Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/20733
DOI URL: http://dx.doi.org/10.21474/IJAR01/20733

RESEARCH ARTICLE

ENHANCING CLUSTERING PERFORMANCE: A HYBRID GENERALIZED K-MEANS APPROACH

Nwoye, O. N.[1] and Okoli, C. N.[2]

1. Procurement Department, National Engineering Design Development Institute (NEDDI), Nnewi, Anambra State, Nigeria.
2. Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Uli, Anambra State, Nigeria.

Manuscript Info

Manuscript History
Received: 15 Febuary 2025
Final Accepted: 18 March 2025
Published:April 2025

Key words:-
Generalized K means, Clustering algorithm, Data segmentation, Pattern recognition, Computational efficiency

Abstract

This study developed a hybrid Generalized K means clustering algorithm to boostclustering accuracy, robustness and computational efficiency across diverse datasets. The proposed method integrates multiple clustering techniques, including Forgy, Lloyd, MacQueen, Hartigan and Wong, Likas and Faber, improving initialization, assignment, and updating processes. Advanced distance metrics, particularly the Mahalanobis distance, are incorporated to account for variable correlations and variances, ensuring precise cluster assignments. The algorithm's effectiveness is validated using datasets from the World Bank Commodity Price Publication 2022 and the R console repository, including Edgar Anderson's Iris data set, COVID-19 mortality outcomes with hydroxychloroquine and chloroquine, and nicotine replacement therapy studies for smoking cessation. The methodology combines robust initialization strategies with iterative assignment and centroid update mechanisms, ensuring convergence to optimal clustering solutions. Performance comparisons with traditional K-means methods revealed the hybrid algorithm's superior accuracy, stability and efficiency, particularly in data sets with varying dimensions, distributions and complexities. By leveraging secondary data from reliable sources, the study ensures comprehensive analysis and generalization of findings. The study's findings have implications for improved pattern recognition, data segmentation and decision-making across domains, showcasing the algorithm's potential as a robust alternative to existing clustering techniques.

Introduction:-

Clustering is a fundamental task in data analysis, aimed at organizing large data sets into meaningful subgroups based on similarity metrics. Among various clustering techniques, the K-means algorithm stands out as the most widely used partitioning-based method due to its simplicity and effectiveness in practical applications (Estivill-Castro, 2002). The algorithm iteratively partitions a database of N objects into K disjoint clusters, optimizing the

Corresponding Author:-Nwoye, O. N.
Address:-Procurement Department, National Engineering Design Development Institute (NEDDI), Nnewi, Anambra State, Nigeria

within-cluster squared error criterion to measure clustering quality (Yuan & Yang, 2019). Despite its popularity, K-means has inherent limitations, such as sensitivity to initial centroid selection, susceptibility to local optima, and challenges in handling high-dimensional and large-scale data sets. The classic K-means algorithm, first introduced by Forgy (1965), minimizes the average squared Euclidean distance between data points and their respective cluster centroids. Forgy's approach initializes centroids randomly, leading to variable clustering outcomes. Lloyd (1982) refined this by treating data distribution discretely, while MacQueen (1967) introduced an online version of the algorithm that updates centroids dynamically during iterations. Further modifications by Hartigan and Wong (1979) sought to optimize the within-cluster sum of squares (SSE) by reassigning data points across clusters iteratively. These foundational algorithms underscore the iterative two-phase process of centroid updates and data point assignment, which continues until convergence (Oti et al., 2021).

