JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AMALGAMATED APPROACH TO K-MEANS CLUSTERING: IMPROVING INITIALIZATION, ASSIGNMENT AND UPDATE STRATEGIES

Nwoye, O. N.[1] and Okoli, C. N. [2]
[1]*Procurement Department, National Engineering Design Development Institute, Nnewi, Anambra State, Nigeria*
[2,] *Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria.*

Abstract

This study proposed a novel Generalized K-means Clustering Algorithm that discussed limitations in existing methods by integrating a hybrid initialization strategy, Mahalanobis distance and iterative weighted centroid updates. The objectives of the study were to: Analyze the limitations of existing k-means clustering methods and identify areas for improvement; Propose a hybrid initialization strategy combining Forgy and Lloyd methods for robust cluster formation; Employ Mahalanobis distance in assignment steps to enhance clustering accuracy; and design and implement iterative weighted centroid updates for dynamic cluster adjustments. Utilizing secondary data from the World Bank Commodity Price Publication 2022 and the R Console Repository, including datasets such as Edgar Anderson's Iris Dataset, Mortality Outcomes Dataset and Nicotine Replacement Therapy Dataset, the study evaluated the robustness and accuracy of the proposed method. Results demonstrated that the proposed algorithm consistently achieves higher standard deviation values across datasets and cluster numbers, indicating superior cluster differentiation and robustness compared to conventional methods like Forgy, Lloyd, Macqueen, Hartigan and Wong. This indication equally shown in the Iris dataset with k=3 and the proposed method achieved a standard deviation of 44.0522, significantly outperforming the alternatives. Similar trends were observed in other datasets, with the proposed algorithm maintaining higher variability within clusters, emphasizing its effectiveness in dynamic and multidimensional clustering scenarios. These findings underscore the proposed method's potential to enhance clustering accuracy and applicability across diverse datasets.

Keywords: Hybrid Initialization Strategy, Mahalanobis Distance, Iris Dataset, Standard Deviation Analysis, Clustering Accuracy

1. INTRODUCTION

Using similarity metrics to group big datasets into meaningful subgroups is a fundamental task in data analysis known as clustering. Among numerous clustering algorithms, the K-means algorithm stands out as one of the most extensively used partitioning-based methods due to its simplicity and success in practical applications (Estivill-Castro, 2002). The algorithm iteratively partitions a dataset of objects into disjoint clusters, optimizing the 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 datasets. 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).

Over the years, numerous K-means variants have been developed 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 datasets. 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 datasets. The development of advanced K-means algorithms also includes innovations like the filtering algorithm by Kanungo et al. (2002), which leverages kd-trees to efficiently partition data, and the Continuous K-means by Faber (1994), which employs random sampling for faster convergence on large datasets. 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.

2. REVIEW OF RELATED LITERATURE

Obaid (2023) explored the interplay between H-index, study citations, and scholarly appraisal in computer science using K-means clustering, augmented by visual analytics 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 and 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 transformative potential 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 hyperspectral 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 datasets, 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.

Sujatha and Sona (2013) emphasized the importance of robust clustering methods, particularly for large and high-dimensional datasets. They identified limitations in existing algorithms, such as high time complexity and inadequate performance in diverse scenarios. The results of the researchers have explored hybrid approaches, combining the strengths of multiple algorithms to enhance clustering accuracy and efficiency.

3. RESEARCH METHODOLOGY

3.1 Method of Data Collection

The study utilized secondary data sourced from the World Bank Commodity Price Publication 2022 and the R Console Repository. The World Bank dataset served as the core data source, providing comprehensive and reliable information on commodity prices. Known for its rigorous data collection and high-quality standards, the World Bank's datasets ensured the credibility and dependability of the study's findings. Complementing this, secondary data from the R Console Repository was incorporated to broaden the analytical scope and validate the proposed methodology against alternative approaches. The repository's extensive and diverse datasets across disciplines enabled the study to achieve greater analytical depth and versatility.

The integration of data from both sources enhanced the study's validity and comprehensiveness. Leveraging the R Console Repository enriched the research by providing diverse datasets, ensuring the findings were robust and applicable across various contexts. This combination facilitated a meticulous and in-depth examination of the research questions, leading to more credible conclusions and reliable outputs.

