Improved K-Means Algorithm: Integrating Density Peaks and Adaptive K-Value for Mall Customer Segmentation

Gong Junmei , Lai Dan , Liu Yang *

Posted Date: 17 April 2026

doi: 10.20944/preprints202604.1238.v1

Keywords: K-means algorithm; density peak; adaptive K-value; outlier detection; customer segmentation; data mining

 doi:10.20944/preprints202604.1238.v1

Article

Improved K-Means Algorithm: Integrating Density Peaks and Adaptive K-Value for Mall Customer Segmentation

Gong Junmei [1], Lai Dan [1] and Liu Yang [2,*]

[1] Chengdu Jincheng College

[2] University of Electronic Science and Technology of China

* Correspondence: ly2015@uestc.edu.cn

Abstract

Customer segmentation is a core application of data mining in the retail industry. Tra-ditional K-means clustering is widely adopted here for its simple principle and high computational efficiency, yet it has notable drawbacks: random initial clustering centers easily lead to local optimal solutions, it is highly sensitive to abnormal data, and the cluster number K relies on manual experience, resulting in unstable clustering perfor-mance. This paper designs an improved K-means algorithm, which filters outliers through a two-layer mechanism combining Local Outlier Factor and distance threshold. It also constructs a multi-index system with Silhouette Coefficient, Calinski-Harabasz and Davies-Bouldin indices to automatically determine the optimal K-value, optimizes initial centers via density peak clustering, and introduces weighted Euclidean distance to enhance clustering compactness. Experiments on the extended large-scale mall customer dataset (1,500 samples compare the proposed algorithm with traditional K-means, K-medoids and DBSCAN. Results show it achieves a Silhouette Coefficient of 0.5821, a CH index of 1025.36 and a DB index of 0.5107, outperforming all comparison algorithms in all indi-cators with more reasonable and stable clustering results. Applied to mall customer segmentation, this algorithm divides customers into 5 groups with distinct character-istics, providing solid data support for malls to formulate scientific and differentiated marketing strategies.

Keywords: K-means algorithm; density peak; adaptive K-value; outlier detection; customer segmentation; data mining

MSC: 62H30; 68T10; 90B60

1. Introduction

To enhance core competitiveness, malls rely on data mining technology for precise customer segmentation. With the continuous promotion of the new retail model and the accumulation of massive consumption data, this method has become a key means for mall operation [1]. By mining the characteristics of customers' consumption behavior, grouping customers with similar attributes into the same category, and formulating exclusive product recommendations, promotional activities and service plans for different groups, malls can effectively improve customer stickiness and consumption conversion rate [2].

Clustering analysis is one of the core algorithms in unsupervised learning and the mainstream technical means for customer segmentation. The K-means algorithm is the most widely used in retail customer segmentation scenarios due to its low computational complexity, strong scalability and fast convergence speed [3]. However, the inherent defects of the traditional K-means algorithm lead to great limitations in practical applications: random selection of initial clustering centers easily causes the algorithm to converge to local optimal solutions, resulting in significant differences in clustering

 doi:10.20944/preprints202604.1238.v1

results from different experiments; the number of clusters K must be manually specified, relying on the researcher's empirical judgment without objective data support; the algorithm is extremely sensitive to outliers, and a small number of outlier samples will seriously interfere with the calculation of clustering centers and reduce clustering accuracy [4,5]. Scholars at home and abroad have conducted a large number of improvement studies on these problems, focusing on three directions: initial center optimization, adaptive K-value selection and outlier processing.

For the optimization of initial centers, existing studies mostly use heuristic algorithms or density clustering algorithms to adjust the selection logic, such as selecting the farthest initial centers based on the K-means++ algorithm [6] and choosing high-density samples as centers according to density peaks [7], which reduce the probability of local optimal problems. For the determination of adaptive K-value, most studies use a single index (such as the elbow method and Silhouette Coefficient) to find the optimal K-value [8], but such indexes are easily disturbed by data distribution, leading to inadequate evaluation results. For outlier processing, most studies only use a single outlier detection algorithm (such as LOF and Z-score) [9], which lacks the ability to filter outliers in the face of complex datasets. Most existing improved algorithms only optimize a single problem, failing to comprehensively solve multiple defects of the K-means algorithm, and there are few targeted verifications in the retail customer segmentation scenario.

