

The Improved K-Means Algorithm for Clustering Optimization: A Comparative Study on VRP Dataset

Nguyen Le Phuong Thao

Master of Science, Teaching Assistant, Industrial System and Engineering
School of Industrial Engineering and Management, International University, Vietnam National
University, Vietnam
nlpthao@hcmiu.edu.vn

Le Nguyen Hoang Vinh

Master of Business Administration, Department of Industrial Management
School of Management
National Taiwan University of Science & Technology (NTUST)
Taiwan
lenguyenhoangvinh3955@gmail.com

Vincent F. Yu

Professor, Department of Industrial Management
School of Management
National Taiwan University of Science & Technology (NTUST)
Taiwan
vincent@mail.ntust.edu.tw

Phan Nguyen Ky Phuc

Associate Professor, Senior Lecturer, Industrial System and Engineering
School of Industrial Engineering and Management, International University, Vietnam National
University, Vietnam
pnkphuc@hcmiu.edu.vn

Abstract

Clustering algorithms are critical in data analysis and optimization, particularly for segmenting datasets in practical applications. This study proposes an improved variant of K-means clustering algorithm and do evaluation on the performance of nine popular clustering methods—DBSCAN, HDBSCAN, Spectral Clustering, Hierarchical Clustering, OPTICS, Mean-Shift Clustering, Self-Organizing Map (SOM), K-Means, Gaussian Mixture Model (GMM), and the proposed Improved K-Means algorithm—using Vehicle Routing Problem (VRP) dataset. Implemented in Python, the methods are compared using the Calinski-Harabasz Index and Silhouette Index to assess cluster quality. Results show that the proposed improved K-Means algorithm, with its optimized parameter configuration through the initialization process of the parameter K and cluster centroid set, performs exceptionally well compared to other methods. These findings highlight the importance of parameter tuning and algorithm selection in clustering tasks, offering valuable insights for researchers and practitioners. This work provides practical guidance

for selecting and refining clustering approaches, contributing to advancements in applications such as vehicle routing and beyond.

Keywords

Clustering algorithm, Improved K-means, Vehicle Routing Problem (VRP), Calinski-Harabasz index, Silhouette index.

1. Introduction

Clustering algorithms are fundamental in data analysis and optimization, providing a means to segment datasets into meaningful groups for improved decision-making. Their applications span numerous domains, including logistics (Battarra et al., 2014; Le et al., 2022; Yücenur & Demirel, 2011), healthcare (Haraty et al., 2015; Negi & Chawla, 2021; Ogbuabor & Ugwoke, 2018), and supply chain optimization (Bai & Sun, 2021; Prabhu et al., 2020; Srinivasan & Moon, 1999). There are three primary strategies for classification: supervised, semi-supervised, and unsupervised. Supervised classification involves using predefined class labels or categories for certain data points to create a training set, which the algorithm uses to derive classification rules. Semi-supervised methods, on the other hand, utilize a combination of labeled and unlabeled data, which is particularly useful when labeling data manually is expensive or time-consuming. Unsupervised classification, commonly referred to as clustering, focuses on grouping data into clusters without prior knowledge of class labels. The goal of clustering is to form groups where objects within a cluster are more similar to one another than to those in other clusters (Rodriguez et al., 2019). This method not only models the data by identifying key patterns but also simplifies its properties, offering a clearer understanding of complex datasets. Although clustering is typically more computationally intensive than supervised classification, it often yields deeper insights into intricate data structures. Clustering algorithms often require tuning multiple parameters, operating in high-dimensional environments, and must handle challenges such as noisy, incomplete, or sampled data. As a result, their effectiveness can differ significantly depending on the application and dataset (Rodriguez et al., 2019). To address these challenges, numerous clustering methods have been introduced in the literature (Camastra & Verri, 2005; Jing et al., 2007; Suzuki & Shimodaira, 2006). Among clustering techniques, K-means remains a cornerstone due to its simplicity and computational efficiency (Xu & Tian, 2015). However, its reliance on parameter initialization and sensitivity to outliers often result in suboptimal clustering outcomes, particularly for datasets with complex structures (Karna & Gibert, 2022). These limitations have led to the development of various alternative clustering algorithms, such as DBSCAN, HDBSCAN, and Spectral Clustering, each tailored to address specific challenges (Rodriguez et al., 2019).

In logistics and optimization, datasets like the Vehicle Routing Problem (VRP) present unique clustering challenges due to their high dimensionality and complex data patterns. Although significant progress has been made in clustering algorithm development, comparative evaluations focused on optimization-specific datasets remain sparse. Furthermore, existing research frequently overlooks the importance of robust parameter tuning and its impact on clustering outcomes (Frías et al., 2023; Geetha et al., 2009). These gaps underscore the need for more systematic analyses of clustering methods tailored to real-world optimization problems.

This study aims to address these gaps by proposing an improved variant of the K-means algorithm that optimizes the initialization of the parameter K and cluster centroids. To evaluate its performance, the study systematically compares the proposed method against nine popular clustering algorithms—DBSCAN, HDBSCAN, Spectral Clustering, Hierarchical Clustering, OPTICS, Mean-Shift Clustering, Self-Organizing Map (SOM), Gaussian Mixture Model (GMM), and the traditional K-means—using the VRP dataset as a benchmark. The evaluation employs the Calinski-Harabasz Index and Silhouette Index, recognized metrics for assessing clustering quality, to ensure a robust and fair comparison. The findings of this study demonstrate that the proposed improved K-means algorithm outperforms existing methods in terms of clustering quality and scalability. These results highlight the critical role of parameter optimization and algorithm selection in achieving effective clustering outcomes, particularly for applications in logistics and beyond. By offering practical insights and advancing the methodology for evaluating clustering algorithms, this research contributes to the broader field of data analysis and optimization.

