1012
9	Immunotherapy	2	0.3977

Table 2 above shown that the DBI value of the method proposed by Likas et al [12] has the best performance when the number of clusters is equal to four on the Breast Tissue dataset. Meanwhile, the worst performance is obtained when the number of clusters is equal to two. From Table 2 it is known that based on DBI measurements on the Breast Tissue dataset the best performance was obtained when the number of clusters was equal to two, while the worst performance was obtained when the number of clusters was equal to six.

TABLE 3
TESTING PHASE MEAN-BASED K-MEANS METHOD [14]

#	Dataset	Number of Clusters	DBI
1	Breast Tissue	2	2.3231
2	Breast Tissue	3	2.1067
3	Breast Tissue	4	1.8335
4	Glass	2	0.6690
5	Glass	3	0.8293
6	Glass	4	0.8573
7	Glass	5	0.9301
8	Glass	6	1.1012
9	Immunotherapy	2	0.3977

Table 3 above shown that based on DBI measurements on the Glass dataset, the method proposed by Goyal and Kumar [14] has the best performance when the number of clusters is equal to three. Meanwhile, the worst performance is obtained when the number of clusters is equal to two. Then it is known that based on DBI measurements on the Breast Tissue dataset, the best performance was obtained when the number of clusters was equal to two, while the worst performance was obtained when the number of clusters was equal to five.

TABLE 4
TESTING PHASE PROPOSED METHOD

#	Dataset	Number of Clusters	DBI
1	Breast Tissue	2	2.5144
2	Breast Tissue	3	2.1067
3	Breast Tissue	4	1.9358
4	Glass	2	0.6690
5	Glass	3	0.8293
6	Glass	4	0.8573
7	Glass	5	1.1308
8	Glass	6	1.1012
9	Immunotherapy	2	0.3977

Table 4 above shown that based on DBI measurements, the method proposed in this study has the best performance when the number of clusters is equal to two on the Immunotherapy dataset. Meanwhile, the worst performance is obtained when the number of clusters is equal to two on the Breast Tissue dataset.

TABLE 5
COMPARISON THE METHOD

#	Dataset	Number of Clusters	Mean-based method	Number of Clusters	Mean-based method
1	Breast Tissue	3	2.1067	2.1067	2.1067
2	Glass	2	0.6690	0.6690	0.6690
3	Glass	3	0.8293	0.8293	0.8293
4	Immunotherapy	2	0.3977	0.3977	0.3977

Based on Davies-Bouldin Index (DBI) measurements, it is known that based on DBI measurements, the proposed method has the same superior performance compared to the Mean Based method and the Global K-means method when applied to the Glass, Immunotherapy and Breast Tissue datasets. The method proposed method in this study has the best performance when the number of clusters is equal to two on the Immunotherapy dataset, while the worst performance was obtained when the number of clusters was equal to three on the Breast Tissue dataset.

Comparison of Davies-Bouldin Index Measurement Performance Methods in Friedman Test

In a study, a comparative analysis was conducted to evaluate the statistical significance of the differences between the proposed method and existing methods [31].

TABLE 6
THE METHOD FRIEDMAN TEST ON DBI

Q (Observed value)	1,5294
Q (Critical value)	5,9915
DF	2
p-value (Two-tailed)	0,4655
alpha	0,05

H0: There is no difference in mean values between the compared methods.
Ha: There is a difference in mean values between the compared methods.

A p-value greater than the alpha value means rejecting Ha and accepting H0. Therefore, a Nemenyi post-hoc test is not necessary.

IV. CONCLUSION

Based on the results of nine experiments with the number of clusters [2,3,4,5,6], it can be seen that the method proposed in this study has the same superior performance compared to the Mean Based method and compared to the Global K-means method. The performance of the method proposed in this study has superior performance compared to the Mean Based method on the Breast Tissue dataset with the number of

clusters equal to two, and on the Glass dataset with the number of clusters equal to five. The method proposed in this study has superior performance compared to the Mean Based method on the Breast Tissue dataset and on the Glass dataset, and has the same superior performance on the Immunotherapy dataset.

ACKNOWLEDGEMENT

The author would like to thank the Directorate of Research, Technology, and Community Service for the financial support that has made this research possible.

REFERENCES

[1] F. Bagaswara, M. A. Muthalib, and R. Meiyanti, "Clustering of Futsal Interest Level Among Students K-Means Method," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 41–50, 2025, doi: 10.52088/ijesty.v5i3.879.

[2] P.-N. Tan *et al.*, *Pang.N I Ng Tan*. 2006.

[3] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1476–1482, 2014, doi: 10.1016/j.eswa.2013.08.044.

[4] K. M. Kumar and A. R. M. Reddy, "An efficient k-means clustering filtering algorithm using density based initial cluster centers," *Inf. Sci. (Ny).*, vol. 418–419, pp. 286–301, 2017, doi: 10.1016/j.ins.2017.07.036.

[5] S. Wang and W. Shi, *Data mining and knowledge discovery*. 2012. doi: 10.1007/978-3-540-72680-7_5.

[6] C. Zhang, D. Ouyang, and J. Ning, "An artificial bee colony approach for clustering," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4761–4767, 2010, doi: 10.1016/j.eswa.2009.11.003.

[7] H. Xie *et al.*, "Improving K-means clustering with enhanced Firefly Algorithms," *Appl. Soft Comput. J.*, vol. 84, p. 105763, 2019, doi: 10.1016/j.asoc.2019.105763.