Aiming at the shortcomings of existing research, this paper develops an improved K-means algorithm, which integrates two-layer outlier filtering, multi-index adaptive K-value, density peak initial centers and weighted distance to optimize the performance of the traditional K-means algorithm, and applies it to the mall customer segmentation scenario. The experiment adopts the public Mall Customer Segmentation dataset, compares the improved algorithm with other algorithms to verify its reliability, and analyzes the characteristics of different customer groups based on clustering results to provide practical reference for mall operation decisions. The innovations of this paper are as follows: (1) Construct a two-layer outlier filtering mechanism of LOF + distance threshold to improve the robustness of the dataset; (2) Automatically select the objective K-value using three core clustering evaluation indexes to avoid errors caused by manual experience; (3) Combine the density peak algorithm with the weighted Euclidean distance to solve the problems of random initial centers and insufficient clustering compactness simultaneously; (4) Verify the effectiveness of the algorithm in the real scenario of mall customer segmentation, integrating theoretical research with practical application. Additionally, this study incorporates insights from recent research, including Wang and Samonte (2026) [23], Ding et al. (2025) [21], Tatikonda et al. (2025) [18], Wang (2025) [19], Ling and Weiling (2025), and Zhao (2024) [21], to further enhance the algorithm's robustness and applicability.

2. Related Theoretical Foundations

2.1. Traditional K-Means Clustering Algorithm

Proposed by MacQueen in 1967, the K-means algorithm is a partitioning-based clustering algorithm that divides all samples into K completely independent clusters, maximizing the similarity of samples within the same cluster and minimizing the similarity of samples between different clusters[10]. The algorithm takes the sum of squared errors (SSE) from samples to the corresponding cluster centers as the objective function, as shown in Formula (1):

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \| x - \mu_i \|^2 \quad (1)$$

where C_i represents the i-th cluster, μ_i is the clustering center of the cluster, and $\| x - \mu_i \|$ refers to the Euclidean distance between the sample x and the cluster center μ_i.

The operation process of the traditional K-means algorithm is as follows: (1) Randomly select K samples as initial clustering centers; (2) Calculate the Euclidean distance from each sample to the K

 doi:10.20944/preprints202604.1238.v1

clustering centers, and assign the sample to the cluster with the nearest distance; (3) Recalculate the sample mean of each cluster and update the clustering centers; (4) Repeat steps (2) and (3) until the clustering centers no longer change significantly or the maximum number of iterations is reached, and the algorithm converges.

The time complexity of the traditional K-means algorithm is $O(nKt)$ (where n is the number of samples, K is the number of clusters, and t is the number of iterations), which still has high computational efficiency on large datasets. However, the defects of random initial centers, manual K-value specification and sensitivity to outliers lead to difficulty in guaranteeing the stability and accuracy of its clustering results.

2.2. Density Peak Clustering (DPC) Algorithm

Proposed by Rodriguez and Laio in 2014, the density peak clustering algorithm is based on the core idea that clustering centers are usually located in regions with high sample density and far from other high-density samples [7]. The algorithm calculates two characteristic quantities for each sample: local density and relative distance, and selects samples with both large local density and relative distance as clustering centers. It does not require pre-specifying the number of clusters and has good clustering performance for datasets with non-spherical distribution.

The local density ρ_i represents the density of samples around a given sample δ_i, calculated using the cutoff distance d_c, as shown in Formula (2):

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c) \tag{2}$$

where d_{ij} is the Euclidean distance between sample i and sample j, $\chi(x)$ is an indicator function with $\chi(x) = 1$ if $x < 0$, otherwise $\chi(x) = 0$. The relative distance represents the minimum distance from a sample to all samples with higher density than it, as shown in Formula (3):

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

For the sample with the maximum density in the dataset, its δ_i is the maximum distance from this sample to all other samples.

