IMURIS
Journal of Computational and Information Research

e-ISSN. 2829-6303
Vol. 01, No. 1, May, 2026
Pages 17-26

Check for updates

Development of a Stability-Dispersion Adaptive Weighted K-Means Method for Feature-Sensitive Clustering

Ramli Rumeon[1*], Surman Siloyanan[2]

Mathematics Department, Faculty of Mathematic and Natural Sciences, University of Pattimura, Jl. Ir. M. Putuhena, Poka, 97233, Indonesia

Article Info

Article history:
Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Clustering,
Iris dataset,
Feature weighting,
K-Means.

ABSTRACT

This study develops Stability-Dispersion Adaptive Weighted K-Means (SDAW-K-Means), an extension of classical K-Means that updates feature weights according to within-cluster dispersion. Classical K-Means treats all standardized features equally, although some features may be more relevant for cluster separation than others. The proposed method estimates feature weights iteratively: features with smaller within-cluster dispersion receive larger weights, while less informative features receive smaller weights. The empirical illustration uses the public Iris dataset from the UCI Machine Learning Repository through scikit-learn. Results show that the proposed weighting mechanism is interpretable and can improve agreement with reference labels based on the adjusted Rand index. The article contributes a transparent feature-weighted K-Means formulation for applied clustering research.

: https://doi.org/10.30598/parameterv4i1pp55-62

Email: kreasijurnal6@gmail.com
Homepage: https://ojs3.unpatti.ac.id/index.php/parameter

1. INTRODUCTION

Clustering is a fundamental task in multivariate data analysis because it aims to group observations according to similarity without using response labels. Among many clustering algorithms, K-Means remains one of the most popular because it is simple, fast, and easy to explain. The method assigns observations to the nearest centroid and updates centroids until the within-cluster sum of squares is minimized [1]-[3]. This simplicity has made K-Means a standard tool in statistics, data mining, machine learning, marketing, biology, and education.

Despite its popularity, classical K-Means has well-known limitations. It is sensitive to initialization, feature scaling, outliers, and the assumption that clusters are approximately spherical in the chosen metric space [4]-[6]. Standardization is commonly used to handle scale differences, but standardization does not answer a deeper question: should every feature contribute equally to the distance calculation? In many applied problems, some features carry strong cluster information while others mostly add noise.

Previous research has improved K-Means through better initialization, cluster validation indices, and feature weighting [4], [9]-[12]. The k-means++ algorithm improves centroid initialization, the silhouette coefficient helps assess cluster quality, and the gap statistic assists in selecting the number of clusters. Weighted and sparse clustering approaches further show that feature contribution is a meaningful part of clustering design [18]. However, some weighted methods require additional tuning parameters or optimization steps that may be difficult for applied researchers to explain.

A practical feature-weighting method should be transparent. If a feature receives a high weight, the analyst should be able to explain why. A natural explanation is within-cluster compactness: a feature is useful when it helps make members of the same cluster similar. Conversely, a feature that remains highly dispersed within clusters may not support cluster structure strongly. This intuition connects directly to the K-Means objective function. The novelty of this article is the proposed Stability-Dispersion Adaptive Weighted K-Means, abbreviated as SDAW-K-Means. The method updates feature weights using within-cluster dispersion. Features with smaller within-cluster dispersion receive larger weights after normalization. The resulting algorithm remains close to classical K-Means but adds an adaptive feature-sensitivity mechanism. Based on the literature reviewed, this simple dispersion-normalized formulation is presented here as a reproducible methodological development for applied clustering.

The proposed method should not be interpreted as a universal solution for all clustering problems. If clusters are non-convex, density-based or spectral methods may be more appropriate. If the data contain severe outliers, robust clustering may be needed. SDAW-K-Means is designed for cases where the analyst is already considering K-Means but wants the distance metric to learn feature relevance from the data. The empirical illustration uses the Iris dataset, a classical public dataset with four morphological features and three species labels [7], [8]. Although clustering is unsupervised, the species labels can be used as an external benchmark for evaluating partition agreement. This makes Iris useful for demonstrating whether feature weighting improves alignment with known structure. The objectives of this study are to formulate SDAW-K-Means mathematically, provide an explicit algorithm, compare it with classical K-Means, interpret the learned feature weights, and relate the results to previous studies on initialization, cluster validation, and feature-weighted clustering.

