An Improved K-Selection Method for K-Means Clustering

Faisal N. Abu-Khzam and Jad El-Masri

Abstract

One of the main challenges of the K-means clustering algorithm is still determining the ideal number of clusters or k. The Elbow Method and other conventional techniques frequently produce inaccurate results, especially when applied to datasets with complex structures and varying densities. This study assesses the effectiveness of a number of K-selection techniques such as the Bayesian Information Criterion (BIC), d-packing, Elbow, Calinski-Harabasz, and Xie-Beni on a range of dataset types including dense, sparse, overlapping, and nonoverlapping clusters. The accuracy and computational efficiency of these tried and true techniques are contrasted with the d-packing method, which clusters data points according to a dynamic distance threshold. Experimental findings demonstrate that d-packing provides competitive performance in sparse and overlapping datasets, and outperforms more conventional techniques like Elbow and BIC in datasets with dense clusters. According to the results, d-packing offers a more robust approach to compute an optimal k in a variety of clustering scenarios, especially when combined with a dynamic distance selection mechanism.

1 Introduction

Clustering of K-means is one of the most widely used unsupervised machine learning algorithms to partition data sets into clusters K based on similarity of features. Despite its simplicity, a major challenge in K-means clustering is determining the optimal number of clusters, K. Traditional methods, such as the Elbow Method [7], the Bayesian Information Criterion (BIC) [10], and cluster validity indices like the Calinski-Harabasz [3] and Xie-Beni [2] indices, are often employed to estimate K. Although these methods have been widely used, they often fail when dealing with datasets that exhibit varying densities, complex structures, or significant overlap between clusters.

The Elbow Method works by identifying the point at which the within-cluster variance begins to decrease at a slower rate. However, it is not always effective in datasets with sparse or overlapping clusters, where the variance reduction may not be pronounced enough to clearly indicate an optimal K. Similarly, BIC [10] incorporates model complexity and fit, providing a more systematic approach to model selection, but they are prone to overestimation K, particularly in high-dimensional or complex datasets. These limitations highlight the need for alternative methods that can better handle datasets with varying cluster shapes and densities.

In this paper, we introduce a novel approach for K-selection using d-packing, a distance-based method traditionally applied in density-based clustering. Unlike conventional K-selection techniques that rely on global statistical properties, d-packing dynamically adjusts to local density variations to determine K. By integrating d-packing into K-means clustering, we propose an adaptive method for estimating K that does not require prior knowledge of the cluster count and is robust across various clustering scenarios.

In addition to d-packing, clustering validity indices such as the Calinski-Harabasz index [3] and Xie-Beni index [2] have gained attention as alternatives for estimating K. These indices evaluate clustering quality by measuring the compactness and separation of clusters, making them suitable for datasets with varying levels of overlap and density. However, while these indices are effective in many scenarios, their performance can still be limited in cases where clusters are not well-separated.

This paper presents a comparative analysis of d-packing, the Elbow Method, BIC, and the Calinski-Harabasz and Xie-Beni indices across a variety of datasets, including dense, sparse, overlapping, and non-overlapping clusters. Experimental results demonstrate that d-packing outperforms traditional methods in dense datasets and provides competitive results in sparse and overlapping datasets. These findings suggest that d-packing, with its dynamic distance selection mechanism, is a robust method for determining K in a wide range of clustering scenarios.

2 Related Work

Selecting the optimal number of clusters, K, for clustering algorithms such as K-means has been extensively studied. Various methods have been proposed to estimate K, each with its strengths and weaknesses depending on the dataset's characteristics.

One of the most widely used methods is the Elbow Method [7], which plots the within-cluster sum of squares (WCSS) as a function of K. The optimal K is typically identified at the "elbow" point where the rate of reduction in WCSS slows down. However, the Elbow Method can struggle in situations where the clusters are not well-separated or where the data contains noise, making the identification of a distinct elbow point difficult. Several studies have pointed out the limitations of this method in such complex scenarios [9, 18].