This paper consists of several sections as follows. Section 2 is the literature review, which describes the overview of the research gaps of previous studies. This paper consists of several sections as follows. Section 2 provides a comprehensive literature review, discussing the research gaps identified in previous studies. Section 3 presents the proposed improved K-means clustering algorithm for the Vehicle Routing Problem (VRP). Section 4 includes detailed

data instances used as benchmarks for performance comparison. Section 5 presents numerical experimental results and analysis. Finally, Section 6 summarizes the conclusions and discusses future research directions.

1.1 Objectives

The primary objective of this research is to bridge the gap in clustering optimization by introducing an improved variant of the K-means clustering algorithm. This study aims to enhance clustering performance, particularly in VRP-related datasets, by addressing common challenges such as parameter initialization, sensitivity to cluster count, and centroid placement. Specifically, the research focuses on: (1) Developing an enhanced K-means algorithm with optimized parameter initialization; (2) Comparing the proposed method against existing clustering techniques using benchmark datasets; (3) Evaluating the clustering quality based on widely used performance metrics such as the Silhouette Index and Calinski-Harabasz Index.

By fulfilling these objectives, the study contributes to advancing clustering methodologies and provides a robust solution for optimization problems in the VRP field.

2. Literature Review

Clustering algorithms are pivotal in unsupervised learning, offering valuable methods to segment datasets for meaningful pattern discovery. These techniques are widely applied across domains such as logistics, healthcare, and optimization problems, including the Vehicle Routing Problem (VRP). Despite their effectiveness, clustering methods face challenges due to sensitivity to parameter configurations, noise in data, and variability in performance across applications. This review examines key clustering algorithms and their relevance to VRP, focusing on recent developments and evaluation techniques.

The K-means algorithm is a fundamental method known for its simplicity and efficiency. However, it is sensitive to initialization and requires the number of clusters K to be predetermined, which can lead to suboptimal results. To address these limitations, K-means++ introduces a more robust seeding strategy, significantly improving clustering accuracy and computational performance by minimizing the average squared distance between points and centroids (Arthur & Vassilvitskii, 2006). Arthur and Vassilvitskii (2006) improves K-means by using a probabilistic seeding technique based on squared distances, significantly enhancing accuracy and efficiency. They obtain $O(\log k)$ -competitiveness with the optimal solution and consistently outperforms standard K-means in speed and clustering quality. In contrast, hierarchical clustering offers a dendrogram-based approach, providing a multi-level perspective of the dataset (Ward Jr, 1963). However, determining the optimal number of clusters often requires human expertise or automated techniques like the Calinski-Harabasz index to enhance interpretability (Wang & Xu, 2019).

Density-based clustering methods, such as DBSCAN and HDBSCAN, excel at detecting clusters of arbitrary shapes and handling noisy data (Rodriguez et al., 2019). These algorithms are particularly effective when dealing with spatial data but require careful tuning of parameters like neighborhood radius and minimum points. Similarly, Spectral Clustering and Gaussian Mixture Models (GMM) are powerful tools for capturing complex data distributions, though their scalability is limited by high computational demands. Rodriguez et al. (2019) evaluated nine clustering algorithms on diverse artificial datasets, highlighting that no single method consistently excels across all scenarios. Spectral clustering performed well on high-dimensional data, while K-means proved efficient for lower-dimensional cases, emphasizing the importance of algorithm selection and parameter tuning.

In the context of VRP, clustering serves as a pre-processing step to segment customers into manageable groups, facilitating route optimization. The combination of clustering with metaheuristic algorithms, such as Ant Colony Optimization, has demonstrated success in addressing VRP variants like the Energy Minimizing Vehicle Routing Problem (EMVRP) (Frias et al., 2023; Kara et al., 2007). Hybrid approaches, such as those integrating K-means with ACO proposed by Frias et al. (2023), have shown promising results in reducing energy consumption and computational costs. Additionally, the use of clustering for VRP with time windows (VRPTW) has proven effective in optimizing route planning under capacity and time constraints. The evaluation of clustering algorithms is a critical aspect of their application. Metrics like the Silhouette Index and Calinski-Harabasz Index are widely used to assess cluster quality, focusing on intra-cluster cohesion and inter-cluster separation (Ashari et al., 2023; Wang & Xu, 2019). Enhanced validation methods, such as the Peak Weight Index (PWI), have been proposed by Wang and Xu (2019) to improve the robustness of these evaluations. Despite these advancements, systematic comparisons of clustering algorithms tailored to optimization-specific datasets, such as VRP, remain limited. This gap underscores the need for

research focusing on algorithmic improvements and robust evaluation frameworks designed for real-world applications. By addressing these challenges, clustering methods can better support complex problems in logistics and beyond.

3. Methods

This section details the working methodology, and the process involved in The Improved K-means Clustering with two enhancements on the parameter initialization process. The performance comparison of the Improved K-means clustering with the original K-means clustering algorithm can be found in section 5.

The overall behavior of the proposed algorithms in this work can be observed in Figure 1. In the flowchart of the clustering process, the improved variant of K-means algorithm is enhanced in centroids set initialization using K-mean++ algorithm and value K initialization using the combination of Calinski – Harabasz index and Silhouette index. After then, the initialized parameters are input into K-means clustering algorithm to obtain the final clusters.

