DOI: 10.24996/ijs.2026.67.2.35

ISSN: 0067-2904

Optimal Number of Clusters by Using Four Indexes

Hanin Haqi Ismail*, Tareef Kamil Mustafa
Computer Science, Collage of Science, University of Baghdad, Baghdad, Iraq

Received: 17/ 9/2024 Accepted: 2/2/2025 Published: 28/2/2026

Abstract*:*
In data analysis, "Clustering" has emerged as a mechanism applied in machine learning to group analogous data points or objects together based on their features, attributes, or characteristics. Clustering attempts to detect underlying patterns or structures in data without prior knowledge of group labels. Many algorithms are used in clustering like K-means, one of the most widely used clustering algorithms whose performance depends on the initial point and the value of K. Most clustering techniques need to determine the number of clusters in the beginning. However, in most cases, predicting that value is a high computational cost task. In this paper, an algorithm is designed to compute the proper number of dataset clusters using various cluster validity indexes. The most popular CVIs (clustering validation indexes) are: Elbow method, Silhouette, Gap statistic, and Davis-Bouldin. The paper also proposes a new technique for estimating the appropriate number of clusters (k) depending on their indexes and ranks. The best result of the (ONC) algorithm obtained by the average of silhouette is: (0.501).

Keywords: clustering, K-means, machine learning, Elbow method, Silhouette score, Gap statistic, Davis-Bouldin index.

إيجاد العدد الأمثل للمجموعات باستعمال اربع دلائل

حنين حقي اسماعيل*, طريف كامل مصطفى
علوم حاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصه
في تحليل البيانات، يُعد التجميع الية مطبقة في التعلم الآلي لجمع نقاط البيانات أو الكائنات المماثلة معاً وذلك بناءً على ميزاتها أو سماتها أو خصائصها. تحاول عملية التجميع اكتشاف الأنماط أو الهياكل الأساسية في البيانات دون معرفة مسبقة بتسميات المجموعة. هناك العديد من الخوارزميات المستخدمة في التجميع مثل (K–means) وهي احدى خوارزميات التجميع الأكثر استخداماً والتي يعتمد ادائها على تحديد النقطة الاولية وقيمة (K). تحتاج غالبية تقنيات التجميع الى تحديد عدد المجموعات في البداية. ومع ذلك، في معظم الحالات، يكون التنبؤ بهذه القيمة مهمة ذات تكلفة حسابية عالية لذلك يتناول هذا البحث تصميم خوارزمية لإيجاد العدد الامثل لمجموعات البيانات باستخدام أربع دلائل وأشهرها: Elbow method, Silhouette, Gap statistic and Davis–Bouldin. كما يقترح البحث طريقة جديدة لتقدير العدد المناسب للعناقيد اعتماداً على الدلائل ورتبتها.

*Email: Hanin.Haqi2201m@sc.uobaghdad.edu.iq

Introduction

Machine learning (ML) is founded on the principles of computer science and artificial intelligence (AI)[1]. Clustering is one of the most common unsupervised machine learning techniques. It partitions input datasets into distinct groups[2]. In the context of huge datasets, this method is essential for data-driven knowledge discovery[3]. Because of its versatility to organize data into meaningful categories, clustering is one of the most important techniques in data mining. It is also utilized in many other disciplines, including economics, marketing, medicine, Image processing and pattern recognition[4] and [5].

The k-means algorithm is the clustering method most commonly used. It has gained popularity due to its efficient theory, straightforward algorithms, fast grouping, and ability to handle big data sets[6] and [7]. The process of allocating points to groups and updating the centroids continues in many iterations until convergence is achieved. This approach aims to minimize the within-cluster sum of squared distances, sometimes called inertia or distortion, which assesses cluster compactness. To find a distance between points and centroid, various distance metrics that play a vital role in clustering are calculated using K-means to allocate these items to related clusters, such as Euclidean, Manhattan, Minkowski, and Cosine Distance. It is important to note that the K-means algorithm can converge to a local minimum, meaning the quality of the clustering result can be sensitive to the initial random centroid initialization. To alleviate the issue, the algorithm is frequently executed numerous times, but the difference is the initializations each time, and the best clustering result depends on a chosen evaluation criterion. CVI, which stands for Cluster Validity Index, is an evaluation metric employed to assess the quality and validity of clustering findings. CVIs provide quantitative measurement to compare different clustering solutions and determine the optimal number of clusters. The choice of CVI is based on the specified demand of the clustering and dividing task and the characteristics of the data (points)[8] [9].