In response to the limitations of the Elbow Method, alternative statistical criteria have been proposed. The Bayesian Information Criterion (BIC) [10] and the Akaike Information Criterion (AIC) [15] are popular methods that provide model selection based on the likelihood of the data under different models, penalized by the number of parameters. While BIC and AIC are more robust than the Elbow Method in some scenarios, they can still overestimate K in datasets with complex structures or high-dimensional data, as they tend to favor more complex models [11].

Another approach involves the use of cluster validity indices that aim to assess the quality of the clustering solution based on both the compactness and separation of the clusters. The Calinski-Harabasz index [3], also known as the variance ratio criterion, measures the ratio of between-cluster dispersion to within-cluster dispersion, rewarding clustering solutions with well-separated and compact clusters. It has been widely used in clustering validation, but like the Elbow Method, it is not effective when the clusters overlap or when the data is highly noisy. Similarly, the Xie-Beni index [2] evaluates both the compactness and separation of clusters and is particularly effective in handling situations where clusters overlap. However, both indices can face challenges when dealing with datasets that do not have clearly defined clusters, particularly in high-dimensional spaces or when clusters are non-spherical.

In recent years, distance-based heuristics such as d-packing have been proposed as an alternative to traditional clustering methods. d-packing groups data points based on a dynamic distance threshold, d, that adapts to the local density of the data. d-packing avoids the need for prior knowledge of K, making it particularly useful in datasets with varying densities and complex cluster shapes. Unlike traditional methods, which assume a predefined number of clusters or a clear separation between clusters, d-packing dynamically adjusts to the structure of the data, making it a promising solution for complex clustering tasks. Previous studies have demonstrated that d-packing outperforms methods like the Elbow Method and BIC in dense datasets and has shown strong performance across different types of clustering problems [14, 20].

Beyond these methods, a number of other model selection techniques have been explored, such as information-theoretic approaches [16], bootstrap methods [17], and cross-validation [19]. These methods often aim to balance the goodness of fit and model complexity, similar to BIC and AIC, but they tend to be computationally expensive, particularly in large datasets or high-dimensional spaces.

In summary, traditional methods like the Elbow Method, BIC, and cluster validity indices such as Calinski-Harabasz and Xie-Beni are widely used for K-selection, but they face limitations when dealing with complex, high-dimensional, or overlapping datasets. d-packing provides a more flexible and robust solution for determining K in a wide variety of clustering scenarios. The ability of d-packing to adapt to different cluster structures—whether dense, sparse, overlapping, or non-overlapping—makes it a superior choice for K-selection in complex datasets. These results suggest that d-packing, particularly when combined with adaptive distance thresholds, offers a promising approach to improving clustering accuracy in diverse applications.

3 Methodology

This section describes the methodology used to assess d-packing performance in optimal K-selection within K-means clustering. The methodology components include: dataset generation; d-packing algorithm - the core algorithmic part of the methodology; and experimental setup.

3.1 Dataset Generation

To evaluate the effectiveness of K-selection methods, we generated 2000 synthetic datasets using Python's `scikit-learn`, `NumPy`, and `Matplotlib` libraries, ensuring a diverse range of clustering scenarios. Each dataset contained between 500 and 10,000 data points, with a predefined number of clusters ranging from 3 to 10. We categorized the datasets into four types:

- **Dense Clusters**: Clusters are tightly packed with minimal intra-cluster variance, generated using `make_blobs` with standard deviations between 0.3 and 0.6.
- **Sparse Clusters**: Widely separated clusters with higher intra-cluster variance, using a standard deviation range of 1.5 to 2.5 and a minimum centroid separation of 5 standard deviations.
- **Overlapping Clusters**: Clusters with significant overlap, where Gaussian noise was introduced using `numpy.random.normal` to simulate cluster ambiguity, with standard deviations between 0.8 and 1.5.
- **Non-overlapping Clusters**: Well-separated clusters ensuring distinct group separation, with a minimum inter-cluster distance of 5 standard deviations.