Numerous K-means variants have been developed over the years to address its limitations. For instance, Jancey (1966) proposed a modification to accelerate convergence, while Bagirov and Mardaneh (2006) introduced the Modified Global K-means (MGKM) algorithm to enhance performance on gene expression data sets. Weighted K-means, proposed by Huang et al. (2005), incorporated variable weights to prioritize relevant features in high-dimensional data. Similarly, Amorim (2012) and Amorim and Mirkin (2012) developed the Restricted Minkowski Weighted K-means to compute cluster-specific feature weights, demonstrating its adaptability to complex data sets.The development of advanced K-means algorithms also includes innovations like the filtering algorithm by Kanungo et al. (2002), which leverages k-dtrees to efficiently partition data and the Continuous K-means by Faber (1994), which employs random sampling for faster convergence on large data sets. Additionally, global optimization techniques such as the Global K-means algorithm by Likas et al. (2003) use K-means as a local search method to overcome initialization dependency. Despite all the contributions by the authors, there are still some challenges need to be addressed, hence this study.

Related Literature Review.

The reviewed studies collectively underscore the evolution and adaptability of the K-Means clustering algorithm in diverse applications.Obaid (2023) explored the interplay between H-index, paper citations, and scholarly appraisal in computer science using K-Means clustering, augmented by visual analytic tools like Orange Data Mining and Power BI. This study highlighted the role of machine learning in data exploration, providing valuable insights into academic influence. Hao et al. (2023) tackled association rule mining challenges by integrating matter-element theory with an improved K-Means algorithm, demonstrating enhanced efficiency and accuracy in rule extraction. Addressing inherent flaws in K-Means. Liu et al. (2023) improved robustness and clustering accuracy through Turkey rules and advanced centre point selection. Kim et al. (2023) applied K-Means to analyze student engagement in online learning, revealing actionable insights to foster inclusive educational environments.Wang et al. (2023) utilized K-Means in smart city initiatives, segmenting power consumers to enhance electricity demand forecasting by 85.25%, showcasing its transformativepotential in urban planning. Kotun et al. (2023) emphasized the challenges of K-Means, including reliance on user-defined parameters and Euclidean distance, while El-Sharkawy et al. (2024) employed K-Means for precise breast cancer diagnosis using hyper spectral imaging. Fox et al. (2024) addressed faulty centre scenarios in clustering, presenting fixed-parameter tractable algorithms with scalable and resilient solutions. Rungruang et al. (2024) proposed a hybrid approach combining formal concept analysis (FCA) with the recency, Frequency and Monetary (RFM) model for customer segmentation, bridging the gap between data insights and actionable marketing strategies.Vishwakarma et al. (2024) highlighted K-Means' superiority in analyzing genetic data sets, leveraging the Calinski-Harabaz Index to demonstrate its efficacy. Mahmud et al. (2024) introduced a distributed clustering framework, using innovative methods like density peak-based clustering and firefly-inspired algorithms, achieving improved scalability and stability in big data clustering. These studies collectively advance the discourse on K-Means clustering, addressing its limitations and extending its applications across domains, thereby enriching the field of clustering and data analytic.Despite these advancements, challenges persist in improving clustering performance. Sujatha and Sona (2013) emphasized the importance of robust clustering methods, particularly for large and high-dimensional data sets. They identified limitations in existing algorithms, such as high time complexity and inadequate performance in diverse scenarios. To address these gaps, researchers have explored hybrid approaches, combining the strengths of multiple algorithms to enhance clustering accuracy and efficiency. This study proposes a Hybrid Generalized K-means Approach to improve clustering performance. Building on the strengths of traditional K-means and its variants, this hybrid method introduces novel centroid initialization and iterative update strategies to mitigate common limitations. By integrating techniques from recent advancements, the proposed approach aims to achieve superior intra-cluster variance minimization, accuracy, and computational efficiency, as demonstrated through both simulated and real-world data sets.

Research Methodology.

Method Of Data Collection.