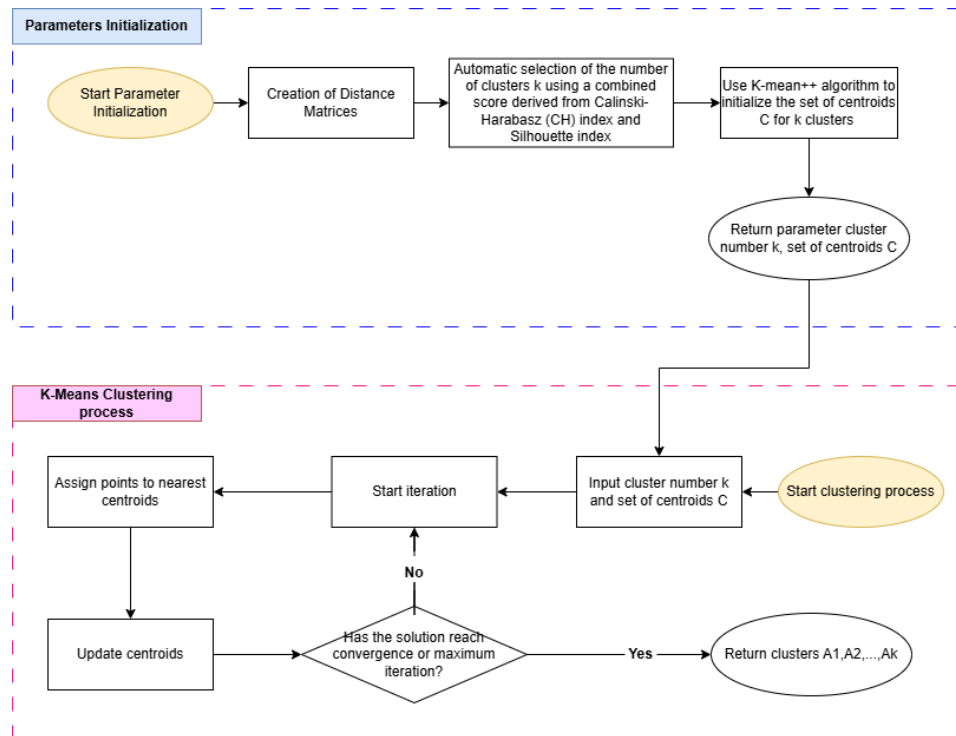


Figure 1: Generalized flowchart of the proposed Improved K-Mean clustering algorithm

3.1. Problem formulation

This section formally defines the clustering problem. In this research, focusing on Vehicle Routing Problem (VRP), customer locations are used as numerical attributes for clustering, and routing is applied within each cluster. Consider a generic VRP defined by a graph $G = (V, A)$, where $V = v_0, v_1, v_2, \dots, v_N$ is the set of customers, with v_0 serving as the depot. Each customer has its own location coordinates $d_i = (x_i, y_i)$ in a Cartesian plane. Clustering algorithm is to cluster N customers into k clusters where k is known advanced. The objective of the algorithm is to minimize the formula (1) and group customers into each separate cluster, having distinct distances among clusters:

$$\sum_{i=1}^N D(d_i, c_k) \quad (1)$$

Where $D(d_i, c_k)$ is the Euclidean distance between a data point d_i and the cluster centroid c_k , $c_k \in C = \{c_1, c_2, \dots, c_K\}$ – the set of K clusters.

3.2. The improved K-Means clustering algorithm

K-means clustering is one of the famous clustering algorithms based on Partition, with the basic idea of regarding the center of data points as the center of the corresponding cluster (Xu & Tian, 2015). The k-means clustering process is conducted through iteratively updating the center of cluster, until compact clusters are formed. It is outstanding for high computation effectiveness and low time complexity. However, one of the weaknesses of it is that the number of clusters needs to be preset, and the clustering result is sensitive to the number of clusters. Wang and Xu (2019) shows *Silhouette* index and *Calinski – Harabasz* index (in this study, we use the abbreviation for them as *S* index and *CH* index) are two common strategies selecting the optimal number of clusters K . However, the two indexes are sensitive to the characteristics of dataset. Additionally, the traditional K-means clustering algorithm initializes the centroids center set randomly which can affect the solution quality (Arthur & Vassilvitskii, 2006).

This study proposes an improved K-means clustering algorithm with enhancement on parameter initialization. Overall, the improved K-Means algorithm proposed in this paper is enhanced by adding two parameter initialization process: (1) Automatic hierarchical tree cutting and selection of the number of clusters k ; (2) Initialization of the set of centroids using K-Mean++ algorithm. After parameter initialization process is done, the parameters are input into the original K-mean clustering algorithm to obtain the clustering set of data points. As shown in Figure 1, in the improved variant of k-means, the value K initialization is proposed efficiently using the combination of *CH* index and *S* index. This combination of the characteristics in two indexes is supposed to help improve the fluctuation of clustering result in the dataset (Wang & Xu, 2019). Then, the centroid set initialization is processed using K-means ++ algorithm. On the work of Frías et al. (2023), it demonstrates that the initial position of the centroids directly affects the result of the clustering process. This parameter initialization process helps to augment the K-means algorithm to obtain the competitive optimal clustering result (Arthur & Vassilvitskii, 2006; Frías et al., 2023)

The proposed improved K-means clustering algorithm is shown in algorithm 1. Firstly, Agglomerative clustering method as described in Ward Jr (1963) is used to obtain the hierarchical clustering. Then, we do the horizontal cut on the hierarchical clustering into K clusters ($K \in 2, 3, \dots, K_{max} - 1$) and compute the according *CH* index and *S* index. K_{max} is theoretically $N - 1$ where N is the number of data points, however, Karna and Gibert (2022) suggests up to 7-10 clusters is meaningful for practical considerations and computational effectiveness. *CH* index for cluster K is computed by the formula as follows:

$$CH(K) = \frac{B(K)/(K - 1)}{W(K)/(N - 1)} \quad (2)$$

Where $B(K)$ and $W(K)$ indicate between-cluster sum of squares and within-cluster sum of squares respectively, K is the number of clusters and N is the total size of the data. The detail mathematical formular of $B(K)$ and $W(K)$ can be found in the appendix. And *S* index for cluster K is computed by the formula as follows:

$$\bar{S}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \right) \quad (3)$$