There are various clustering types, like partitioning-based, density-based, hierarchical-based and grid-based[10]. The separability between the clusters and the compactness within them are the two cluster properties that serve as the foundation for the definitions of CVIs. CVIs are often calculated as the ratio of a compactness measure's value to a separation measure's value or vice versa. Another way to describe CVIs is through the linear combination of the two metrics. The maximum or minimum of this validity index can be calculated to determine the appropriate number of clusters for a specific dataset[11].
Silhouette Coefficient[12], Elbow method[13], Gap statistic[14], Davis-Bouldin Index[15], and other CVIs are utilized to locate the ideal number of clusters in the points of the dataset. Finding the optimal number of clusters in a dataset is challenging, and it has many limitations and difficulties. Different clustering algorithms and evaluation metrics may yield different results; different metrics may lead to different optimal cluster numbers, and choosing the same evaluation metric may introduce bias and influence the results.

In some cases, the data may not have a clear natural clustering structure, making it difficult to set the optimal number of clusters. The underlying patterns may be complex, overlapping, or spread out, leading to ambiguity in selecting the suitable number of clusters. The computational complexity of clustering algorithms increases with the number of points and dimensions. As the dataset grows in size or complexity, searching for the optimal number of clusters becomes more computationally intensive and time-consuming.[8]

Literature Review

Many previous studies have addressed the issue of finding the optimal number of clusters. The study in[16] proposed a novel discriminant for elbow point. This approach is offered to produce a statistical measure that predicts the appropriate cluster number while dividing a set of points into clusters. The rate of distortion determined by using the Elbow approach is standardized on a scale of (0 - 10). The found values determine the (cos) cosine of the intersected angles of elbow points. The obtained cosine of intersection angles is combined with the theorem of arccosine to calculate the intersection angles; the estimated possible ideal cluster number is determined using the index of the previously calculated least value of intersected angles between elbow points. The outcome results from an experiment based on simulated datasets, and a well-recognized standard dataset (Iris Dataset) applied that the predicted optimal cluster number found by the newly suggested method is superior to the used on wide-scale Silhouette coefficient.

Dinh, Duy-Tai, Tsutomu Fujinami, and Van-Nam Huynh[17] offered a method called k-SCC (k-means- clustering algorithm based on the silhouette analysis approach to estimate the optimal number of clusters in categorical data clustering, namely k-SCC.) for predicting the best k in dividing data clustering. While the clustering is executed, cluster centers are defined by the algorithm using the kernel density estimation approach. Furthermore, it leverages the dissimilarity theorized by the information so that the distance between centers and other items is measured in all clusters. The quality of the clustering results in the previous stage and the optimal value of k are calculated using an approach based on the silhouette analysis. The k-SCC algorithm was tested by comparative experiments on both synthetic and actual world datasets to check which techniques tested were more effective for clustering than three other clustering algorithms. The results show that k-SCC does better than other algorithms to find the best number of clusters for each data set.

Joo, Yeongin[18] suggested a new technique to improve the gap statistic method to estimate k. It has been applied to many datasets, and the findings are superior to the original gap statistics. The novel approach is also suitable for other clustering algorithms that need the integer k. This is because the new methodology, like gap statistics, is compatible with any clustering method.

Rena Nainggolan et al.[19] resolved the weaknesses in the K-Means algorithm through enhancements to improve clustering quality. To find the number of clusters, their method applied the Sum of Squared Error (SSE) technique, where groups have a high degree of member similarity. They also employed the elbow method to enhance the performance of this process by obtaining cluster determination, thereby further boosting the performance of the K-Means algorithm.

In a similar vein, Sagala, Noviyanti TM, and Alexander Agung Santoso Gunawan[20] applied various techniques such as Elbow, Silhouette, Gap Statistics, and Nb-Clust on Major Crime Indicators dataset (MCI) from 2014 to 2019. Their experiments showed that the Elbow, Silhouette, and Nb-Clust methods always found two as the optimal number of clusters. The results of these were verified by the average Silhouette method, where two clusters were found to be the best fit for the data with a Silhouette value of 0.73, indicating very strong cluster cohesion[21].

This research offered a divide-and-conquer technique for completing the same activity in less time. K-means clustering has been found in the time complexity for O (n2). But this algorithm is based on the silhouette score optimal cluster formation, which has to execute at least $\sqrt{N}$ times to identify the best number of the cluster. Therefore, the total complexity of the algorithm comes out to be $O(n2) \times O(\sqrt{n}) \approx O(n5/2)$. The task was completed 2.5 times

faster with the proposed methodology than the iterative method;, but some memory parts costs are needed to store the results. Finally, in this paper, the elbow method, gap statistic, silhouette coefficient, and Davis-Bouldin model are used, and by ranking the results, we can predict the optimal number of clusters in 6 datasets. Table 1 shows the previous studies.