2.3. Clustering Evaluation Indexes

Proposed by Rodriguez and Laio in 2014, the density peak clustering algorithm is based on the core idea that clustering centers are usually located in regions with high sample density and far from other high-density samples [7]. The algorithm calculates two characteristic quantities for each sample: local density and relative distance, and selects samples with both large local density and relative distance as clustering centers. It does not require pre-specifying the number of clusters and has good clustering performance for datasets with non-spherical distribution.

To objectively judge the actual clustering performance, this paper selects the Silhouette Coefficient, Calinski-Harabasz (CH) index and Davies-Bouldin (DB) index as the core evaluation indexes. These three indexes belong to different dimensions, can measure the effectiveness of clustering, and are classic and commonly used evaluation indexes in the field of unsupervised clustering [11]

(1) Silhouette Coefficient: It considers both intra-cluster compactness and inter-cluster separation, with values in the interval [−1,1]. The closer the value is to 1, the better the clustering performance, and the higher the intra-cluster compactness and inter-cluster separation of samples. The Silhouette Coefficient of a single sample is shown in Formula (4), and the overall Silhouette Coefficient is the average of the Silhouette Coefficients of all samples:

 doi:10.20944/preprints202604.1238.v1

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4}$$

where $a(i)$ refers to the average distance between the i-th sample and all other samples in the same cluster, and $b(i)$ is the minimum average distance from the i-th sample to all other clusters.

(2) CH Index: It judges the clustering performance by the ratio of between-cluster variance to within-cluster variance. A larger value indicates greater differences between clusters, more concentrated distribution of samples within clusters, and better clustering performance, calculated as shown in Formula (5):

$$CH = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n - K}{K - 1} \tag{5}$$

where $tr(B_k)$ is the trace of the between-cluster scatter matrix, $tr(W_k)$ is the trace of the within-cluster scatter matrix, n is the number of samples, and K is the number of clusters.

(3) DB Index: It calculates the average similarity between all clusters, with a value range of $[0, +\infty)$. The closer the value is to 0, the lower the similarity between clusters, and the better the clustering performance, calculated with reference to Formula (6):

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{6}$$

where s_i represents the average distance from all samples in the i-th cluster to the cluster center, and d_{ij} represents the Euclidean distance between the i-th cluster center and the j-th cluster center.

3. Related Theoretical Foundations

Aiming at four deficiencies of the traditional K-means algorithm, this paper optimizes outlier processing, adaptive K-value selection, initial center optimization and distance calculation, and proposes an improved K-means algorithm. The overall framework of the algorithm is shown in Figure 1.

First, the raw dataset is standardized and processed with two-layer outlier filtering to obtain a high-quality cleaned dataset. Then, the optimal number of clusters K is automatically selected by the multi-index comprehensive evaluation system. Next, the K optimal initial clustering centers are determined by the density peak clustering algorithm. Finally, the weighted Euclidean distance is introduced to optimize the iteration process of the traditional K-means algorithm to obtain the final clustering results.