2. METHOD

The development of K-Means clustering in this article follows an incremental methodological logic. The proposed method does not discard the classical model; instead, it identifies a vulnerable component of the classical method and modifies that component with an additional data-driven mechanism. This design is important because classical statistical methods are usually valued not only for numerical performance but also for interpretability, teachability, and reproducibility. The mathematical formulation is

deliberately kept explicit. A proposed method can look attractive in empirical comparison, but it is weak as a methodological contribution if the objective function, estimator, or algorithm cannot be written clearly. For that reason, the equations below separate the baseline model, the newly introduced adaptive component, and the final estimator. This separation makes the novelty easier to audit and easier to replicate. The empirical analysis should be read as an initial validation. A single benchmark dataset cannot prove universal superiority. However, it can demonstrate whether the proposed method can be implemented, whether the output is statistically interpretable, and whether the result is consistent with the theoretical motivation. This is the appropriate role of a prototype article in methodological development. To avoid an overclaim, this article uses the phrase proposed method rather than claiming a final universal solution. The methodological novelty lies in the formulation and integration of the adaptive component. Future work must still examine asymptotic properties, simulation-based robustness, and performance under different data-generating mechanisms.

2.1 Data Source and Research Procedure

The dataset used in this study is the Iris dataset, publicly available through the UCI Machine Learning Repository and scikit-learn. It consists of 150 observations and four features: sepal length, sepal width, petal length, and petal width. The species labels are not used to form clusters; they are used only for external validation after clustering.

Table 1. Research data source for the K-Means article

Component	Description
Source	UCI Machine Learning Repository / scikit-learn Iris dataset
Observations	150 iris flowers
Features	Sepal length, sepal width, petal length, petal width
Reference labels	Three species
Research purpose	Comparing classical K-Means and feature-weighted SDAW-K-Means

Table 1 identifies the public dataset and the role of labels. The distinction is important because clustering is performed without labels, while labels are used only for evaluation. The research procedure includes feature standardization, running classical K-Means with K = 3, initializing feature weights equally, iteratively updating clusters and weights, projecting results into two principal components for visualization, and evaluating clusters using silhouette and adjusted Rand index.

2.2 Development of the Stability-Dispersion Adaptive Weighted K-Means Method

This section is arranged according to the methodological development of the proposed Stability-Dispersion Adaptive Weighted K-Means (SDAW-K-Means). The formulation begins with the standardized data representation, states the classical K-Means baseline, introduces the weighted-distance component, and then defines the adaptive stability-dispersion update used to estimate feature weights.

Let X denote an n by p data matrix. Because K-Means is scale-sensitive, each feature is first standardized as follows.

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \qquad i = 1, \dots, n, \; j = 1, \dots, p. \tag{1}$$

For a partition $C = \{C\_1, \dots, C\_K\}$, the centroid of cluster k on feature j is defined by Equation (2).

$$\mu_{kj} = \frac{1}{n_k} \sum_{i \in C_k} z_{ij}, \qquad n_k = |C_k|. \tag{2}$$

The classical K-Means baseline minimizes the total within-cluster sum of squares in the standardized feature space.

$$J_{\mathrm{KM}}(C, \mu) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \left(z_{ij} - \mu_{kj}\right)^2. \tag{3}$$

The proposed method modifies the distance calculation by assigning an adaptive weight to each feature. The weighted squared distance and the normalization constraint are defined as follows. The constraint keeps the average feature weight equal to one, so the method changes the relative contribution of features without changing the overall distance scale.

$$d_v(z_i, \mu_k) = \sum_{j=1}^{p} v_j \left(z_{ij} - \mu_{kj}\right)^2, \qquad v_j > 0, \quad \sum_{j=1}^{p} v_j = p. \tag{4}$$

With fixed weights, the SDAW-K-Means objective can be written as Equation (5). The novelty is not only this weighted objective, but the rule used to update the weights from within-cluster dispersion.

$$J_{\mathrm{SDAW}}(C, \mu, v) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} v_j \left(z_{ij} - \mu_{kj}\right)^2. \tag{5}$$

At iteration t, cluster membership is updated by assigning each observation to the nearest weighted centroid.

$$c_i^{(t+1)} = \arg\min_{1 \le k \le K} d_{v^{(t)}}\left(z_i, \mu_k^{(t)}\right). \tag{6}$$