Table 1: Literature Review to find the optimal number of clusters

Name of paper	Problem	Technique (Algorithm)	dataset	result	The value of k
A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm (2021)	Solve the problem of smooth curve of elbow method	Elbow method Silhouette coefficient Cosine law	Iris dataset	The estimated optimal cluster number obtained by our newly proposed method is better than the widely used Silhouette method.	The best value of k to Iris dataset is 3
Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient (2019)	Estimating the number of clusters (k)	K-SCC k-modes	Synthetic and real datasets	Experimental results show that *k*-SCC outperforms the compared algorithms in determining the number of clusters for each dataset.	$K \in [2,10]$
A new approach to determining the optimal number of clusters based on the gap statistic (2020)	Solve the overlapping problem	Gap statistic	Synthetic datasets	Improve the gap statistic method	K=2,3
Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method (2019)	Determining the optimal number of clusters by using the elbow method	k-means SSE (sum square error)	20 data for patient	Build a cluster is better for finding the most optimum cluster center	K=3
Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and Nb-Clust Methods (2021)	Estimate the number of clusters	k-means elbow method gap statistics silhouette nb-clust method	MCI (Major Crime Indicators)	Finding the best optimal number of clusters for Toronto's MCI datasets	K=2
Binning-based silhouette approach to find the optimal cluster using k-means (2022)	Time consuming	k-means silhouette elbow method central limit theorem	Real- world dataset	Proposed approach the task has been completely faster in comparison to the iterative method	

Clustering Types

There are several types of clustering, which can be classified into the following categories: Partitions clustering: It aims to divide the data into K distinct groups or clusters; K-means is one of the most widely used clustering algorithms. The algorithm assigns each data point to the nearest centroid, iteratively refining the centroids to decrease the within-cluster sum of squares[22]

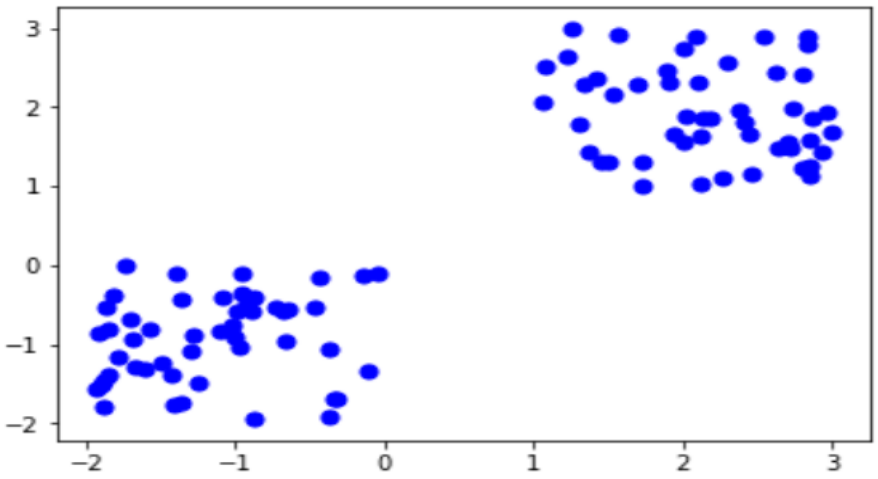

Figure 1: Partition clustering

Hierarchical Clustering: It builds a cluster hierarchy (agglomerative) bottom-up or (divisive) approach. In agglomerative clustering, each data point is given to a separate cluster, and similar clusters are iteratively merged until a stopping condition is met. Divisive clustering starts with all data points in a single cluster and then separates them recursively into smaller groupings[23]