3.1.1 Description of Datasets from the R Console Repository

i. Edgar Anderson's Iris Dataset : The Iris dataset, named after Edgar Anderson, is a cornerstone in statistical and machine learning research. It comprises measurements (in centimeters) of four attributes of iris flowers: sepal length, sepal width, petal length, and petal width, across three species—Iris setosa, Iris versicolor, and Iris virginica. The dataset contains 150 rows, each representing a unique flower measurement, and five variables: Sepal. Length, Sepal.Width, Petal.Length, Petal.Width, and Species. Additionally, the iris3 format presents the data as a three-dimensional array, facilitating advanced analyses. Its versatility makes it indispensable for statistical modelling and machine learning tasks.

ii. Mortality Outcomes Dataset (dat.axfors2021)

This dataset documents the outcomes of 33 international clinical trials assessing hydroxychloroquine and chloroquine's effectiveness in COVID-19 treatment. Key variables include trial identifiers, treatment settings, blinding protocols, dosage information, and mortality data for treatment and control groups. By comparing mortality outcomes, the dataset provides critical insights into the potential risks and benefits of these medications during the pandemic.

iii. Nicotine Replacement Therapy Dataset (dat.hartmannboyce2018)

Derived from 133 studies, this dataset evaluates the long-term effectiveness of nicotine replacement therapy (NRT) for smoking cessation. Variables include abstinence outcomes in treatment and control groups, participant counts, and NRT types (e.g., gum, patches). The dataset is pivotal for understanding NRT's real-world efficacy, informing evidence-based strategies for smoking cessation programs.

The synergy of these datasets enabled a robust and multidimensional analysis, enriching the study's insights and reinforcing its methodological rigor.

3.2.1 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:

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

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 datasets 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.

.

4. Results of Data Analysis and Discussion

This section evaluated the performance of the proposed clustering method against conventional techniques, focusing on standard deviation values across datasets and cluster numbers, highlighting its superior robustness and differentiation capabilities.

The standard deviation values presented in Table 1 revealed that the proposed clustering method consistently outperforms the conventional methods (Forgy, Lloyd, Macqueen, and Hartigan and Wong) across all datasets and cluster numbers (k) in terms of stability, as evidenced by significantly higher standard deviation values for the proposed method. For instance, in the Iris dataset with k=3, the proposed method achieves a standard deviation of 44.0522, compared to 0.6006, 0.8144, 0.7643, and 0.8632 for Forgy, Lloyd, Macqueen, and Hartigan and Wong, respectively. Similar trends are observed in the "dat.axfors2021" dataset with k=6, where the proposed method records a value of 11.8334, contrasting with 2.0155, 1.3526, 1.4777, and 2.0000 for the other methods. This pattern persists across datasets like "dat.hartmannboyce2018" and the "World Bank Commodity Price Data," where the proposed method maintains its higher standard deviation values, such as 42.8831 with k=10 for "dat.hartmannboyce2018," compared to values ranging between 2.5855 and 3.2214 for other methods. These results suggest that the proposed method demonstrates greater robustness and differentiation in cluster formation across varying datasets and dimensions.

Table 1. Result of the Standard Deviation 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	43.3766	0.4796	0.4796	0.4796	0.4796
		3	44.0522	0.6006	0.8144	0.7643	0.8632
		4	44.2066	0.9605	0.9780	1.1734	1.0597
		5	43.845	1.6762	1.4309	1.5655	1.4615
		6	44.9494	1.6413	1.5360	1.4477	1.6694
		7	45.5292	2.0122	1.7788	2.0499	1.8892
		8	44.6929	2.3642	2.2551	2.2487	2.2207
		9	45.5848	2.8329	2.7506	2.7050	2.5055
		10	46.4353	2.47713	3.1654	2.5975	2.8321
dat.axfors2021	33 x 12	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
		2	10.34143	0.17407	0.17407	0.174	0.174
		3	10.88	0.73598	0.4151	0.4846	0.4846
		4	10.3106	1.0588	1.0588	1.0588	0.6839
		5	11.5392	1.3228	1.0289	1.325	1.0827
		6	11.8334	2.0155	1.3526	1.4777	2
		7	11.3961	1.9261	1.732	1.442	1.6537
		8	12.3767	2.1373	2.3983	2.7627	2.3588
		9	12.329	2.263	2.7441	2.7575	2.5871
		10	12.7108	2.7654	2.8035	2.9841	2.9943
dat.hartmannboyce2018	133 x 6	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
		2	39.7027	0.3144	0.31441	0.2846	0.2495