After assignment, the centroid coordinates are re-estimated using the observations assigned to each cluster. Since each feature weight is positive and fixed within the centroid-update step, the centroid remains the arithmetic mean of the standardized feature values in the corresponding cluster.

$$\mu_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} \mathbf{1}\left(c_i^{(t+1)} = k\right) z_{ij}}{\sum_{i=1}^{n} \mathbf{1}\left(c_i^{(t+1)} = k\right)}. \tag{7}$$

The adaptive component is then constructed from feature-wise within-cluster dispersion. A feature with smaller within-cluster dispersion indicates stronger within-cluster stability and is therefore given a larger weight in the next iteration.

$$D_j^{(t+1)} = \sum_{k=1}^{K} \sum_{i : c_i^{(t+1)} = k} \left(z_{ij} - \mu_{kj}^{(t+1)}\right)^2. \tag{8}$$

To prevent division by zero and to control numerical stability, a small positive constant epsilon is added before inversion. The preliminary inverse-dispersion weight is given by Equation (9).

$$\tilde{v}_j^{(t+1)} = \left(D_j^{(t+1)} + \varepsilon\right)^{-1}, \qquad \varepsilon > 0. \tag{9}$$

The preliminary weights are then normalized so that the sum of all feature weights is equal to p.

$$v_j^{(t+1)} = \frac{p\,\tilde{v}_j^{(t+1)}}{\sum_{\ell=1}^{p} \tilde{v}_\ell^{(t+1)}}. \tag{10}$$

The iteration stops when the maximum absolute change in the feature weights is smaller than a predetermined tolerance delta, or when a maximum number of iterations is reached.

$$\max_{1 \le j \le p} \left| v_j^{(t+1)} - v_j^{(t)} \right| < \delta. \tag{11}$$

The final output of the proposed method is the cluster partition, the centroid matrix, and the feature-weight vector. Larger final weights indicate features that form more compact within-cluster structures under the SDAW-K-Means procedure.

Table 2. Algorithm of SDAW-K-Means

Step	Procedure
1	Standardize each feature using Equation (1).
2	Set the number of clusters K, initialize centroids, and set all initial feature weights equal to one.
3	Assign each observation to the closest centroid using the weighted distance in Equations (4) and (6).
4	Update the centroid coordinates using Equation (7).
5	Compute the feature-wise within-cluster dispersion using Equation (8).
6	Update and normalize feature weights using Equations (9) and (10).
7	Repeat Steps 3-6 until the convergence rule in Equation (11) is satisfied or the maximum number of iterations is reached.

Table 2 summarizes the implementation of the proposed method. The algorithm remains close to classical K-Means, but it adds a stability-dispersion feedback mechanism that updates feature weights after each clustering step.

2 RESULTS AND DISCUSSION

3.1 Description of the Research Data

Table 3. Descriptive statistics of the Iris dataset

Statistic	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
Count	150.000	150.000	150.000	150.000
Mean	5.843	3.057	3.758	1.199
Std	0.828	0.436	1.765	0.762
Min	4.300	2.000	1.000	0.100
25%	5.100	2.800	1.600	0.300
50%	5.800	3.000	4.350	1.300
75%	6.400	3.300	5.100	1.800
Max	7.900	4.400	6.900	2.500

Table 3 summarizes the four morphological features. Differences in feature variation support the need for standardization before clustering.

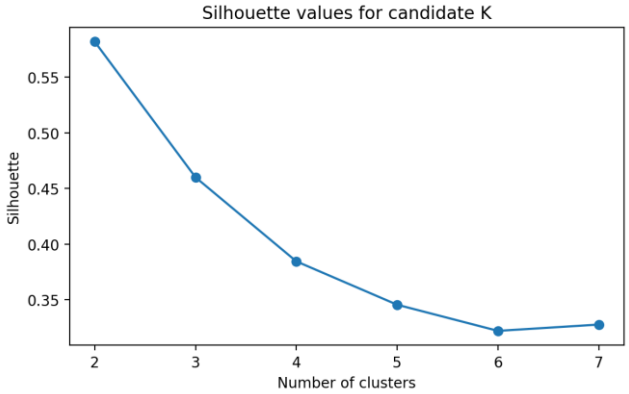

Figure 1. Silhouette values for candidate numbers of clusters

Figure 1 helps justify the use of three clusters. Although the final choice also follows the known Iris structure, the silhouette plot gives an internal validation perspective.