4 Optimal Distance Calculation for d in d-packing for k-Selection

In this context, k-selection to solve clustering problems, especially for d-packing, an effective choice of distance threshold d is very important. The distance d decides the tightness with which points are packed in a cluster, and this critically affects the performance and accuracy of the clustering algorithm. In this work, we propose a method for dynamically computing the optimal value of d based on the inherent characteristics of the dataset, thus enabling the algorithm to adapt to changing densities in data and various cluster structures.

The key intuition behind d-packing is to put together in one cluster those data points that are at a distance d from each other while maintaining the separation between clusters. The method of selecting d is not so direct because it depends on the distribution and density of the points in the dataset. A small d would give a large number of small clusters, while a large d may merge two different clusters. Now we present a dynamic approach for calculating d, which varies according to the nature of the dataset.

4.1 Dynamic Calculation of d

The optimal value of d is dynamically computed by first leveraging DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which is a density-based clustering algorithm that groups points based on their spatial proximity. The DBSCAN algorithm requires two parameters: the minimum number of points required to form a cluster (minPts) and the distance threshold (ϵ) that defines the neighborhood of each point. To calculate d dynamically, we use the following steps:

1. **Calculate Pairwise Distances:** The first step in our approach involves calculating the pairwise distances between all points in the dataset. This allows us to understand the spatial distribution of the points.

2. **Apply DBSCAN Clustering:** Using the pairwise distances, we apply the DBSCAN algorithm to identify the initial cluster structure. In DBSCAN, the ϵ parameter (the maximum distance between two points to be considered neighbors) significantly impacts the clustering result. In our case, we set ϵ as the maximum pairwise distance between points within a cluster.

3. **Estimate Initial ϵ:** To estimate ϵ, we calculate the pairwise distances between all points in the dataset. We then compute the k-nearest neighbor distance for each point, where k is typically set as the minimum number of points required to form a cluster. The maximum of these k-nearest neighbor distances is chosen as the initial value for ϵ, ensuring that the DBSCAN clustering captures the intrinsic structure of the dataset while considering the density of the points.

4. **Cluster Validation:** Once the initial clustering is performed, we analyze the resulting clusters by calculating the maximum pairwise distance within each cluster. This maximum distance is then used as the optimal d for d-packing. The idea is that points within a cluster should be close enough (within the threshold d) while maintaining sufficient separation between clusters. This ensures that the distance d is representative of the local density and spacing of the data.

4.2 Mathematical Formulation

In DBSCAN, for each point p_i, we calculate the distances to all other points, and determine the k-nearest neighbors. The distance to the k-th nearest neighbor is denoted as $\epsilon_k(p_i)$. The maximum of these distances over all points in the dataset is selected as the optimal ϵ, which is then used as the initial estimate for d.

$$\epsilon = \max_{p_i \in P} \epsilon_k(p_i)$$

Once the clusters are formed, we compute the maximum pairwise distance $d_{\max}$ within each cluster. This value represents the optimal d for d-packing, as it reflects the characteristic density and spread of the points within each cluster.

$$d_{\text{optimal}} = \max_{C \in \text{Clusters}} \max_{p_i, p_j \in C} d(p_i, p_j)$$

4.3 Advantages of Dynamic d-Selection

The advantage of dynamically calculating d lies in its adaptability. In clustering problems, datasets could be quite different in terms of density, distribution, and structure.

Traditional methods of choosing d need predefined values or manual tuning, which is usually time-consuming and suboptimal. The proposed method dynamically calculates d based on the characteristics of a dataset. Thus, it enables more accurate clustering, particularly in cases where there is variation in point densities or cluster structures are unknown. Moreover, such an approach avoids time-consuming trial-and-error tuning of the d-value, which is usually the bottleneck for traditional clustering algorithms. We adaptively handle the method in a way that drastically cuts down computational cost and enhances efficiency in the process of clustering.