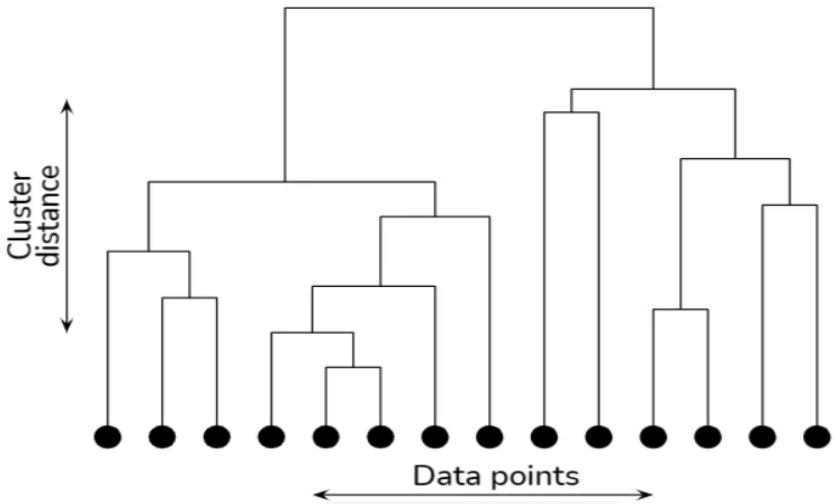

Figure 2: Hierarchical clustering

Density-based Clustering: Density-based clustering algorithms group data points based on their density. They define clusters as areas of higher density separated by lower density areas. The mechanism includes Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points to Identify the Clustering Structure (OPTICS)[24]

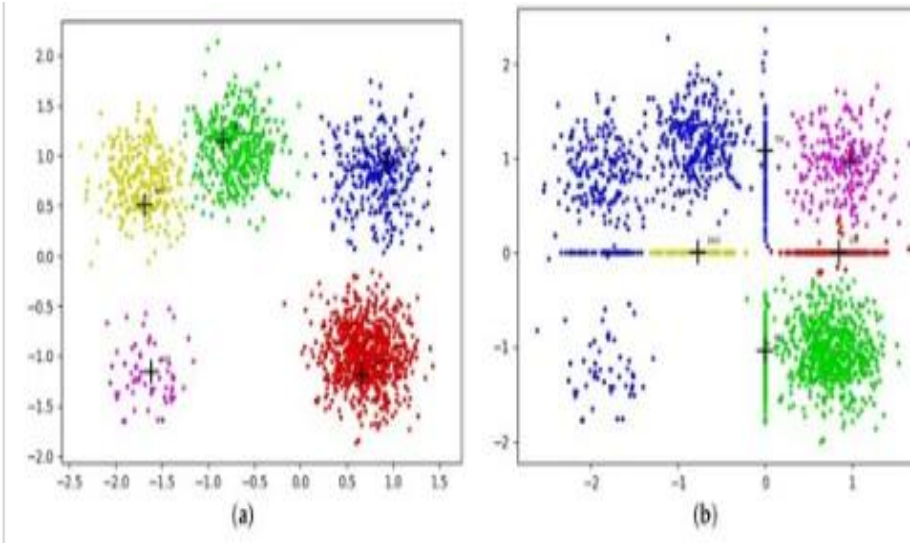

Figure 3: Density-based clustering

Gaussian Mixture Models (GMM): It is a probability model that assumes the data points are created by a mixture of Gaussian distributions. It estimates the parameters of the Gaussian components and assigns data points to different clusters based on their probability of belonging to each component[25]

Fuzzy Clustering: Fuzzy clustering enables data points to be assigned to different clusters with varied membership levels. Instead of assigning data points to a single cluster, fuzzy clustering assigns membership values to each data point, indicating the degree of association with each cluster. Fuzzy C-means (FCM) is a well-known fuzzy clustering algorithm[26] [27]

Spectral Clustering: Spectral clustering utilizes the eigenvalues and eigenvectors of a similarity matrix derived from the data. It involves transforming the data into a lower-dimensional space and performing clustering in that space. Spectral clustering is particularly useful for grouping non-convex or complex-shaped clusters[28]

Distance metrics

Euclidean Distance: Euclidean distance is the most extensively used distance measurement. It is defined as the straight-line distance between two points in Euclidean space[29] [30]

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

d: Euclidean distance
x1,y1: the coordinate of the first point
x2,y2: the coordinate of the second point

Manhattan Distance: The total of absolute differences is also known as city block distance between the coordinates of two points[31] [32]

$$d(x,y) = \sum_{i=1}^{k} |x_i - y_i| \tag{2}$$

d (x, y): Manhattan distance.
(x_i, y_i): The coordinate of points.

Minkowski Distance: is a generalized distance measure that encompasses both Euclidean and Manhattan distances[33]

$$D(x,y) = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q\right)^{1/q} \quad (3)$$

D (x, y): Minkowski distance.
(x_i, y_i): The coordinate of points.
In the Minkowski distance formula, the parameter "q" is a value that determines the type of distance metric being calculated. Depending on the value of "q," the Minkowski distance formula can represent different distance metrics
When q=1, the Minkowski distance reduces to the Manhattan distance.
When q=2, the Minkowski distance represents the Euclidean distance.