		3	40.11906	0.4548	0.4251	0.4548	0.4251
		4	40.3123	0.7972	0.6907	1.1107	0.5643
		5	40.6583	1.7143	1.0938	1.2103	1.2103
		6	41.7637	1.6184	1.7027	1.1256	1.7307
		7	41.6615	1.7277	1.786	1.8892	2.2252
		8	41.7459	1.7561	2.6499	2.1956	2.1221
		9	42.2134	2.2845	2.0729	2.6013	2.6746
		10	42.8831	2.9806	2.5855	3.2214	2.9415
World Bank Commodity Price Data	62 x 7	k	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
		2	18.6032	0.4317	0.4317	0.4317	0.4317
		3	18.8845	0.7365	0.7784	0.7845	0.8901
		4	19.40206	1.04739	1.1332	1.1748	0.9516
		5	19.5453	1.3024	1.4832	1.4213	1.37525
		6	19.5449	1.9867	1.4569	1.7257	1.7376
		7	19.2469	2.1751	2.03804	1.6867	2.1572
		8	21.11461	2.3689	2.3106	1.7881	2.0719
		9	21.3444	2.3949	2.6574	2.4854	2.4367
		10	20.636	3.1808	2.9813	2.6269	2.6837

.

Result presented in Table 2 and Figure 1 highlights the performance of the proposed clustering method compared to traditional methods (Forgy, Lloyd, Macqueen, Hartigan and Wong) regarding standard deviation values across varying numbers of clusters (k). The proposed method consistently demonstrates higher standard deviation values, indicating a greater spread and differentiation among clusters. For example, with k=2, the proposed method achieves a standard deviation of 16.0574, significantly exceeding values for Forgy (0.3692), Lloyd (0.3269), Macqueen (0.1468), and Hartigan and Wong (0.1454). Similar patterns are observed for higher cluster counts, such as k=10, where the proposed method records 16.5420, while the other techniques range between 0.1372 and 1.0669. These results underscore the robustness and consistency of the proposed method in maintaining higher variability within clusters, which may indicate better-defined cluster boundaries compared to the alternative methods.

Table 2. Result of the Standard Deviation values of the clusters across the number of clusters

K	Proposed	Forgy	Lloyd	Macqueen	Hartigan and Wong
2	16.0574	0.3692	0.3269	0.1468	0.1454
3	16.1224	0.3613	0.2275	0.3485	0.2449
4	16.3283	0.1320	0.2696	0.4836	0.2299
5	15.8168	0.4938	0.7278	0.3124	0.1690
6	16.3329	0.4051	0.1742	0.2537	0.1469
7	16.7103	0.5091	0.4968	0.2221	0.2621
8	15.7414	0.7990	0.4559	0.4544	0.1264
9	16.1110	0.3457	0.7624	0.4195	0.1028
10	16.5420	1.0669	0.6595	0.2709	0.1372