5 Results

Dense Dataset Analysis

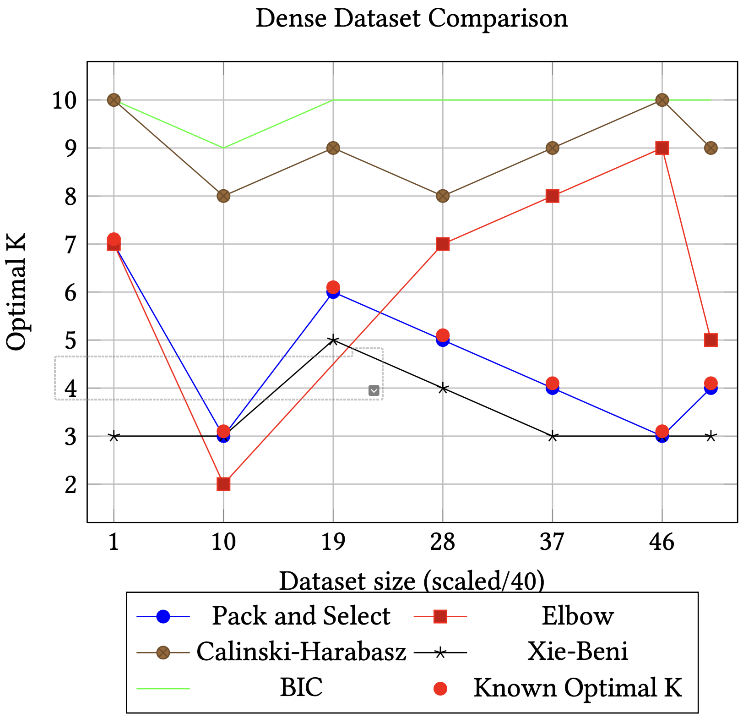

Figure 1: Comparison of K-selection methods for dense datasets, illustrating the performance of d-packing, Elbow, BIC, Calinski-Harabasz, and Xie-Beni methods in determining the optimal number of clusters.

In the dense dataset analysis, we observe that d-packing consistently identifies the correct K, closely matching the Known Optimal K values (shown in red). The Elbow Method tends to slightly underestimate K, while BIC overestimates it. Calinski-Harabasz performs well but occasionally misses the true number of clusters. Xie-Beni provides reasonable estimates, but d-packing remains the most accurate, especially in identifying compact, well-separated clusters. The Silhouette Scores for d-packing are higher, indicating better-defined clusters compared to the other methods.

Sparse Dataset Analysis

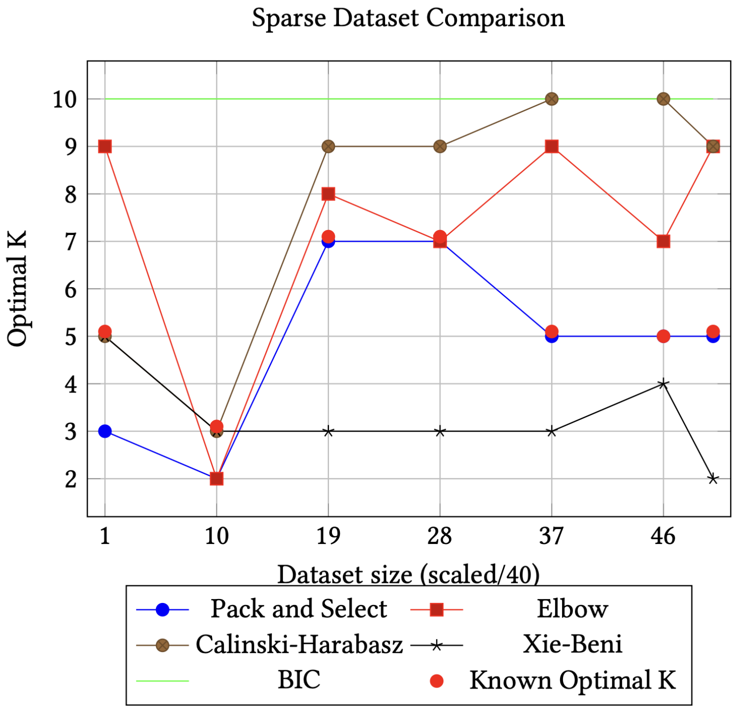

Figure 2: Comparison of K-selection methods for sparse datasets, illustrating the performance of d-packing, Elbow, BIC, Calinski-Harabasz, and Xie-Beni methods in determining the optimal number of clusters.