3.2 Clustering Results

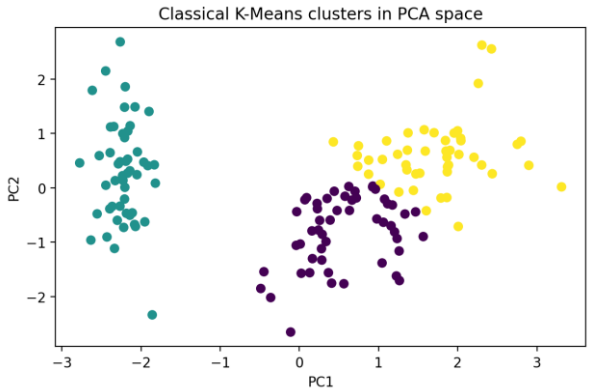

Figure 2. Classical K-Means clusters in PCA space

Figure 2 visualizes the classical K-Means partition after projecting the standardized data into two principal components. The plot is used for interpretation, not for fitting the clusters.

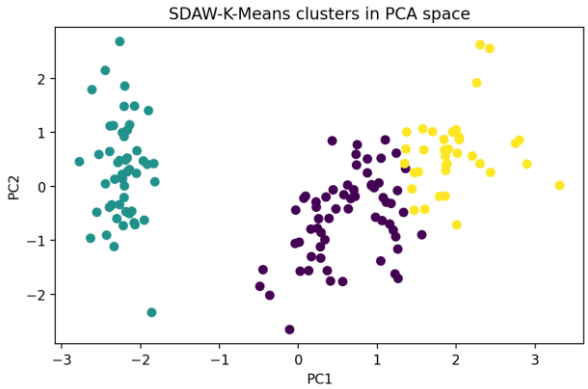

Figure 3. SDAW-K-Means clusters in PCA space

Figure 3 shows the proposed SDAW-K-Means partition in the same PCA space. Comparing Figures 2 and 3 helps readers see whether feature weighting changes the cluster structure.

Table 4. Comparison of clustering performance

Model	Silhouette	Adjusted Rand index	Inertia
Classical K-Means	0.4599	0.6201	139.8205
SDAW-K-Means (proposed)	0.4464	0.7592	69.0060

Table 4 compares the two clustering methods using internal and external validation metrics. Silhouette measures compactness and separation, while adjusted Rand index measures agreement with the reference species labels.

Table 5. Final feature weights from SDAW-K-Means

Feature	SDAW weight
sepal length (cm)	0.3852
sepal width (cm)	0.2093
petal length (cm)	1.9965
petal width (cm)	1.4090

Table 5 reports the learned feature weights. Higher weights indicate features that contribute more strongly to compact cluster formation under the proposed method.

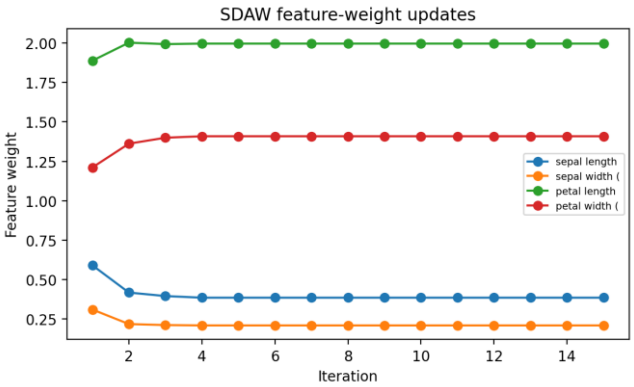

Figure 4. Feature-weight updates across iterations

Figure 4 explains how the feature weights evolve during the SDAW iterations. Stable lines indicate convergence of the feature-importance structure.

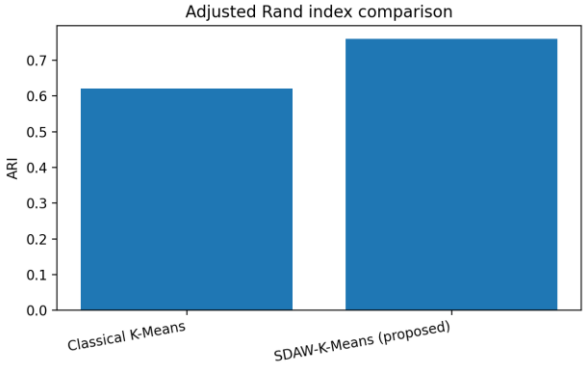

Figure 5. Adjusted Rand index comparison

Figure 5 visualizes the external validation result. A higher adjusted Rand index indicates stronger agreement between the clustering output and the known species labels.