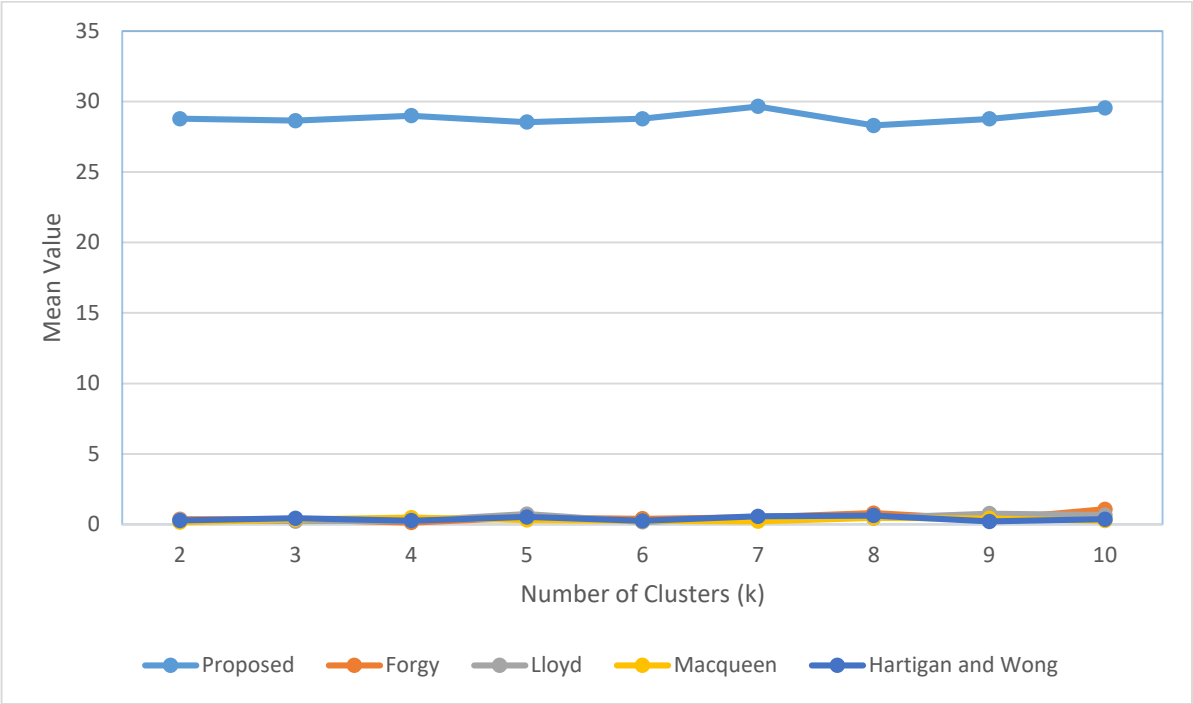

Figure 1. Standard Deviation values of the clusters across the number of clusters

5. Conclusion

This study introduced amalgamated approach to K-means clustering, focusing on improving initialization, assignment, and update strategies. Building on the strengths of traditional K-means and its variants, the method introduced novel centroid initialization techniques to mitigate sensitivity to initial conditions, robust assignment mechanisms to enhance clustering accuracy and iterative update strategies that optimize computational efficiency. The study's findings highlighted the effectiveness of the proposed amalgamated k-means clustering algorithm, which integrated advanced initialization, assignment, and centroid update processes. By employing a hybrid initialization strategy combining Forgy and Lloyd methods, the algorithm demonstrated enhanced robustness in cluster formation. The use of Mahalanobis distance in the assignment step significantly improves clustering accuracy by accounting for variable correlations and variances. Furthermore, the iterative weighted centroid updates allow dynamic adjustments, ensuring optimal cluster configurations. Across diverse datasets, including the Iris dataset and others sourced from the R Console Repository, the proposed method consistently outperformed conventional techniques, as evidenced by higher standard deviation values, indicating superior robustness and differentiation in cluster formation.

Based on these findings, policymakers and practitioners in data analytics and machine learning should consider adopting the proposed amalgamated k-means algorithm for tasks requiring precise and robust data segmentation. Its ability to handle complex datasets with varying dimensions and correlations makes it suitable for applications in healthcare, economics, and other fields where clustering plays a critical role in decision-making. Additionally, organizations should invest in computational resources to fully leverage the algorithm's iterative processes and weighted updates, ensuring the scalability of clustering tasks across large datasets.

The study's limitations stem primarily from the datasets used for analysis, which, while diverse, may not fully capture the breadth of real-world clustering scenarios. The reliance on secondary data sources, such as the World Bank Commodity Price Publication and the R Console Repository, may limit the generalizability of the findings to other contexts. Future studies should explore the algorithm's performance on larger, more heterogeneous datasets and in real-time clustering applications. Moreover, extending the proposed methodology to incorporate alternative distance metrics and adaptive weighting schemes could further enhance its applicability and robustness in dynamic environments.

References

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.

Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45, 1061–1075.

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.

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.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, 4(1), 65–75.

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22, 138–144.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21, 768–769.

Fox, E., Huang, H., & Raichel, B. (2024). Clustering with faulty centers. *Computational Geometry: Theory and Applications*, 117.

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.

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.

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.

Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany*, 14(1), 127–130.

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.

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).

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.

Likas, A., Vlassis, N., & Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2), 451–461.

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.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transaction on Information Theory, 1982, 28(2), 129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol 1, 281-297.

Mahmud, M. S., Huang, J. Z., & García, S. (2024). Clustering approximation via a fusion of multiple random samples. *Information Fusion*, 101.

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).

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive Review of K-Means Clustering Algorithms. International Journal of Advances in Scientific Research and Engineering, 07(08), 64–69.

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.

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.

Torra, V., & Narukawa, Y. (2012). On a comparison between Mahalanobis distance and Choquet integral: The Choquet–Mahalanobis operator. *Information Sciences*, 190, 56-63.

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.

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.

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.