[8] Y. Li, K. Zhao, X. Chu, and J. Liu, "Speeding up k-Means algorithm by GPUs," *J. Comput. Syst. Sci.*, vol. 79, no. 2, pp. 216–229, 2013, doi: 10.1016/j.jcss.2012.05.004.

[9] H. Xue, Q. Yang, and S. Chen, *Nugroho*, vol. 6, no. SVM. 2009. doi: 10.1007/s10115-007-0114-2.

[10] R. T. Aldahdooh and W. Ashour, "DIMK-means 'Distance-based Initialization Method for K-means Clustering Algorithm,'" *Int. J. Intell. Syst. Appl.*, vol. 5, no. 2, pp. 41–51, 2013, doi: 10.5815/ijisa.2013.02.05.

[11] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, pp. 281–297, 1967, doi: 10.1007/s11665-016-2173-6.

[12] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003, doi: 10.1016/S0031-3203(02)00060-2.

[13] X. Wang and Y. Bai, "The global Minmax k-means algorithm," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-3329-4.

[14] M. Goyal and S. Kumar, "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability," *J. Inst. Eng. Ser. B*, vol. 95, no. 4, pp. 345–350, 2014, doi: 10.1007/s40031-014-0106-z.

[15] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," *Knowledge-Based Syst.*, vol. 203, p. 106167, 2020, doi: 10.1016/j.knosys.2020.106167.

[16] J. Liu, Y. Guo, D. Li, Z. Wang, and Y. Xu, "Kernel-based MinMax clustering methods with kernelization of the metric and auto-tuning hyper-parameters," *Neurocomputing*, vol. 359, pp. 173–184, 2019, doi: 10.1016/j.neucom.2019.05.056.

[17] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A Novel Algorithm for Initial Cluster Center Selection," *IEEE Access*, vol. 7, pp. 74683–74693, 2019, doi: 10.1109/ACCESS.2019.2921320.

[18] N. Han, S. Qiao, G. Yuan, P. Huang, D. Liu, and K. Yue, "A novel Chinese herbal medicine clustering algorithm via artificial bee colony optimization," *Artif. Intell. Med.*, vol. 101, p. 101760, 2019, doi: 10.1016/j.artmed.2019.101760.

[19] A. Ilham, D. Ibrahim, L. Assaffat, and A. Solichan, "Tackling Initial Centroid of K-Means with Distance Part (DP-KMeans)," *Proceeding - 2018 Int. Symp. Adv. Intell. Informatics Revolutionize Intell. Informatics Spectr. Humanit. SAIN 2018*, pp. 185–189, 2019, doi: 10.1109/SAIN.2018.8673364.

[20] Srividya, S. Mohanavalli, N. Sripriya, and S. Poornima, "Outlier Detection using Clustering Techniques," *Int. J. Eng. Technol.*, vol. 7, no. 3.12, p. 813, 2018, doi: 10.14419/ijet.v7i3.12.16508.

[21] X. Huang, X. Yang, J. Zhao, L. Xiong, and Y. Ye, "A new weighting k-means type clustering framework with an l2-norm regularization," *Knowledge-Based Syst.*, vol. 151, pp. 165–179, 2018, doi: 10.1016/j.knosys.2018.03.028.

[22] G. Wang, Y. Wei, and P. Tse, "Clustering by defining and merging candidates of cluster centers via independence and affinity," *Neurocomputing*, vol. 315, pp. 486–495, 2018, doi: 10.1016/j.neucom.2018.07.043.

[23] S. K. Majhi and S. Biswal, "Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer," *Karbala Int. J. Mod. Sci.*, vol. 4, no. 4, pp. 347–360, 2018, doi: 10.1016/j.kijoms.2018.09.001.

[24] S. A. Sajidha, S. P. Chodnekar, and K. Desikan, "Initial seed selection for K-modes clustering – A distance and density based approach," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018, doi: 10.1016/j.jksuci.2018.04.013.

[25] D. Yu, G. Liu, M. Guo, and X. Liu, "An improved K-medoids algorithm based on step increasing and optimizing medoids," *Expert Syst. Appl.*, vol. 92, pp. 464–473, 2018, doi: 10.1016/j.eswa.2017.09.052.

[26] M. A. Masud, J. Z. Huang, C. Wei, J. Wang, I. Khan, and M. Zhong, "I-nice: A new approach for identifying the number of clusters and initial cluster centres," *Inf. Sci. (Ny).*, vol. 466, pp. 129–151, 2018, doi: 10.1016/j.ins.2018.07.034.

[27] E. Zhu and R. Ma, "An effective partitional clustering algorithm based on new clustering validity index," *Appl. Soft Comput. J.*, vol. 71, pp. 608–621, 2018, doi: 10.1016/j.asoc.2018.07.026.

[28] H. Yan, L. Wang, and Y. Lu, "Identifying cluster centroids from decision graph automatically using a statistical outlier detection method," *Neurocomputing*, no. xxxx, 2018, doi: 10.1016/j.neucom.2018.10.067.

[29] W. Cai, "A dimension reduction algorithm preserving both global and local clustering structure," *Knowledge-Based Syst.*, vol. 118, pp. 191–203, 2017, doi: 10.1016/j.knosys.2016.11.020.

[30] R. S. Xue W, Yang RL, Hong XY, Zhao N, "A novel k-Means based on spatial density similarity measurement".

[31] A. Mahiruna, E. H. Rachmawanto, and D. Istiawan, "Analysis of Time Optimization for Watermark Image Quality Using Run Length Encoding Compression," *J. Intell. Comput. Heal. Informatics*, vol. 4, no. 2, p. 35, 2023, doi: 10.26714/jichi.v4i2.12058.