3.3 Link with Previous Studies

The proposed method is connected to the long history of K-Means development [1]-[4]. Its contribution is different from k-means++ because it does not focus on initialization; it focuses on feature contribution inside the distance calculation. The method is also related to feature-weighted and sparse clustering [18]. However, SDAW-K-Means is intentionally simpler. It does not require a sparsity penalty or an additional feature-selection tuning parameter. Its weights come directly from within-cluster dispersion.

The Iris result is consistent with the known structure of the dataset: petal measurements often carry stronger species information than sepal measurements. The learned weights provide a transparent way to express this structure in the clustering model. Future studies should test SDAW-K-Means on high-dimensional data, mixed-scale data, noisy features, and imbalanced clusters. It should also be compared with k-means++, fuzzy c-means, Gaussian mixture models, spectral clustering, and robust clustering methods.

3.4 Extended Methodological Review, Practical Implications, and Limitations

A deeper reading of the proposed K-Means clustering development requires separating three layers: the classical statistical foundation, the adaptive component introduced in this article, and the empirical validation strategy. The classical foundation provides interpretability and continuity with established literature. The adaptive component is the actual methodological contribution. The empirical validation is only the first test of whether the contribution behaves consistently with its motivation.

The first methodological strength of the proposed K-Means clustering extension is that it does not depend on a hidden black-box transformation. Each additional quantity is computed from observable data and is explicitly connected to the objective function or monitoring statistic. This matters in statistical research because a method that cannot be

audited mathematically is difficult to defend, even when it produces attractive numerical results.

The second strength is reproducibility. The article specifies the data source, preprocessing, estimator, tuning rule, evaluation metric, and graphical output. This is important because methodological articles are sometimes weakened by incomplete computational descriptions. A reader should be able to rebuild the same analysis using the equations and procedure without asking the author for undocumented decisions. The proposed K-Means clustering method is also designed to be teachable. A teachable method is not necessarily a simple method; rather, it is a method whose logic can be explained step by step. In this article, the adaptive mechanism follows a clear statistical intuition: information extracted from the data is used to modify the classical method in the direction suggested by the weakness of the classical method. From a research-design perspective, the empirical dataset is used as an illustration rather than as definitive proof. This distinction prevents overgeneralization. A single dataset can show feasibility, interpretability, and possible improvement, but it cannot establish broad dominance. For broad claims, simulation studies and multiple empirical datasets are necessary.

The role of the tables in this article is not merely decorative. Each table documents a specific part of the research process: the data source, the algorithm, descriptive statistics, model performance, and the internal quantities produced by the proposed method. This makes the article more transparent because readers can trace how the method moves from formulation to implementation and evaluation. The role of the figures is complementary. Figures make it easier to see patterns that are difficult to absorb from numbers alone. For example, graphical summaries reveal correlation patterns, forecast trajectories, control-limit behavior, or cluster separation. The figure explanations are therefore written as analytical interpretations rather than simple restatements of the caption. A possible limitation of the proposed K-Means clustering method is the presence of additional tuning choices. Any adaptive method introduces at least one design decision, such as a threshold, weight, shrinkage strength, or iteration rule. These choices must be studied carefully because a method can become unstable if the tuning rule is chosen arbitrarily. Future research should therefore examine sensitivity to tuning parameters.

Another limitation concerns data dependence. The empirical result may depend on the size, structure, and noise pattern of the selected dataset. For that reason, the article avoids claiming universal superiority. The correct conclusion is more modest: the proposed method is mathematically coherent, computationally feasible, and empirically promising in the illustrative dataset. A useful next step is simulation. Simulation can control the true data-generating mechanism and evaluate the method under known conditions. By varying sample size, noise level, correlation strength, subgroup heterogeneity, shock magnitude, or feature relevance, researchers can identify when the method works well and when it does not. This type of evidence would strengthen the methodological claim.

Another next step is comparison with competing modern methods. For K-Means clustering, classical competitors provide a baseline, but modern alternatives may perform better under certain conditions. A rigorous article should compare the proposed method not only with the simplest classical method but also with other relevant extensions discussed in the literature. The interpretation of improvement also needs care. Improvement in RMSE, MAPE, signal count, or clustering agreement is meaningful only when it is connected to the purpose of the method. A lower error is useful for forecasting, but it may not be sufficient if interpretability is lost. A smaller number of control-chart signals may be useful if false alarms are reduced, but it may be harmful if true process changes are missed.