Where $a(i)$ and $b(i)$ represent the average distance of node i to other nodes in the cluster and the minimum distance of the sample from the node i to the other clusters. The best value of K is chosen if its associated value *combined_score* is the highest maxima among the set of *combined_score*. *Combined_score* is calculated using formula (4). It is noted that *CH* index is normalized for bias mitigation when combining with *S* index. Mathematically the decision rule is demonstrated as follows:

$$combined_score[K] = \bar{S}(K) + \frac{CH[K]}{\max(CH)} \quad (4)$$

Where $\max(CH)$ represents the maximum value of *CH* among the considered cases of value K . After determining the optimal number of clusters K , its clustering centroids are initialized using the K-means++ algorithm, as shown in the work of D. Arthur and S. Vassilvitskii (Arthur & Vassilvitskii, 2006). K-means ++ algorithm is proved to be competitive in computational time, which is $O(\log k)$ competitive (Arthur & Vassilvitskii, 2006). The results shown in Arthur and Vassilvitskii (2006) prove that with careful seedings on initial clustering centroids, the performance of clustering process is improved significantly.

After parameter initialization process is completed, the number of clusters K is defined and the set of initial cluster centroids is input into the K-means clustering algorithm to obtain the final clustering assignment. Different from the original K-mean clustering algorithm, the improved variant proposed in this paper has decision rules defining the number of clusters K and the set of centroids as in previous sections to desire a more logical and automatic way of parameter initialization.

Algorithm: The proposed Improved K-means Clustering	
1:	Input: Set of elements $N = \{e_1, e_2, \dots, e_N\}$
2:	Initialize hierarchical clustering using agglomerative clustering method
Procedure: Parameter initialization	
3:	Determine the optimal number of clusters K for K in range $(2, K_{max} - 1)$ Calculate CH index $CH(K)$ and S index $\bar{S}(K)$ using (2) and (3) Calculate $Combined_score[K]$ using (4)
4:	Determine the optimal number of clusters K : $K = argmax(combined_score), K \geq 2$
5:	Initialize the set of centroids for K clusters using K-means++ clustering algorithm
Procedure: Cluster assignments using K-mean clustering	
6:	Input parameter K and the set of centroids from Parameter initialization
7:	for iteration in range $(0, max_iter)$: if $distance(e_i, centroid_k)$ is minimum for all $k = \{2, \dots, K\}$: Assign e_i to the nearest centroid based on Euclidean distance Update $centroid_k = mean(points\ in\ cluster\ k)$ iteration += 1 End
Output: Set of K clusters $\{A_1, A_2, \dots, A_K\}$	

4. Data Collection

The algorithm running and comparison is conducted with benchmarks from the CVRPLIB library. “Set A - Augerat, 1995,” and “Set 2 - Golden et al., 1998.” are used in this study. The customer 2D coordinates are extracted from each .txt data file and forms into a coordinates array as the input into the clustering process. Data from set A consisting of 22 instances with data sizes from 30 to 78 nodes. Set B consists of 10 groups of instances ranging from 200 to 483 nodes (excluding the depot). The details of the dataset can be found in the open source CVRPLIB library.

5. Results and Discussion

In this section, the measurement metric for the evaluation of the quality of the generated partitions among clustering algorithms is presented. CH score is presented to evaluate the algorithm performance. This metric is introduced as the metric measuring the ratio of between – cluster dispersion to within-cluster dispersion. A higher score indicates better-defined clusters. Our proposed improved K-means clustering algorithm is compared with the other nine common clustering methods: DBSCAN, HDBSCAN, Spectral clustering, Hierarchical clustering, OPTICS, Mean-shift, SOM, K-means and GMM programed by Scikit-learn – a common Python package.

The parameter configurations for each algorithm are defined in Table 1 to ensure consistent functionality and comparability. Common parameters like initialization runs, maximum iterations, and cluster counts are uniformly set across methods. Algorithm-specific parameters, such as the neighborhood function and learning rate for Self-Organizing Maps (SOMs) and minimum sample size for OPTICS, are appropriately configured to match their respective operational requirements. The number of clusters K for those algorithms requiring the preset value is set using Silhouette score. These settings provide a robust foundation for evaluating clustering performance.

Table 1. Parameter configuration for comparative clustering algorithms

Algorithm	Parameter	Value
All algorithm	Number of times the algorithm is run with different initializations	15
	Maximum iterations in one run	300 (default)
	Random seed for reproducibility	42
	The number of clusters to form	Determined using Silhouette score
Self-Organizing Maps (SOMs) algorithm	Spread of the neighborhood function	0.5
	Initial learning rate for updating weights	0.5
	The number of rows of the SOM grid	5
	The number of columns of the SOM grid	5
OPTICS algorithm	The minimum samples within one cluster	5

5.1 Numerical Results

To augment the significant improvement thanks to the combination of S and CH indices, we develop the three versions of the improved K-means: the proposed **Improved K-Means (Combined)**, the Improved K-means (Silhouette) which deploys only S index, the improved K-means (CH) which uses only CH index. As shown in Figures 2 and 3, the proposed **Improved K-Means (Combined)**, in general, demonstrates superior clustering performance by integrating the strengths of the S and CH indices. Through the analysis of various datasets, the Combined variant consistently outperforms both the baseline K-Means and other improved variants, achieving higher S and CH scores across diverse clustering challenges. This dual-objective optimization enables the algorithm to effectively balance intra-cluster cohesion and inter-cluster separation, ensuring robust performance even in complex and noisy datasets. The results highlight its adaptability and reliability, making it a strong candidate for real-world applications where clustering quality is critical.