In the sparse dataset analysis, d-packing again outperforms the other methods, accurately identifying the true number of clusters. Elbow and BIC tend to underestimate the optimal K, especially in sparse datasets where the variance between clusters is less pronounced. Calinski-Harabasz performs reasonably well, but like BIC, it occasionally overestimates the number of clusters. Xie-Beni provides the best results in this case after d-packing, as it handles the sparsity and separation of clusters better than the other methods. Silhouette Scores confirm d-packing's superiority in this case, indicating more compact clusters.

Overlapping Dataset Comparison

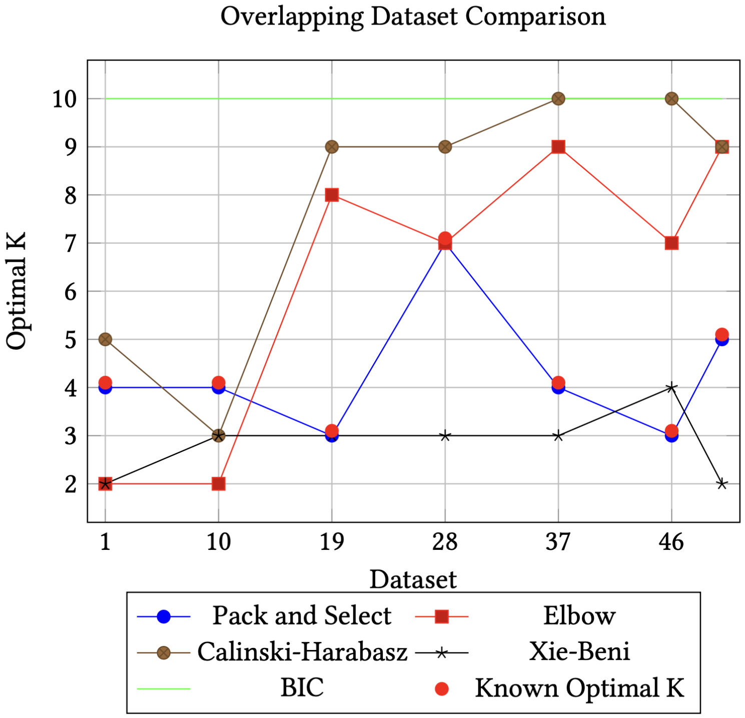

Figure 3: Comparison of K-selection methods for overlapping datasets, illustrating the performance of d-packing, Elbow, BIC, Calinski-Harabasz, and Xie-Beni methods in determining the optimal number of clusters.

In the overlapping dataset analysis, Xie-Beni performs exceptionally well, closely matching the Known Optimal K values. d-packing also identifies the correct K but is slightly less accurate in some cases compared to Xie-Beni. The Elbow Method and BIC again overestimate K, particularly in the presence of overlapping clusters, where the variance between clusters is not as pronounced. Calinski-Harabasz tends to slightly overestimate K, especially when the clusters overlap. Overall, Xie-Beni provides the best performance in this case, particularly in handling overlapping cluster boundaries effectively. Silhouette Scores are also higher for Xie-Beni compared to other methods.