Types of indexes

When evaluating the quality of clustering algorithms, several validation indexes can be used to assess the performance of the clustering results[34]
Elbow method: It is a graphical method for estimating the optimal number of clusters in a dataset. The method helps to know the "elbow" point in a plot within-cluster sum of squares (WCSS) or the distortion against the number of clusters. The elbow point illustrates a trade-off between the number of clusters and their compactness[35]

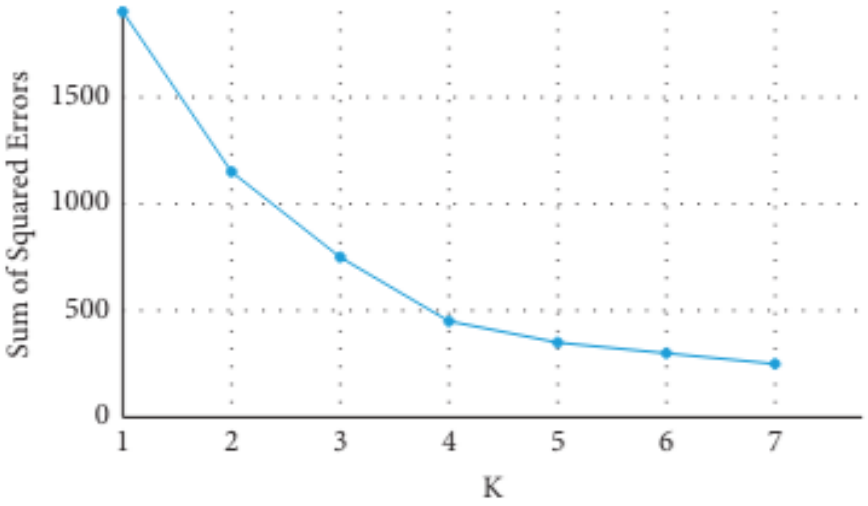

Figure 4: Elbow method [36]

- Silhouette Coefficient: It assesses the compactness and separation of clusters and calculates the average silhouette coefficient for every data point, ranging from -1 to 1. A coefficient Close to 1 suggests well-separated clusters. A value Close to 0 means overlapping status. A minus number implies that data points may be assigned to the wrong clusters[17]

Davies-Bouldin Index: The Davies-Bouldin Index evaluates the average similarity between clusters while considering cluster separation and compactness. It measures the ratio of the average distance between clusters to the average within-cluster distance. Smaller results of this index indicate better clustering, with well-separated and compact clusters[37]
Gap statistic: A statistical approach used to assess the ideal number of clusters in a dataset. It compares the within-cluster dispersion of the data to a reference null distribution to determine if the clustering structure is significant. The method helps identify the number of clusters where the clustering structure is significantly stronger than random [16][38]

Our proposal

To find the optimal number of clusters, which is considered an NB hard problem[39], the Python programming language and its libraries were used in this study. We follow several steps depending on the set of indexes: elbow method, silhouette coefficient, gap statistic, and Davies-Bouldin. Six datasets were used, three of which are standard datasets: iris dataset, wine dataset, and digit dataset. The other three datasets were randomly selected, the first one contains 100 points, the second one contains 10,000 points, and finally, the last one contains 1,000,000 points, representing big data.

The algorithm is a sequence of steps. Initially, we take the dataset and divide the points into groups depending on the distance between them by using the k-means algorithm in a range of clusters from 2 to 11 clusters. After that, we do some calculations, i.e., applying the silhouette coefficient formula for each k (k is the number of clusters), then we use the gap statistic, elbow method, and Davis-Bouldin for all the range of clusters 2, 3, 4…11.

The values resulting from these methods are then arranged and saved in tables, and the silhouette coefficient result is sorted in descending order and ranked from 1 to 10. The value with rank 1 represents the maximum number of silhouette coefficients, the matching number of clusters represents the optimal number in this method, and so on in the gap statistic method and the lowest value of Davis-Bouldin represents the optimal number of clusters.

Algorithm1: Optimal Number of Clusters (ONC)

Input: set of points of any dataset
Output: optimal number of clusters
1: Draw all the points in 2D coordinates
2: for k=2 to k=11 do
3: Apply the k-means to divide the points of the dataset into k clusters and find the center of each cluster
4: Draw the distribution of points and determine each point belonging to any cluster
5: Calculate the silhouette coefficient and gap statistic and davis-bouldin
6: End for
7: Apply the elbow method
8: Calculate the ssd (Sum of Squared Distances)
9: Calculate the difference between the ssd (Sum of Squared Distances)
10: Sort the silhouette coefficient descending and rank the result from 1 to 10, and save the results in a table
11: Sort the gap statistic descending and rank the result from 1 to 10 and save the results in the same table
12: Sort the davis-bouldin ascending and rank the result from 1 to 10 and save the results in the same table
13: Calculate sum of rank in each number of clusters
14: the maximum rank represents the worst value and the matching number of clusters represents the bad number of clusters
15: The minimum rank represents the best value and the matching number of clusters represents the optimal number of clusters.