Among the nine common clustering algorithms, as indicated in Figure 4 and 5, the Improved K-Means consistently ranks among the top-performing algorithms, and also standard K-means clustering algorithm, in terms of CH Index and S index. It excels particularly in complex datasets like the "Golden" datasets, where it achieves high S index and CH scores, indicating strong inter-cluster separation and intra-cluster cohesion.

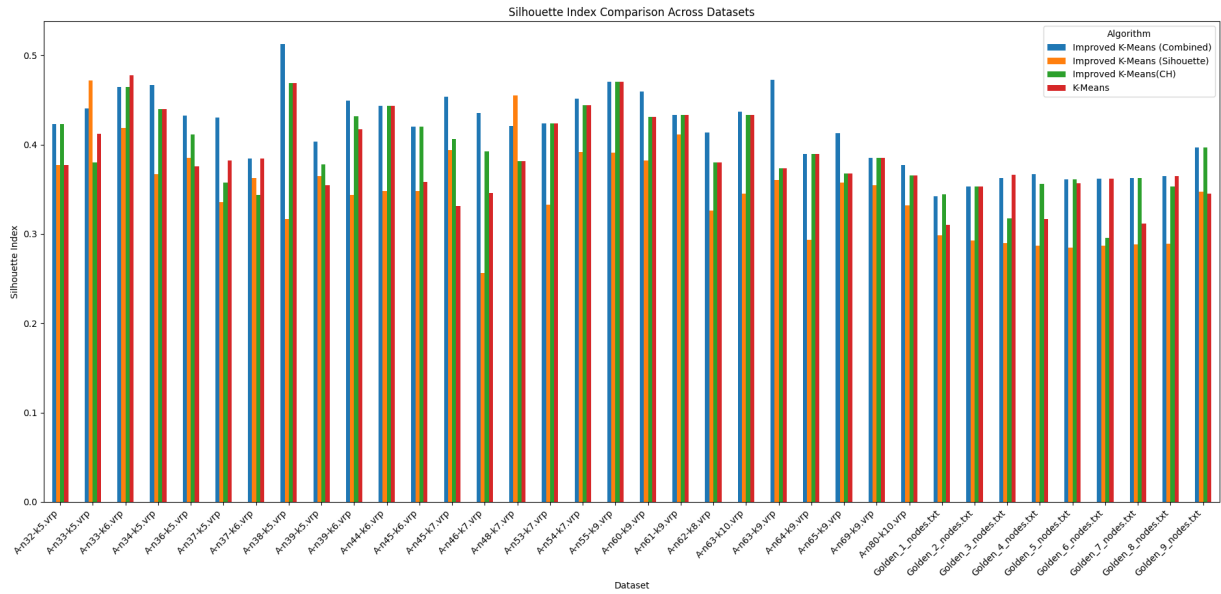


Figure 2: *S* index Comparison graph among three proposed improved clustering algorithm and the original clustering algorithm

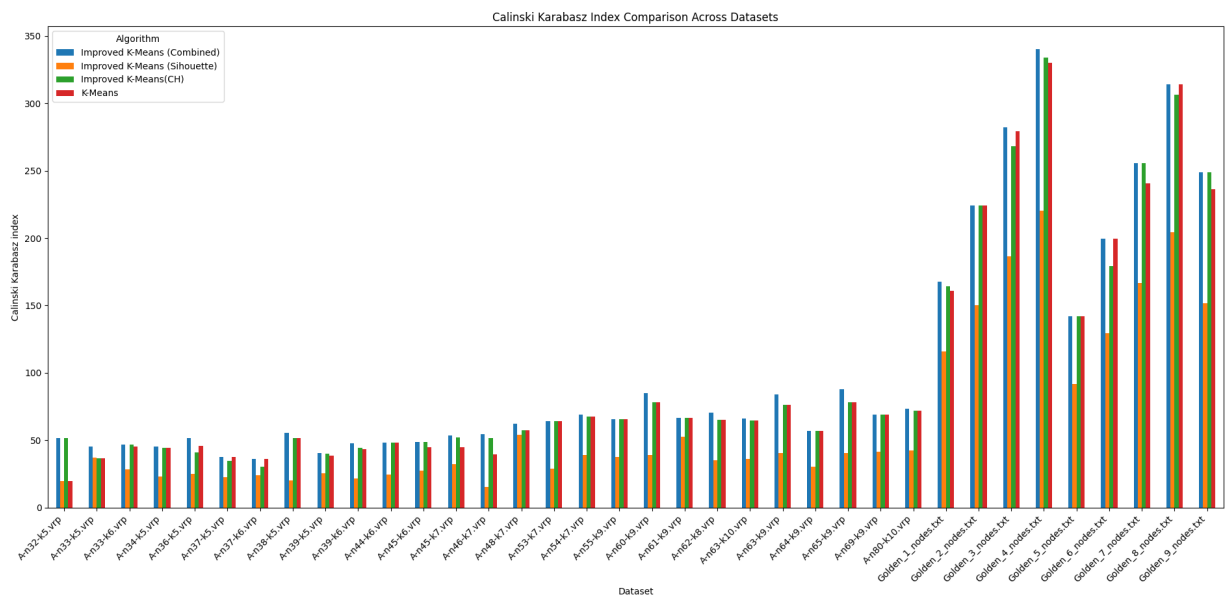


Figure 3: *CH* index Comparison graph among the three proposed improved clustering algorithm and the original clustering algorithm