 doi:10.20944/preprints202604.1238.v1

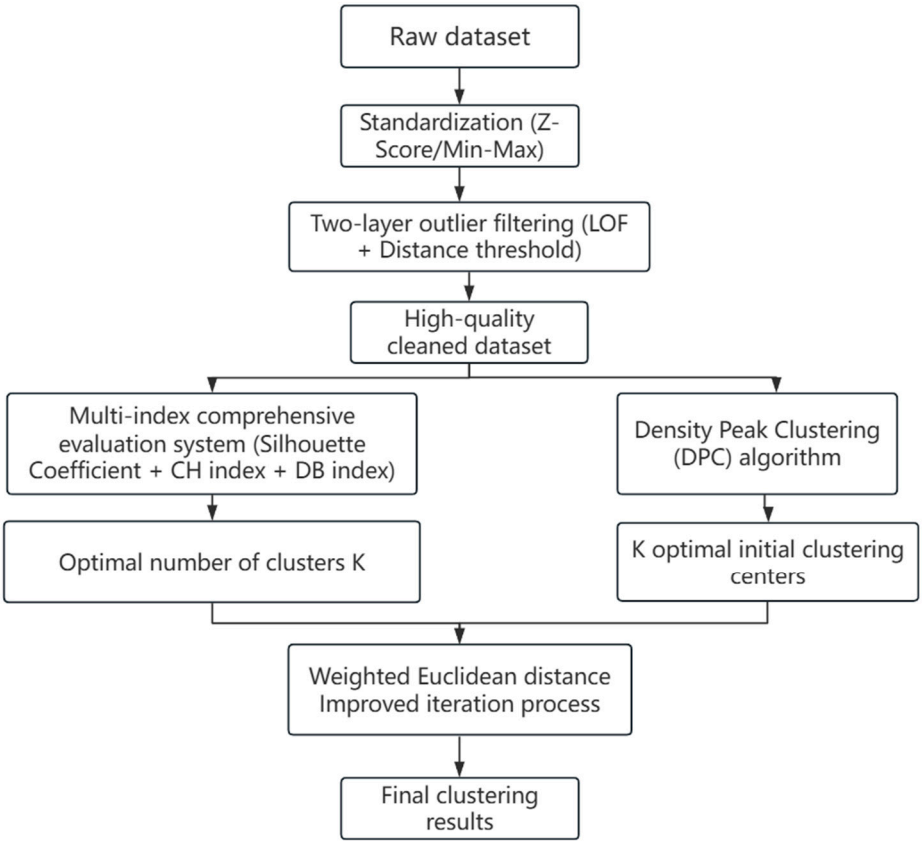


Figure 1. Overall Execution Framework of the Improved K-means Algorithm.

3.1. Traditional K-Means Clustering Algorithm

3.1.1. Data Standardization

Mall customer data has different dimensions for different features (e.g., annual income is in thousands of US dollars, and spending score is a dimensionless value from 1 to 100). Direct calculation will make the distance results affected by dimensions. This paper adopts the Z-score standardization method for data normalization, mapping all features to a standard normal distribution with a mean of 0 and a variance of 1, as shown in Formula (7):

$$x^* = \frac{x - \mu}{\sigma} \tag{7}$$

where x represents the value of the original feature, μ is the mean of the feature, σ is the standard deviation of the feature, and x^* is the standardized feature value.

3.1.2. Two-Layer Outlier Filtering

Outliers refer to outlier samples in the dataset that deviate significantly from the overall distribution. The existence of such samples will interfere with the calculation of clustering centers of the K-means algorithm and distort the clustering results. Traditional single outlier detection algorithms have insufficient filtering effects. This paper proposes a two-layer outlier filtering mechanism of LOF + distance threshold, which first detects local outlier samples with the LOF algorithm and then filters global outlier samples with the distance threshold to achieve accurate outlier filtering.

 doi:10.20944/preprints202604.1238.v1

(1) First layer: LOF local outlier detection: The LOF algorithm calculates the Local Outlier Factor of each sample to judge the deviation degree of the sample from the surrounding data. A higher LOF value indicates a higher abnormal degree of the sample[9]. In this paper, the outlier ratio is set to 4%, and samples with LOF values exceeding the threshold are labeled as outliers.

(2) Second layer: Distance threshold global filtering: Taking the overall center of the dataset as the reference, this paper calculates the Euclidean distance from each sample to the center, marks samples with distances exceeding the 95th percentile as global outliers, and filters outlier samples far from the main body of the data.

Finally, only samples that are normal in both rounds of detection are retained to construct the input dataset for subsequent clustering analysis, thereby improving data quality and avoiding the interference of outliers on clustering results.