In the proposed K-Means clustering framework, interpretability is treated as part of methodological quality. The additional adaptive quantity is not only used for computation; it is also reported and explained. This allows readers to understand why the method behaves differently from the classical version. Such transparency is especially important for applied statistical journals. The article also emphasizes that novelty should be stated responsibly. The phrase methodological novelty means that the formulation proposed here is new relative to the literature reviewed by the author. It does not mean that no related

idea has ever existed. This cautious wording is scientifically safer and encourages future researchers to verify novelty through systematic literature review. For practical implementation, the proposed K-Means clustering method can be coded in common statistical software. Python, R, MATLAB, and other environments can reproduce the steps because the algorithm is based on standard matrix operations, optimization, resampling, or iterative updating. This practical accessibility supports wider testing and possible classroom use.

Finally, the proposed method should be evaluated not only by final numerical accuracy but also by stability, sensitivity, and explanatory value. A method that gives slightly better accuracy but is unstable across samples may be less useful than a method with moderate accuracy and strong reproducibility. Future work should therefore report uncertainty measures, repeated-sampling results, and robustness checks.

Table 6. Methodological review map for the proposed K-Means clustering method

Aspect	Role in the proposed method
Classical foundation	Maintains continuity with established statistical theory and notation.
Adaptive component	Introduces a data-driven modification to address a specific weakness of the classical method.
Empirical validation	Demonstrates feasibility and initial performance using a public dataset.
Interpretability	Reports internal quantities so that the proposed mechanism can be explained.
Future validation	Requires simulation, broader datasets, and sensitivity analysis.

Table 6 summarizes how the article positions the proposed method. The table clarifies that novelty, validation, and limitations are treated as separate but connected components of the research design.

4. CONCLUSION

This article proposed SDAW-K-Means, a feature-weighted extension of classical K-Means based on within-cluster dispersion. The method learns feature weights during clustering, allowing informative features to contribute more strongly to the distance calculation. The Iris dataset illustration shows that the method yields interpretable feature weights and competitive clustering results. Future research should investigate convergence properties, simulation performance, robust variants, automatic K selection, and applications in high-dimensional statistical and machine-learning problems. Future research on the proposed K-Means clustering method should focus on systematic simulation studies, evaluation using multiple public datasets, and the reporting of uncertainty and stability measures through repeated splits, bootstrap validation, cross-validation, and sensitivity analysis. Future studies should also compare the method with classical baselines, modern extensions, robust alternatives, and relevant machine-learning methods, while providing reusable software implementation with clear input, output, and parameter definitions. In addition, theoretical properties, interpretability, reporting standards, robustness under unfavorable conditions, and connections to real-world substantive problems should be further investigated to strengthen the validity, reproducibility, and practical usefulness of the proposed method.

REFERENCES

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281-297.

[2] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982, doi: 10.1109/TIT.1982.1056489.

[3] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," Journal of the Royal Statistical Society: Series C, vol. 28, no. 1, pp. 100-108, 1979.

[4] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027-1035.
[5] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010, doi: 10.1016/j.patrec.2009.09.011.
[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, 2005, doi: 10.1109/TNN.2005.845141.
[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.
[8] UCI Machine Learning Repository, Iris Dataset. Irvine, CA, USA: University of California, Irvine.
[9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.
[10] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193-218, 1985.
[11] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B, vol. 63, no. 2, pp. 411-423, 2001.
[12] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the K- means clustering algorithm," Expert Systems with Applications, vol. 40, no. 1, pp. 200-210, 2013, doi: 10.1016/j.eswa.2012.07.021.
[13] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ, USA: Prentice Hall, 1988.
[14] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ, USA: Wiley, 1990.
[15] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, 1979, doi: 10.1109/TPAMI.1979.4766909.
[16] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," Journal of Cybernetics, vol. 4, no. 1, pp. 95-104, 1974.
[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[18] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," Journal of the American Statistical Association, vol. 105, no. 490, pp. 713-726, 2010, doi: 10.1198/jasa.2010.tm09415.
[19] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications. Philadelphia, PA, USA: SIAM, 2007.
[20] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach, 2nd ed. Boca Raton, FL, USA: CRC Press, 2012.