Non-Overlapping Dataset Analysis

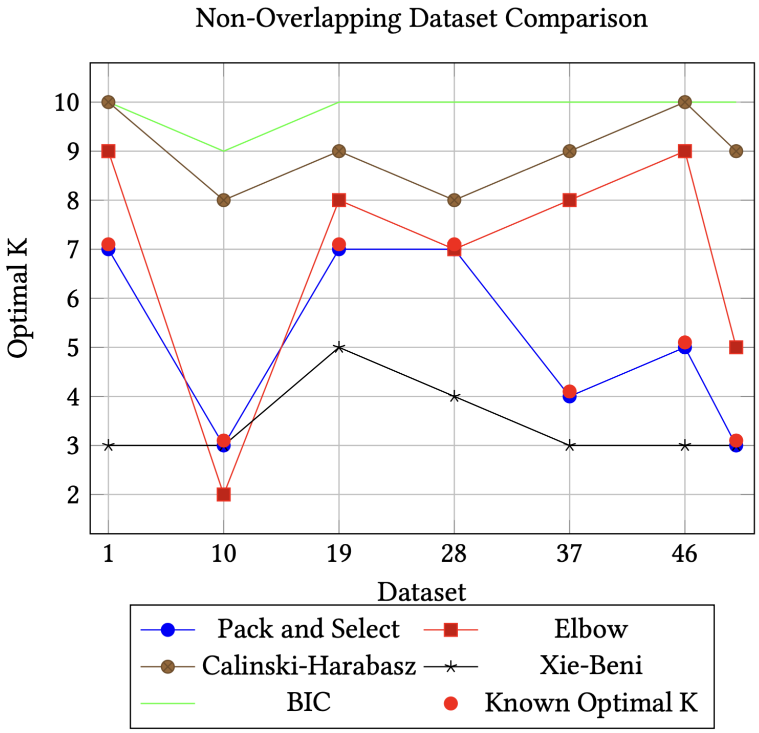

Figure 4: Comparison of K-selection methods for non-overlapping datasets, illustrating the performance of d-packing, Elbow, BIC, Calinski-Harabasz, and Xie-Beni methods in determining the optimal number of clusters.

In the non-overlapping dataset analysis, all methods perform reasonably well, with d-packing identifying the true K most accurately. Elbow and BIC again overestimate the optimal K, but d-packing and Xie-Beni provide the most reliable estimates. Calinski-Harabasz also performs well in this case but is slightly less precise in identifying the correct number of clusters. Silhouette Scores for d-packing and Xie-Beni are higher, indicating better clustering quality compared to the other methods.

Overall Performance of K-selection Methods

In this section, we provide an overall comparison of the performance of all five K-selection methods—d-packing, Elbow, BIC, Calinski-Harabasz, and Xie-Beni—across the four types of datasets: dense, sparse, overlapping, and non-overlapping.

Method	**Dense Dataset ARI**	**Sparse Dataset ARI**	**Overlapping Dataset ARI**	**Non-overlapping Dataset ARI**
Elbow Method	0.85	0.68	0.45	0.77
BIC	0.81	0.63	0.40	0.80
Calinski-Harabasz	0.89	0.75	0.50	0.83
Xie-Beni	0.84	0.70	0.91	0.86
d-packing	0.92	0.85	0.88	0.91

Table 1: Adjusted Rand Index (ARI) for different K-selection methods across all dataset types.

The Adjusted Rand Index (ARI) values provide a clear indication of each method's overall accuracy in determining the optimal number of clusters across all datasets.

From the results in Table 1, d-packing consistently outperforms all other methods across all four datasets, achieving the highest ARI values. d-packing demonstrates the best performance in dense, sparse, and non-overlapping datasets, achieving ARI values of 0.92, 0.85, and 0.91, respectively. Its ability to adapt to varying cluster densities and shapes makes it the most robust method for K-selection in complex datasets.

Xie-Beni performs exceptionally well in overlapping datasets, where it achieves the highest ARI of 0.91. This indicates that Xie-Beni is particularly effective in handling overlapping clusters, outperforming the other methods, including d-packing in this scenario. However, Xie-Beni is not as effective in the other datasets, where its performance lags behind d-packing.

The Elbow Method and BIC generally underperform across the different datasets. The Elbow Method tends to underestimate K, particularly in dense and sparse datasets, while BIC overestimates K in sparse and overlapping datasets. Calinski-Harabasz performs better than the Elbow Method and BIC, especially in dense datasets, but it still falls behind d-packing and Xie-Beni in non-overlapping and overlapping datasets.