Experimental results

This section presents the experimental results obtained using our algorithm on the Iris dataset, a well-known and commonly used dataset in machine learning and statistics. It is often used for classification tasks and is a benchmark dataset for evaluating algorithms and models. The dataset contains 150 instances. The results can be summarized in the following table.

Table 2: The optimal number of Iris dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.681	10	0.404	1	6	1	8
3	0.552	11	0.661	2	1	2	5
4	0.498	10	0.780	3	7	3	13
5	0.488	11	0.805	4	2	4	10
6	0.367	11	0.925	5	3	6	14
7	0.356	9	0.968	6	10	8	24
8	0.361	10	0.921	7	8	5	20
9	0.336	11	0.960	8	4	7	19
10	0.319	10	1.027	9	9	10	28
11	0.316	11	1.025	10	5	9	24

The Wine dataset is a popular dataset that contains information about the physicochemical properties of white and red wine.

Table 3: The optimal number of Wine dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.656	9	0.478	1	4	3	8
3	0.571	2	0.534	2	10	5	17
4	0.561	9	0.546	4	5	10	19
5	0.548	9	0.545	6	6	9	21
6	0.565	9	0.465	3	7	2	12
7	0.557	7	0.461	5	9	1	15
8	0.548	11	0.493	7	1	4	12
9	0.525	11	0.534	8	2	6	16
10	0.520	9	0.544	9	8	8	25
11	0.517	10	0.537	10	3	7	20

The Digits dataset, also known as the MNIST dataset, is a popular benchmark dataset frequently used in machine learning. It consists of a collection of images representing handwritten digits

Table 4: The optimal number of Digits dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.118	2.648	2.528	10	1	10	21
3	0.126	2.601	2.438	9	2	9	20
4	0.122	2.569	2.194	8	3	8	19
5	0.137	2.536	2.054	7	4	7	18
6	0.150	2.498	1.955	6	5	6	17
7	0.162	2.474	1.993	5	6	5	16
8	0.174	2.447	1.840	4	7	2	13
9	0.188	2.427	1.756	1	8	1	10
10	0.182	2.412	1.924	2	9	4	15
11	0.182	2.398	1.921	3	10	3	16

This algorithm was applied to 3 random datasets (the first set has 100 points, the second has 10,000 points, and the last has 1,000,000 points). Each run obtained a good partition, and the partitioning of points into suitable clusters depends on the average silhouette coefficient. As a result, the optimal number of clusters can be found in many datasets.
Table 5 shows the result of the clustering of the synthetic dataset containing 100 points divided into 4 clusters.

Table 5: The optimal number of synthetic dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.592	-5.702	0.518	4	9	4	17
3	0.757	-4.688	0.344	2	8	2	12
4	0.793	-3.760	0.288	1	1	1	3
5	0.704	-3.854	0.511	3	2	3	8
6	0.565	-3.979	0.776	5	3	6	14
7	0.467	-4.076	0.7405	6	5	5	16
8	0.452	-4.071	0.916	8	4	8	20
9	0.465	-4.138	0.857	7	7	7	21
10	0.348	-4.097	0.948	9	6	9	24

Table 6 shows the result of clustering of synthetic dataset containing 10000 points divided into 2 clusters.

Table 6: The optimal number of synthetic dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.822	-2.959	0.250	1	1	1	3
3	0.566	-3.237	0.910	2	2	3	7
4	0.309	-3.484	1.257	7	5	9	21
5	0.315	-3.480	1.080	4	3	8	15
6	0.310	-3.552	1.047	6	7	7	20
7	0.319	-3.481	0.943	3	4	5	12
8	0.309	-3.522	0.942	8	6	4	18
9	0.301	-3.586	0.976	9	8	6	23
10	0.314	-3.596	0.883	5	9	2	16

Table 7 shows the result of the clustering of synthetic dataset containing 1000000 points into 6 clusters.