The World Bank Commodity Price Publication 2022 and the R console repository provided the secondary data used in the study. The World Bank Commodity Price Publication 2022, the data source, provided the study's core data set. This data set offered thorough data on commodity prices. Since the World Bank is well-known for its excellent data sets and exacting data gathering procedures, using their data ensured legitimacy and dependability. Secondary data from the R console repository was used in addition to the World Bank data set to enhance the analysis and evaluate the viability of the suggested approach in comparison to other approaches that were taken into consideration for the study. With so many data sets available in so many different disciplines, the R console repository gives researchers access to a wide range of pertinent and varied data sources. To make sure that the results were reliable and applicable to a variety of data sets, the study sought to improve the analysis's validity and comprehensiveness by integrating secondary data from the R console repository. Hence, the combination of data from the World Bank Commodity Price Publication 2022 and the R console repository enabled a thorough and rigorous examination of the research questions. The research outputs were deemed more credible and reliable due to the study's capacity to form strong conclusions and validate its findings through the utilization of several secondary data sources.

Description Dataset from R console Repository

i. Edgar Anderson's Iris Data

A well-known and frequently used data set in the fields of statistics and machine learning is the iris data set, which was named after Edgar Anderson. The data set offers measures in centimeters for several characteristics of iris flowers, including sepal length, sepal width, petal length, and petal width. There are three distinct species of iris flowers: Iris setosa, Iris versicolor, and Iris virginica. The data set consists of 150 cases, or rows, where each case is a measurement for a single iris flower. Five variables or columns make up the data set: Species, and Sepal.Length, Petal.Length, and Petal.Width. The first four variables represent the measurements of the flowers' physical characteristics, and the species of the iris flower that corresponds to each set of data is indicated by the fifth variable, Species. Furthermore, the data set's iris3 format displays the same data slightly differently. It is organized as a three-dimensional array, with dimensions that match the measurements (Sepal L., Sepal W., Petal L., and Petal W.), the species of iris, and the case number inside each species sub-sample.

The iris data set is an essential part of many research projects and instructional programs since all things considered, it is a useful resource for a variety of statistical analyses, pattern recognition tasks, and machine learning algorithms.

ii. Mortality Outcomes with Hydroxychloroquine and Chloroquine in COVID-19 from an International Collaborative Meta-Analysis of Randomized Trials (dat.axfors2021)

The dat.axfors2021 dataset provides valuable insights into the outcomes of 33 clinical trials investigating the effectiveness of hydroxychloroquine or chloroquine in patients with COVID-19. These trials, both published and unpublished, have been conducted internationally as part of a collaborative effort to understand the potential benefits or risks associated with these medications.

Each row in the data set represents a specific trial, with detailed information provided in several columns:

id: A unique identifier for each trial.

acronym: A shortened registry number for quick reference.

patient_setting: Describes the setting in which the patients were treated.

blinding_exact: Indicates whether the study was conducted with exact blinding protocols.

high_dose: Specifies whether patients received a high or low dose of the medication.

Published: Indicates the publication status of the trial results.

hcq_cq: Specifies the type of medication administered (hydroxychloroquine or chloroquine).

hcq_arm_event: The number of deaths observed in the treatment group.

hcq_arm_total: The total number of patients in the treatment group.

control_arm_event: The number of deaths observed in the control group.

control_arm_total: The total number of patients in the control group.

Control: Describes the type of control group used (Standard of Care or Placebo).

The primary focus of the data set is on the comparison of mortality outcomes between the treatment and control groups across different trials. This information is crucial for understanding the potential impact of hydroxychloroquine or chloroquine on patient outcomes in the context of COVID-19 treatment.

iii. Data on Studies on the Effectiveness of Nicotine Replacement Therapy for Smoking Cessation (dat.hartmannboyce2018)

The dataset provides comprehensive insights into the effectiveness of nicotine replacement therapy (NRT) for smoking cessation based on the findings of 133 studies. These studies have been conducted to evaluate the long-term impact of NRT on individuals who are attempting to quit smoking, with a focus on abstinence outcomes at 6 months or more of follow-up.

Each row in the data set represents a specific study, and several columns provide detailed information:

study: A unique identifier for each study.

x.nrt: The number of participants in the NRT group who were abstinent at the follow-up assessment.

n.nrt: The total number of participants in the NRT group.

x.ctrl: The number of participants in the control group who were abstinent at the follow-up assessment.

n.ctrl: The total number of participants in the control group.

treatment: Specifies the type of NRT provided in the treatment group (e.g., gum, patch, inhalator).