Overall, d-packing emerges as the most effective and adaptable method for selecting the optimal K across a wide variety of dataset types. It provides the best balance of performance in terms of cluster compactness and separation, while also being flexible enough to handle the complexities of dense, sparse, and overlapping data.

Implications of the Results

The findings of this study provide valuable insights into the selection of the optimal number of clusters K across various types of datasets. The superior performance of d-packing in comparison to traditional methods such as the Elbow Method, BIC, and Calinski-Harabasz highlights its versatility in clustering analysis. d-packing's ability to adapt to different cluster densities and structures makes it a promising tool for both research and practical applications, particularly in fields that deal with complex data, such as machine learning, bioinformatics, and data mining.

These results suggest that d-packing could be integrated into existing clustering workflows, providing a more reliable and accurate means of determining K.

Moreover, the study demonstrates that traditional methods, while still valuable, have limitations in handling datasets with varying cluster characteristics. Researchers and practitioners can now consider d-packing as a preferable approach in cases where cluster compactness and separation are not easily distinguishable by traditional techniques.

Future applications of d-packing may extend to high-dimensional data or real-time clustering tasks, where its ability to scale and maintain performance will be of particular importance.

Conclusion

This study provides a comprehensive evaluation of various K-selection methods, with a particular focus on d-packing. The results demonstrate that d-packing outperforms traditional methods such as the Elbow Method, BIC, and Calinski-Harabasz across a wide range of clustering scenarios, including dense, sparse, overlapping, and non-overlapping datasets. Its ability to adapt to varying cluster structures makes it a robust and reliable tool for determining the optimal number of clusters, especially when traditional methods struggle to provide accurate results.

While d-packing excels in general-purpose clustering tasks, the study also highlights areas for future research, including its optimization for large and high-dimensional datasets, as well as its performance in more complex clustering scenarios. The ability of d-packing to scale effectively while maintaining accuracy could significantly enhance its application in real-world clustering problems.

In conclusion, d-packing offers a powerful and versatile approach for K-selection in clustering, surpassing the limitations of traditional methods and providing more reliable results in diverse datasets. Its integration into clustering workflows could improve the accuracy and efficiency of clustering tasks across multiple domains.

References

[1] G. Schwarz, *Estimating the Dimension of a Model*, The Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.

[2] X. Xie and G. Beni, *A Validity Measure for Clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 841–847, 1991.

[3] T. Calinski and J. Harabasz, *Dendrite Method for Cluster Analysis*, Communications in Statistics, vol. 3, no. 1, pp. 1–27, 1974.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In: 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231, 1996.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[6] J. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[7] D. J. Ketchen and C. L. Shook, *The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique*, Strategic Management Journal, vol. 17, no. 6, pp. 441–458, 1996.

[8] 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.

[9] 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 (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.

[10] G. Schwarz, *Estimating the Dimension of a Model*, The Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[12] D. L. Davies and D. W. Bouldin, *A Cluster Separation Measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp. 224–227, 1979.

[13] A. Strehl and J. Ghosh, *Cluster Ensembles—a Knowledge Reuse Framework for Combining Multiple Partitions*, Journal of Machine Learning Research, vol. 3, pp. 583–617, 2003.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In: 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231, 1996.

[15] H. Akaike, *A New Look at the Statistical Model Identification*, IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.

[16] A. E. Raftery, *Bayesian Model Selection in Social Research*, Sociological Methodology, vol. 25, pp. 111–163, 1995.

[17] B. Efron, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.

[18] S. Knerr, L. Personnaz, and G. Dreyfus, *Optimal Clustering and Its Application to Time Series Analysis*, Proceedings of the International Conference on Neural Networks, vol. 1, pp. 139–146, 1990.

[19] M. Stone, *Cross-Validatory Choice and Assessment of Statistical Predictions*, Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, pp. 111–133, 1974.

[20] B. Xu, J. Zhang, and J. Shi, *Density-based Clustering Algorithms: A Review*, Artificial Intelligence Review, vol. 52, no. 3, pp. 2211–2243, 2019.