Table 7: The optimal number of synthetic dataset

No.of cluster	silhouette	Gap statistic	Davis	Rank of silhouette	Rank of gap	Rank of Davis	Sum of ranks
2	0.599	-4.747	0.640	1	7	3	11
3	0.576	-4.773	0.623	3	9	2	14
4	0.532	-4.753	0.671	6	8	5	19
5	0.544	-4.668	0.642	4	6	4	14
6	0.586	-4.160	0.574	2	1	1	4
7	0.537	-4.266	0.745	5	2	6	13
8	0.475	-4.375	0.902	7	3	7	17
9	0.446	-4.453	0.973	8	4	8	20
10	0.412	-4.462	1.048	9	5	9	23

The best result of silhouette is 0.8 when dividing a synthetic dataset containing 1000 points into 2 clusters.

Conclusion

This study investigated a modification of the K-Means algorithm to locate the optimal cluster center using the minimal sum of squared errors (SSE) value. The K-Means cluster is implemented with the best cluster center. The number of clusters can be determined as a critical step in cluster analysis. This study proposes using the sum of ranks for the set of index clustering in the interior validity index to assess the right or most appropriate number of clusters. The Manhattan distance performs better than the Euclidean distance method, which is based on iris, wine, digits, and three other random datasets that become increasingly high and low. Using the elbow method, silhouette coefficient, Davis-Bouldin, and gap statistic methods, most validity indexes can be determined in predicting the optimal number of clusters in any dataset.

References

[1] A. F. Jahwar and A. M. Abdulazeez, "Meta-heuristic algorithms for K-means clustering: A review," *PalArch's J. Archaeol. Egypt/Egyptology*, vol. 17, no. 7, pp. 12002–12020, 2020.

[2] C. Patil and I. Baidari, "Estimating the optimal number of clusters k in a dataset using data depth," *Data Sci. Eng.*, vol. 4, pp. 132–140, 2019. :10.1007/s41019-019-0091-y

[3] F. Hassan and S. F. Behadili, "Modeling Social Networks using Data Mining Approaches-Review," *Iraqi J. Sci.*, pp. 1313–1338, 2022. :10.24996/ijs.2022.63.3.35

[4] D. M. Saputra, D. Saputra, and L. D. Oswari, "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method," in *Sriwijaya International Conference on information technology and its applications (SICONIAN 2019)*, 2020, pp. 341–346. :10.2991/aisr.k.200424.051

[5] I. Alkanani and R. Fawzi, "Fuzzy Linear Discriminant Analysis Clustering With Its Application," *Iraqi J. Sci.*, vol. 54, no. Mathematics conf, pp. 739–743, 2013. https://orcid.org/0000-0003-0106-6688

[6] W. A. Abbas, "Genetic Algorithm-Based Anisotropic Diffusion Filter and Clustering Algorithms for Thyroid Tumor Detection," *Iraqi J. Sci.*, pp. 1016–1026, 2020. 10.24996/ijs.2020.61.5.10

[7] N. Fouad and S. M. Hameed, "Genetic Algorithm based Clustering for Intrusion Detection," *Iraqi J. Sci.*, pp. 929–938, 2017. https://doi.org/10.58496/MJCS/2024/011

[8] **S. Q. Noor and T. K. Mustafa,** "Comparing K-Means, Nearest Neighbor, and Lloyd's clustering algorithms," *Iraqi J. Sci.*, vol. 65, no. 11, pp. 6688–6704, 2024, doi:10.24996/ijs.2024.65.11.40.

[9] Y. Hussein and S. A. Jalil, "Proposed KDBSCAN algorithm for clustering," *Iraqi J. Sci.*, vol. 59, no. 1 A, pp. 173–178, 2018, doi: 10.24996/IJS.2018.59.1A.18. https://doi.org/10.21123/bsj.2024.9516

[10] T. M. Ghazal, "Performances of k-means clustering algorithm with different distance metrics," *Intell. Autom. \& Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021. :10.32604/iasc.2021.019067

[11] E. Zhu, Y. Zhang, P. Wen, and F. Liu, "Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index," *Neurocomputing*, vol. 363, pp. 149–170, 2019. 10.1088/1755-1315/1083/1/012082

[12] N. Rohman and A. Wibowo, "Clustering of popular Spotify songs in 2023 using k-means method and silhouette coefficient," *J. Pilar Nusa Mandiri*, vol. 20, no. 1, pp. 18–24, 2024. :10.33480/pilar.v20i1.4937

[13] L. Pamungkas, N. A. Dewi, and N. A. Putri, "Classification of Student Grade Data Using the K-Means Clustering Method," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 86–91, 2024. 10.54209/jurnalinstall.v16i02.201

[14] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4995–5005, 2024. 10.1007/s00521-023-09339-6

[15] M. Cherradi, "Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques," *Seminars in Medical Writing and Education*, 2024, p. 68.