The data set aims to offer a comprehensive overview of the effectiveness of NRT across various forms and delivery methods, as indicated by the treatment variable. By analyzing the abstinence outcomes in both the NRT and control groups, researchers and healthcare professionals can gain insights into the efficacy of NRT for smoking cessation over an extended period.

Hence, the "dat.hartmannboyce2018"data set serves as a valuable resource for understanding the real-world effectiveness of NRT interventions and informing evidence-based strategies for smoking cessation programs and policies.

The proposed Generalized K-means Clustering Algorithm

Given a matrix or data frame of n observations and m variables and interest is in clustering the data into k number of clusters. This k-means clustering method looks at improving the initialization method, the assignment method and the updating method by employing the combination of existing k-means clustering techniques, including the Forgy, Lloyd, MacQueen, Hartigan and Wong, Likas, and Faber's clustering method, in an effort to improve the initialization, assignment, and updating processes.

The proposed algorithm starts by:

Computing the number of observations (n) and the number of variables (m) in the input data.

Perform initialization based on the selected initialization method:

If the initialization method is "forgy", randomly select k observations from the data as the initial centroids.

Choose k observations at random from the data if the initialization technique is "forgy" to serve as the initial centroids.

If the initialization method is "lloyd", use the first k observations as the initial centroids.

If the initialization method is neither "forgy" nor "lloyd", throw an error.

Start the iteration loop (iteration) from 1 to maximum iterations:

Perform the assignment step based on the selected assignment method:

If the assignment method is "macQueen":

Compute the pairwise distances between the centroids and the data points using the Mahalanobis distance.

The Mahalanobis distance, which accounts for the correlations and variances of the variables, is a measurement of the separation between a point and a distribution. The formula for the Mahalanobis distance between a point X and a distribution with mean μ and covariance matrix Σ is:

$$Mahalanobisdistance = \sqrt{\frac{(X-\mu)^{\tau}(X-\mu)}{\Sigma}} \quad (1)$$

In equation (1), (X - μ) represents the difference between the point X and the mean μ, Σ^{-1} is the inverse of the covariance matrix Σ, and "τ" denotes the transpose operation.

When compared to a standard Euclidean distance, the Mahalanobis distance takes the variables' scales and correlations into account, providing a more precise distance measurement, particularly when working with data sets where the variables are correlated or have different variances (Torra and Narukawa, 2012).

The next step is to assign each data point to the nearest centroid based on the minimum distance.

If the assignment method is "Hartigan & Wong":

For each data point, find the centroid with the minimum sum of squared differences between the data point and the centroid.

Assign the data point to the nearest centroid.

If the assignment method is "Likas":

Compute the pairwise distances between the centroids and the data points using equation (1).
Assign each data point to the nearest centroid based on the minimum distance.
Check if the number of unique clusters is less than k.
If the number of unique clusters is less than k, repeat the initialization step and assignment step until k unique clusters are obtained.
If the assignment method is "Faber":
Compute the pairwise distances between the centroids and the data points using equation (1).
Assign each data point to the nearest centroid based on the minimum distance.
Check if the number of unique clusters is less than k.
If the number of unique clusters is less than k, repeat the initialization step and assignment step until k unique clusters are obtained.
For each cluster (i), iteratively update the centroid:
Initialize weights (w) for each cluster as equal (1/k).
While the weight for cluster i (w[i]) is greater than the tolerance (tol):
the current centroid for cluster i will be stored as old centroid.
Update the centroid for cluster i by computing the weighted mean of the data points assigned to cluster i.
Compute the distances between each data point in cluster i and the updated centroid. The weights (w) based on the inverse of the distances normalized by their sum will be updated.
If the squared difference between the updated centroid and the old centroid is less than the squared tolerance (tol^2), there will be a break in the iteration.
The Update Methods (Centroid Update) was done for each of the methods considered in the study by:
a) MacQueen method ("macqueen"):
This method does not involve explicit centroid updates. It only assigns data points to the nearest centroids based on the pairwise distances.
b) Hartigan-Wong method ("hartigan_wong"):
Since the Hartigan-Wong method does not perform centroid updates, we move on to the next assignment method.
c) Likas method ("likas"):
Repeat the steps of the MacQueen method as described above.
If the number of unique clusters is less than k, repeat the initialization and assignment steps until k unique clusters are obtained.
Note that the Likas method does not involve explicit centroid updates.
d) Faber method ("faber"):
Repeat the steps of the Likas method as described above.
For each cluster (i), iteratively update the centroid using the weighted mean of the data points assigned to that cluster:
Initialize the weight vector (w) for each cluster with equal weights (1/k).
While the weight for cluster i (w[i]) is greater than the tolerance (tol):
the current centroid for cluster i will be stored as the old centroid.
Update the centroid for cluster i by computing the weighted mean of the data points assigned to cluster i:
Update the weights (w) based on the inverse of the distances normalized by their sum:

$$w = \frac{1/distances}{\sum(1/distances)} \qquad (2)$$

If the squared difference between the updated centroid and the old centroid is less than the squared tolerance (tol^2), break the iteration.
Repeat the assignment and update steps for the specified maximum number of iterations. At the end of the iteration loop, return the final cluster assignments (clusters). Hence, this proposed generalized k-means clustering algorithm provides a complete explanation of the proposed clustering algorithm, including the initialization, assignment, and centroid update steps for each method.

Results And Discussion Of The Analysis

The results presented in Tables 1 and 2 summarize the mean values of the clusters for the various data sets analyzed in the study. Table 1 demonstrates that the proposed method consistently outperforms traditional clustering methods (Forgy, Lloyd, MacQueen, Hartigan and Wong) by achieving higher mean values across all data sets and cluster numbers (k). Table 2 further supports these findings, showing that while traditional methods exhibit erratic or non-linear trends in mean cluster values as k changes, the proposed method remains stable and superior.

The results in Table 1 demonstrate that the proposed method consistently achieves higher mean values for clusters across all data sets and numbers of clusters (k) compared to other methods (Forgy, Lloyd, MacQueen, and Hartigan and Wong). This trend is observed across diverse data sets with varying dimensions, such as the Iris data set, dat.axfors2021, dat.hartmannboyce2018, and World Bank Commodity Price Data. While the mean values for traditional methods fluctuate, sometimes increasing or decreasing erratically with k, the proposed method shows a relatively stable and higher performance. This indicates its robustness and superior capability in clustering, as higher mean values suggest better intra-cluster homogeneity and inter-cluster separation. These results highlight the proposed method's effectiveness in identifying meaningful patterns and structures in the data, making it a reliable choice for clustering tasks across diverse data sets.

Table 1. Result of the mean values of clusters for the various dataset considered in the study

Data	Dimension	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
Iris	150 x 5	2	76.1866	1.6466	1.6467	1.64666	1.646667
		3	76.3466	2.04	2.4333	1.92	1.92
		4	78.16	2.5133	2.42	2.2133	2.0666
		5	76.2533	2.7133	2.6266	2.7866	2.6866
		6	77.9133	3.1933	3.6133	3.7733	3.46
		7	79.0733	4.3333	4.1533	4.16	4.2733
		8	77.8333	4.4333	4.96	4.8466	4.5666
		9	79.4733	4.88	4.9333	4.6866	5.1933
		10	80.22	5.2266	6.0066	5.3333	5.2133
dat.axfors2021	33 x 12	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
		2	16.8484	1.0303	1.9696	1.9696	1.0303
		3	17	1.33333	1.8787	1.2121	2.7878
		4	17.606	2.606	2.3939	2.606	2.6969
		5	17.1818	3	4.0606	3.4545	3.7878
		6	18.9697	4	3.7272	3.9393	3
		7	17.606	5.0909	4	4.2727	3.1212
		8	19.0606	5.5454	4.2424	4.5151	5.2424
		9	19.5454	4.606	5.6969	5.6666	5.4545
		10	19.7575	4.0909	6.7878	4.9696	4.8181
dat.hartmannboyce2018	133 x 6	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
		2	69.55882	1.1102	1.8897	1.9117	1.0661
		3	69.9044	1.2279	2.1838	1.2279	1.8161
		4	70.4044	2.7132	2.0661	1.6029	2.2426
		5	71.0367	2.3455	2.4411	3.1544	3.1544
		6	71.4117	3.8676	3.3161	3.3529	2.9338
		7	71.44853	4.7573	5.05147	4.375	4.2647
		8	71.5441	3.6102	5	3.7867	4.7426
		9	72.6544	5.397	6.3823	5.4411	5.5441
		10	73.8823	5.2205	5.2352	4.6911	5.6176
World Bank Commodity Price Data	62 x 7	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong

2	32.0161	1.7581	1.2419	1.75806	1.2419
3	33.1129	1.5806	2.129	1.6774	2.2741
4	33.7741	2.4032	2.7258	2.6451	2.4354
5	34.6935	3.5161	2.88709	3.4354	3.75806
6	33.1612	3.2903	3.5161	3.8064	3.3548
7	33.1774	3.9193	4.75806	4.6774	4.25806
8	35.5322	4.7258	4.1451	4.1774	3.7419
9	36.2903	4.7419	4.7096	5.2258	5.1129
10	34.6129	6.6935	5.6451	5.129	5.5483

.

The results presented in Table 2 and Figure 1 indicate that the proposed method outperforms traditional clustering methods (Forgy, Lloyd, MacQueen, Hartigan and Wong) in terms of mean cluster values across various numbers of clusters (k). While traditional methods exhibit non-linear or erratic relationships between mean values and k, with trends such as initial increases, declines, or erratic variations, the proposed method demonstrates consistently higher and more stable mean values. This stability suggests its robustness and resilience to changes in k, making it effective for diverse clustering scenarios. Higher mean values produced by the proposed method imply improved intra-cluster homogeneity and inter-cluster separation, signifying its superior ability to uncover meaningful and interpretable clusters. Consequently, the proposed method provides better clustering outcomes, enabling researchers to derive more accurate insights and make informed decisions.