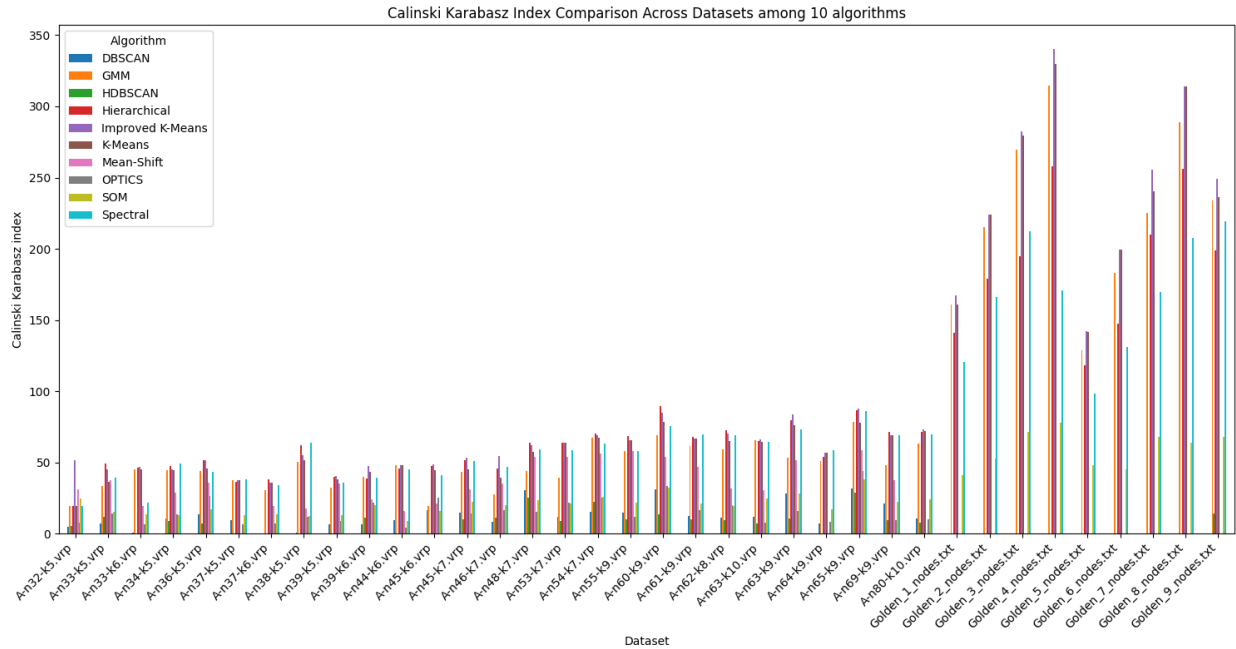


Figure 4: Calinski – Harabasz index Comparison graph among nine clustering algorithms and the proposed improved clustering algorithm

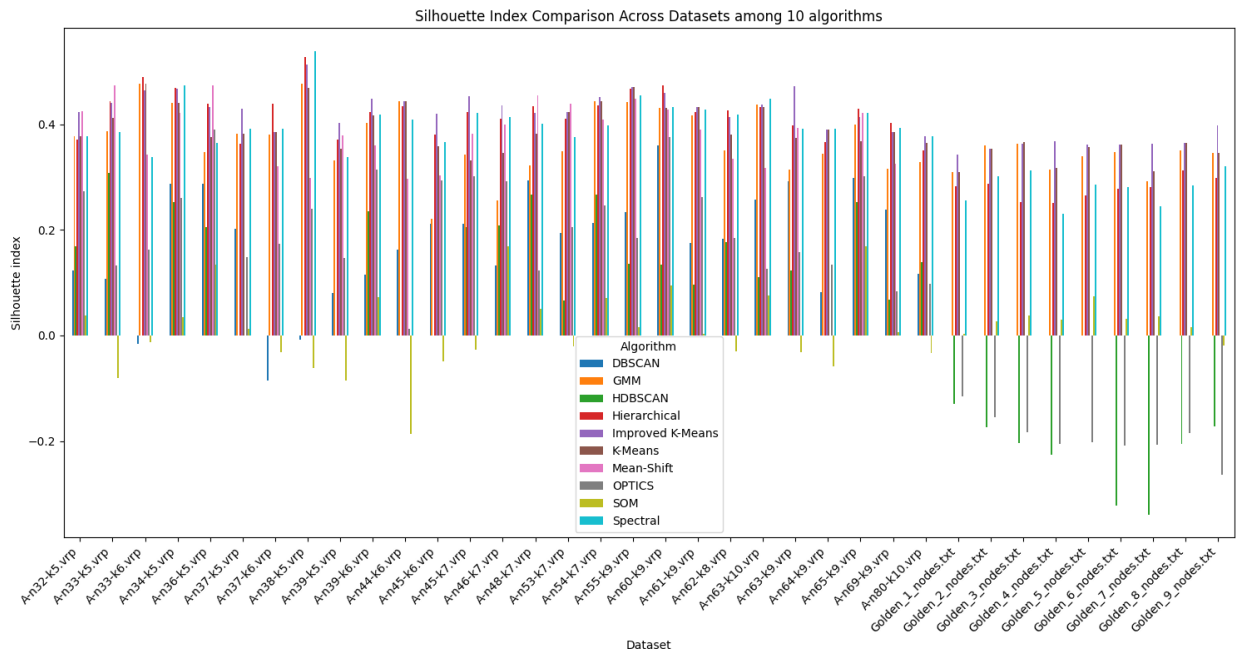


Figure 5: Silhouette index Comparison graph among nine clustering algorithms and the proposed improved clustering algorithm

5.2 Graphical Results

Figure 6 illustrates the clustering results on the dataset “A-n32-k5.vrp” for graphical result demonstration. It shows how the proposed Improved K-means optimally chooses the number of clusters K ($K=9$) that maximizes the intra-cluster cohesion and inter-cluster separation, compared with the traditional K-means choosing $K = 2$ that suffer from under-clustering, likely due to suboptimal selection of K or an inability to adapt to dataset complexity. The improved K-Means demonstrates clear superiority by identifying more meaningful clusters that better reflect the underlying data structure.