10.56294/mw202468

[16] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, pp. 1–16, 2021. 10.21203/rs.3.rs-58011/v3

[17] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient," in *Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29--December 1, 2019, Proceedings 20*, 2019, pp. 1–17.

[18] J. Yang, J.-Y. Lee, M. Choi, and Y. Joo, "A new approach to determine the optimal number of clusters based on the gap statistic," in *International Conference on Machine Learning for Networking*, 2019, pp. 227–239.

[19] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method," in *Journal of Physics: Conference Series*, 2019, p. 12015. :10.1088/1742-6596/1361/1/012015

[20] N. T. M. Sagala and A. A. S. Gunawan, "Discovering the optimal number of crime cluster using elbow, Silhouette, gap statistics, and NbClust methods," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 1–10, 2022.

[21] M. SUBRAMANIAN, "Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means". 10.1109/ACCESS.2022.3215568

[22] M. Sarkar, A. R. Puja, and F. R. Chowdhury, "Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis," *J. Bus. Manag. Stud.*, vol. 6, no. 2, pp. 54–60, 2024. :10.3390/jtaer19030081

[23] L. L. Gao, J. Bien, and D. Witten, "Selective inference for hierarchical clustering," *J. Am. Stat. Assoc.*, vol. 119, no. 545, pp. 332–342, 2024.

[24] F. Amjad, E. B. Agyekum, and N. Wassan, "Identification of appropriate sites for solar-based green hydrogen production using a combination of density-based clustering, Best-Worst Method, and Spatial GIS," *Int. J. Hydrogen Energy*, vol. 68, pp. 1281–1296, 2024. 10.1016/j.ijhydene.2019.10.099

[25] H. Guan *et al.*, "Improved Gaussian mixture model to map the flooded crops of VV and VH polarization data," *Remote Sens. Environ.*, vol. 295, p. 113714, 2023. 10.1016/j.rse.2023.113714

[26] Y. Tang, J. Huang, W. Pedrycz, B. Li, and F. Ren, "A fuzzy clustering validity index induced by triple center relation," *IEEE Trans. Cybern.*, vol. 53, no. 8, pp. 5024–5036, 2023.

[27] **I. H. Alkanani and R. M. Fawzi,** "Fuzzy linear discriminant analysis clustering with its application," *Iraqi J. Sci.*, vol. 54, no. 3, pp. 739–743, 2013.[28] L. Yu, J. Gu, and S. Volgushev, "Spectral clustering with variance information for group structure estimation in panel data," *J. Econom.*, vol. 241, no. 1, p. 105709, 2024.

[28] Q. Liu, J. Ma, X. Zhao, K. Zhang, K. Xiangli, and D. Meng, "A novel method for fault diagnosis and type identification of cell voltage inconsistency in electric vehicles using weighted Euclidean distance evaluation and statistical analysis," *Energy*, vol. 293, p. 130575, 2024.

[29] B. K. Abd, N. A. Z. Abdullah, and Q. K. Abood, "Hand written signature verification based on geometric and grid features," *Iraqi J. Sci.*, pp. 1800–1809, 2015.

[30] H. Sabah Talabani, H. M. T. Abdulhadi, and M. H. Ali, "Obfuscated Malware Memory Detection Employing Lazy Instance Based Learner Algorithm Based On Manhattan Distance Function," *Passer J. Basic Appl. Sci.*, vol. 6, no. 1, pp. 130–137, 2024. 10.33545/27076571.2024.v5.i1a.86

[31] E. F. Nasser and A. A. A. Karim, "Object Tracking and matching in a Video Stream based on SURF and Wavelet Transform," *Iraqi J. Sci.*, pp. 939–950, 2017.

[32] Z. Yang, R. Zhu, and W. Liao, "Minkowski Distance Based Pilot Protection for Tie Lines Between Offshore Wind Farms and MMC," *IEEE Trans. Ind. Informatics*, 2024. :10.1109/TII.2024.3369668

[33] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020. :10.1109/ACCESS.2020.2988796

[34] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, 2018, p. 12017. :10.1088/1757-

899X/336/1/012017

[35] R. Sammouda and A. El-Zaart, "An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, p. 4553832, 2021. 10.1155/2021/4553832

[36] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in *IOP Conference Series: Materials Science and Engineering*, 2020, p. 12128.

[37] A. Satre-Meloy, M. Diakonova, and P. Grünewald, "Cluster analysis and prediction of residential peak demand profiles using occupant activity data," *Appl. Energy*, vol. 260, p. 114246, 2020. 10.1016/j.apenergy.2019.114246

[38] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Appl. Sci.*, vol. 11, no. 23, p. 11202, 2021. 10.3390/app112311202