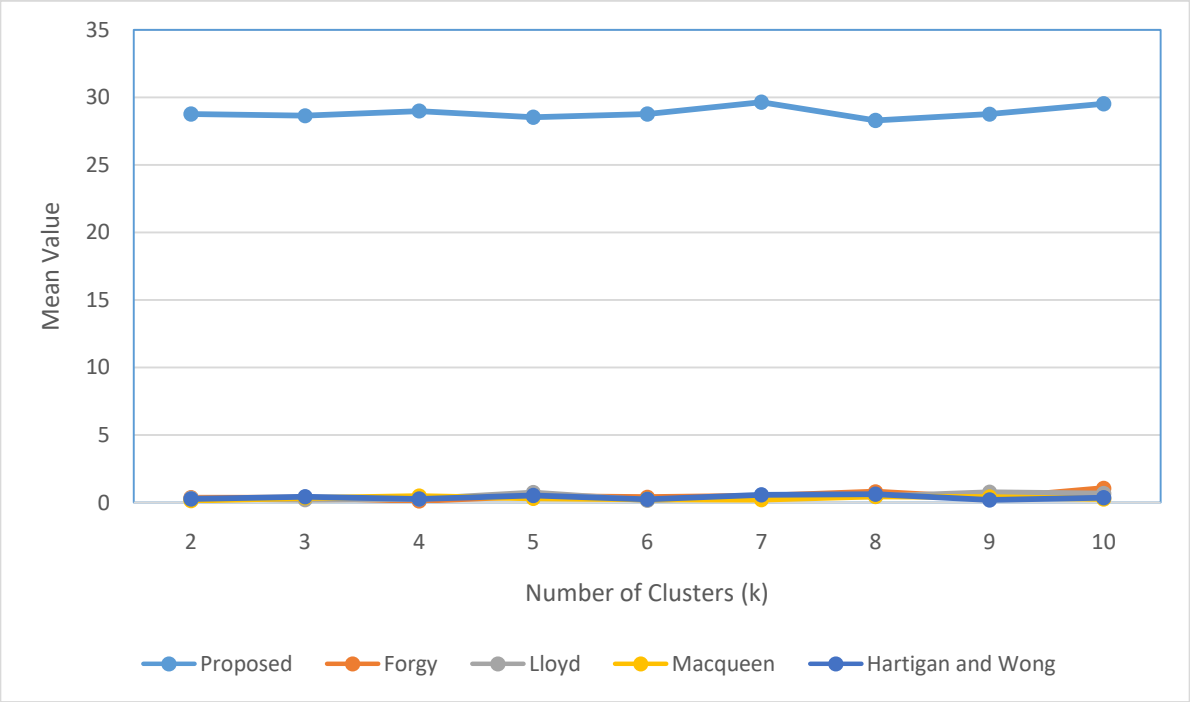

Figure 1. Mean Values of the Methods across the number of clusters (k)

Table 2. Result of the mean values of the clusters across the number of clusters

K	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
2	28.77195	0.36923	0.326951	0.146812	0.282513
3	28.64263	0.361316	0.227517	0.348543	0.438474
4	28.99411	0.132093	0.269607	0.483664	0.270218
5	28.5361	0.493879	0.727864	0.312476	0.528046
6	28.77214	0.405106	0.17424	0.253735	0.259336
7	29.64929	0.509191	0.496847	0.222172	0.572111
8	28.2929	0.799016	0.455969	0.454456	0.623866
9	28.7604	0.345777	0.762408	0.419515	0.205832
10	29.5355	1.066962	0.659568	0.27098	0.366181

Conclusion.

The findings of this study underscore the effectiveness and robustness of the proposed hybrid Generalized K-means clustering method compared to traditional approaches such as Forgy, Lloyd, MacQueen, Hartigan, and Wong. The findings demonstrated that the proposed method consistently achieves higher mean values across various data sets and numbers of clusters (k), signifying its superior intra-cluster homogeneity and inter-cluster separation. This trend is observed across diverse data sets, including the Iris data set, dat.axfors2021, dat.hartmannboyce2018, and the World Bank Commodity Price Data, highlighting the method's adaptability to data sets with varying dimensions and complexities. Unlike traditional methods, which exhibit erratic or non-linear trends in mean cluster values as k changes.The proposed method maintains stability and consistently outperforms its counterparts. This stability and superior performance suggest that the proposed method is less sensitive to variations in initial parameters and can effectively identify meaningful patterns and structures in data. The higher mean cluster values produced by the proposed method further confirm its capability to uncover interpretable and actionable clusters, making it a reliable tool for diverse clustering applications in both academic and practical contexts.

The findings provide compelling evidence that the hybrid Generalized K-means approach is a significant advancement in clustering methodologies. Its ability to deliver consistent and superior outcomes across various data sets positions it as a valuable tool for researchers and practitioners seeking to analyze complex data and derive meaningful insights. The study contributes to the broader discourse on improving clustering techniques, offering a method that bridges the gap between theoretical robustness and practical utility.

Based on the findings of this study, policymakers and decision-makers in fields reliant on data-driven insights are encouraged to adopt advanced clustering methods, such as the proposed hybrid Generalized K-Means, to enhance the accuracy and reliability of their analyses. By leveraging this method's demonstrated ability to produce stable and superior clustering outcomes across diverse data sets, policymakers can identify meaningful patterns, improve resource allocation, and develop targeted interventions. Furthermore, policymakers should invest in capacity-building initiatives to train analysts and researchers on advanced clustering techniques, ensuring the effective application of these tools in addressing complex societal and economic challenges.