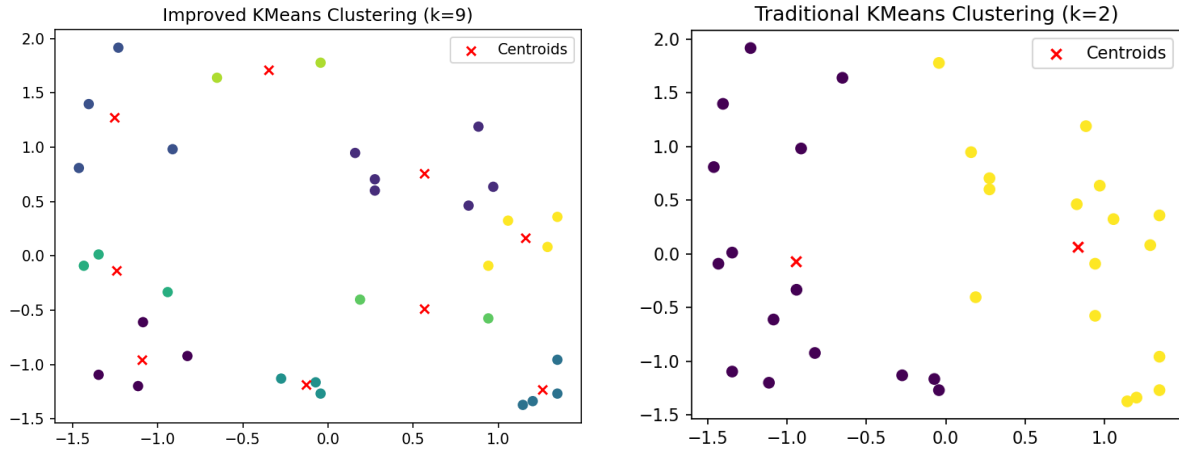


Figure 6: Clustering results obtained from the improved clustering and the original clustering

5.3 Proposed Improvements

As shown in Figure 7, the improved K-means algorithm demonstrates significant as evidenced by an average 8% improvement in S index and a 10% improvement on CH index compared to original K-Means. Figure 6 illustrates the well-separated and cohesive clusters generated by the proposed method. Future work will focus on integrating advanced clustering techniques such as density-based clustering to handle non-linear separations. Additionally, we propose testing the algorithm on larger datasets to assess scalability and robustness."

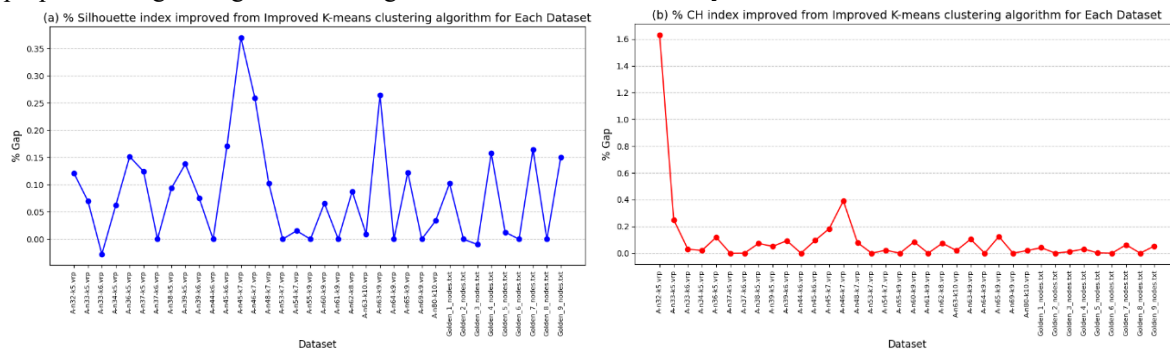


Figure 7: Improvement percentage of clustering performance of the proposed clustering algorithm with the original clustering algorithm on metric: (a) S index (b) CH index

5.4 Validation

This study validates the clustering structure using internal validation. We run the K-means clustering (Scikit-learn) and compare its performance with our proposed improved K-means clustering algorithm on two metrics: S index and CH index. The proposed algorithm enhances S index 8% and CH index 10% improvement compared with the well-known K-means algorithm of Scikit-learn. Besides, we propose the three versions of improved K-means clustering algorithm, as shown in section 5: Improved K-means (Silhouette), Improved K-means (CH) and Improved K-means (Combined) to validate the effectiveness of the combination of S index and CH index in parameter initialization. The datasets “Golden et al. 1998 – Set 2” and “Augerat 1995 – Set A” are well-established in the field of vehicle routing problem (VRP). Their reliability can be assessed based on their origins and VRP-REP Repository quality control.

6. Conclusion

The proposed Improved K-Means Clustering Algorithm successfully fulfills all research objectives, demonstrating significant advancements in clustering accuracy and robustness. By integrating the S Index and CH Index, the algorithm achieves a well-balanced optimization of intra-cluster cohesion and inter-cluster separation. The adoption of K-Means++ for centroid initialization further enhances parameter optimization, resulting in improved clustering quality and computational efficiency. Extensive experimental evaluations on diverse datasets, including highly complex and noisy scenarios such as the "Golden" datasets, illustrate the superior performance of the Improved K-

Means algorithm. It consistently outperforms the original K-Means and other benchmark clustering methods, achieving higher S and CH indices across varied clustering tasks. These results underscore the adaptability and reliability of the proposed approach in addressing real-world challenges.