However, despite its promising results, the study is not without limitations. The data sets analyzed, while diverse, may not comprehensively represent the full spectrum of real-world data scenarios. For instance, the data sets used such as Iris, dat.axfors2021, and World Bank Commodity Price Data are well-structured and relatively clean, which may not reflect the challenges posed by highly noisy, sparse, or unstructured data sets. Furthermore, the proposed method's performance in dynamic environments or streaming data contexts was not evaluated, which could limit its applicability in real-time clustering tasks. Future research should address these limitations by testing the method on more complex and heterogeneous data sets and exploring its adaptability to evolving data streams.

References.

1. Amorim, R. C. (2012). Constrained clustering with Minkowski weighted k-means. *Proceedings of the 13th IEEE International Symposium on Computational Intelligence* and Informatics, 13–17.
2. Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45, 1061–1075.
3. Bagirov, A. M., & Mardaneh, K. (2006). Modified global k-means algorithm for clustering in gene expression datasets. *Conference Proceedings Workshop on Intelligent Systems for Bioinformatics*, 73, 23–28.
4. El-Sharkawy, Y. H., Elbasuney, S., & Radwan, S. M. (2024). Non-invasive diffused reflected/transmitted signature accompanied with hyperspectral imaging for breast cancer early diagnosis. *Optics and Laser Technology*, 169.
5. Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, 4(1), 65–75.
6. Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22, 138–144.
7. Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21, 768–769.
8. Fox, E., Huang, H., & Raichel, B. (2024). Clustering with faulty centers. *Computational Geometry: Theory and Applications*, 117.
9. Hao, L., Wang, T., & Guo, C. (2023). Research on parallel association rule mining of big data based on an improved K-means clustering algorithm. *International Journal of Autonomous and Adaptive Communications Systems*, 16(3), 233–247.
10. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108.
11. Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5), 657–668.
12. Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany*, 14(1), 127–130.
13. Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., & Wu, A. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881–892.
14. Kim, S., Cho, S., Kim, J. Y., & Kim, D. J. (2023). Statistical assessment on student engagement in asynchronous online learning using the k-means clustering algorithm. *Sustainability (Switzerland)*, 15(3).
15. Kotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210.
16. Likas, A., Vlassis, N., & Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2), 451–461.
17. Liu, J., Qiu, Z., Gao, M., & Yu, D. (2023). Improved K means Clustering Algorithm Based on Tukey Rule and Initial Center Point Optimization. *Shuju Caiji Yu Chuli/Journal of Data Acquisition and Processing*, 38(3), 643–651.
18. Mahmud, M. S., Huang, J. Z., & García, S. (2024). Clustering approximation via a fusion of multiple random samples. *Information Fusion*, 101.
19. Obaid, O. I. (2023). Analysis of H-index and Papers Citation in Computer Science Field using K-Means Clustering Algorithm. *Iraqi Journal for Computer Science and Mathematics*, 4(2).
20. Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA[Formula presented]. *Expert Systems with Applications*, 237.
21. Sujatha, S. and Sona, A. S. (2013). New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method.*International Journal of Engineering Research & Technology (IJERT)*, 2(2): 1-9.
22. Torra, V., & Narukawa, Y. (2012). On a comparison between Mahalanobis distance and Choquet integral: The Choquet–Mahalanobis operator. *Information Sciences*, 190, 56-63.
23. Vishwakarma, S., Bhardwaj, S. K., Bihari, A., Tripathi, S., Agrawal, S., & Joshi, P. (2024). Cancer Gene Clustering Using Computational Model. *GMSARN International Journal*, 18(2), 252–257.
24. Wang, S., Song, A., & Qian, Y. (2023). Predicting Smart Cities' Electricity Demands Using K-Means Clustering Algorithm in Smart Grid. *Computer Science and Information Systems*, 20(2), 657–678.
25. Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. *Multidisciplinary, Scientific Journal*, 2019, 2(2), 226-235.