The unique contribution of this research lies in the development of a novel dual-metric optimization framework combined with an enhanced centroid initialization strategy. This innovative approach not only improves clustering outcomes but also provides a robust and scalable solution for practical applications, such as the Vehicle Routing Problem. The findings of this study contribute to advancing clustering methodologies and offer valuable insights for future research and practical implementations in the field.

References

- Arthur, D., & Vassilvitskii, S, *k-means++: The advantages of careful seeding*. 2006.
- Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R, Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta. *Journal of Applied Informatics and Computing*, 7(1), 95-103, 2023.
- Bai, S., & Sun, H, Research on Enterprise Supply Chain Optimization Model and Algorithm Based on Fuzzy Clustering. *Journal of Mathematics*, 2021(1), 4827903, 2021.
- Battarra, M., Erdoğan, G., & Vigo, D, Exact algorithms for the clustered vehicle routing problem. *Operations Research*, 62(1), 58-71, 2014.
- Camstra, F., & Verri, A, A novel kernel method for clustering. *IEEE transactions on pattern analysis and machine intelligence*, 27(5), 801-805. 2005.
- Frias, N., Johnson, F., & Valle, C, Hybrid Algorithms for energy minimizing vehicle routing problem: integrating clustering and ant colony optimization. *IEEE Access*. 2023.
- Geetha, S., Poonthalir, G., & Vanathi, P, Improved k-means algorithm for capacitated clustering problem. *INFOCOMP Journal of Computer Science*, 8(4), 52-59. 2009.
- Haraty, R. A., Dimishkieh, M., & Masud, M., An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of distributed sensor networks*, 11(6), 615740.2023.
- Jing, L., Ng, M. K., & Huang, J. Z, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8), 1026-1041, 2007..
- Kara, I., Kara, B. Y., & Yetis, M. K, Energy minimizing vehicle routing problem. Combinatorial Optimization and Applications: First International Conference, COCOA 2007, Xi'an, China, August 14-16, 2007. Proceedings 1,
- Karna, A., & Gibert, K, Automatic identification of the number of clusters in hierarchical clustering. *Neural Computing and Applications*, 34(1), 119-134 2022.
- Le, T. D. C., Nguyen, D. D., Oláh, J., & Pakurár, M, Clustering algorithm for a vehicle routing problem with time windows. *Transport*, 37(1), 17-27-17-27, 2022.
- Negi, N., & Chawla, G, Clustering Algorithms in Healthcare. In *Intelligent healthcare: Applications of ai in ehealth* (pp. 211-224), 2021. Springer.
- Ogbuabor, G., & Ugwoke, F, Clustering algorithm for a healthcare dataset using silhouette score value. *Int. J. Comput. Sci. Inf. Technol*, 10(2), 27-37, 2018.
- Prabhu, R. M., Hema, G., Chepure, S., & Guptha, M. N, Logistics optimization in supply chain management using clustering algorithms. *Scalable Computing: Practice and Experience*, 21(1), 107-114, 2020.
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. , Clustering algorithms: A comparative approach. *PloS one*, 14(1), e0210236, 2019.
- Srinivasan, M., & Moon, Y. B., A comprehensive clustering algorithm for strategic analysis of supply chain networks. *Computers & industrial engineering*, 36(3), 615-633, 1999.
- Suzuki, R., & Shimodaira, H. (2006). Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), 1540-1542, 2019.
- Wang, X., & Xu, Y, An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. IOP Conference Series: Materials Science and Engineering, 1963,
- Ward Jr, J. H, Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236-244, 2015.
- Xu, D., & Tian, Y, A comprehensive survey of clustering algorithms. *Annals of data science*, 2, 165-193.
- Yücenur, G. N., & Demirel, N. Ç, A new geometric shape-based genetic clustering algorithm for the multi-depot vehicle routing problem. *Expert Systems with Applications*, 38(9), 11859-11865, 2011.

Biographies

Le Nguyen Hoang Vinh is a Ph.D. student at Taiwan Tech under Professor Yu. He earned his Master's and Bachelor's degrees from NTUST and Vietnam National University, respectively. His research focuses on the Vehicle Routing Problem and metaheuristics for logistics optimization.

Phan Nguyen Ky Phuc is a lecturer at International University (VNU-HCM) with expertise in optimization models, machine learning, and AI in production systems. He has authored 41 publications, including 18 ISI/Scopus-indexed papers, and has guided 13 graduate students in industrial engineering.

Vincent F. Yu (喻奉天) is a Professor at the National Taiwan University of Science and Technology (Taiwan Tech) in the Department of Industrial Management and the Graduate Institute of Intelligent Manufacturing Technology. He directs the Global Logistics and Supply Chain Management Laboratory and was a Visiting Scholar at Portland State University. He holds a Ph.D. and M.S. in Industrial and Operations Engineering from the University of Michigan, an M.S. in Mathematics from Michigan State University, and a B.S. in Mathematics from National Taiwan Normal University. His research focuses on operations research, logistics, supply chain management, AI, energy optimization, and blockchain applications.

Nguyen Le Phuong Thao holds an M.S. in Global Production Management and Engineering from the Technical University of Berlin and a B.E. in Logistics and Supply Chain Management from Vietnam National University. She has industry and academic experience, including research on metaheuristics and machine learning in resource optimization. Her work includes a published paper on vehicle routing using Ant Colony Optimization.