3.2. Adaptive K-Value Selection Based on Multi-Index Fusion

The final clustering performance of the K-means algorithm is completely determined by the value of the number of clusters K. The traditional manual specification of K has poor objectivity. Most existing related studies only use a single index to select the K-value, which is easily affected by data distribution and leads to biased results. This paper integrates three core clustering evaluation indexes (Silhouette Coefficient, CH index and DB index) to construct a multi-index comprehensive evaluation system for adaptive and objective selection of the K-value, with the specific steps as follows:

Step 1: Determine the search interval of the K-value $[2,12]$, substitute different K-values one by one to perform clustering with the traditional K-means algorithm, and calculate the corresponding Silhouette Coefficient S, CH index C and DB index D respectively.

Step 2: Normalize the three indexes to eliminate dimensional differences. The Silhouette Coefficient and CH index are positive indexes (larger values indicate better performance) and normalized using Formula (8); the DB index is a negative index (smaller values indicate better performance), which is first negated and then normalized using Formula (8). The normalized indexes S^*, C^* and D^* are finally obtained:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{8}$$

where x represents the original index value, and x_{max} and x_{min} are the maximum and minimum values of the index in the set search interval, respectively.

Step 3: Construct the comprehensive evaluation score, and assign corresponding weights according to the evaluation ability of each index. The Silhouette Coefficient and CH index each account for 40%, and the DB index accounts for 20%, as shown in Formula (9):

$$F = 0.4S^* + 0.4C^* + 0.2D^* \tag{9}$$

Step 4: Find the maximum comprehensive evaluation score among all K-values, take the corresponding K as the optimal number of clusters, and complete the adaptive selection of the K-value.

3.3. Initial Clustering Center Optimization Based on Density Peaks

The selection of initial clustering centers is a core link of the K-means algorithm, which directly determines whether the algorithm will fall into local optimal solutions. This paper draws on the core idea of the density peak clustering algorithm to select samples with high local density and large spacing from other high-density samples as initial clustering centers, ensuring that these centers are representative and well-distributed, with the specific steps as follows:

 doi:10.20944/preprints202604.1238.v1

Step 1: Process the cleaned dataset, calculate the Euclidean distance between every two samples and organize it into a matrix, where $D=[d_{ij}]_{n\times n}$ represents the Euclidean distance between sample i and sample j, and n is the total number of samples;

Step 2: Calculate the local density ρ_i of each sample, and select the 12th percentile of the distance matrix d_c as the cutoff distance to balance the resolution and computational efficiency of density calculation;

Step 3: Calculate the relative distance δ_i between each sample, and take the maximum value of the distance matrix for the sample with the highest density as its relative distance;

Step 4: Calculate the comprehensive score $\gamma_i=\rho_i\times\delta_i$ of each sample, where a higher score γ_i indicates that the sample is more suitable as a clustering center;

Step 5: Select the K samples with the highest comprehensive scores γ_i from all samples as the initial clustering centers of the improved K-means algorithm, where K is the optimal number of clusters selected adaptively in Section 2.2.

3.4. Iterative Optimization Based on Weighted Euclidean Distance

The traditional K-means algorithm uses the Euclidean distance to calculate the similarity between samples and clustering centers, assigning equal weights to all features and being easily disturbed by noise. This paper replaces the original distance calculation method with the weighted Euclidean distance, sets weight coefficients for different features to improve clustering compactness, and enables samples to cluster faster into clusters with higher feature similarity. The formula of the weighted Euclidean distance is shown in Formula (10):

$$d(x,\mu_i)=\|x-\mu_i\|^w=\left(\sum_{k=1}^{m}(x_k-\mu_{ik})^2\right)^{\frac{w}{2}} \tag{10}$$

where m is the number of features, x_k is the k-th feature value of sample x, μ_{ik} is the k-th feature value of the i-th cluster center, and w is the weight coefficient. Through multiple experimental verifications, the clustering performance is optimal when $w=1.3$ in this paper.

During the iteration of the K-means algorithm, the traditional Euclidean distance is replaced with the weighted Euclidean distance, the objective function is redefined, and sample assignment and clustering center update are performed until the algorithm stops iterating.

3.5. Complete Execution Steps of the Improved K-Means Algorithm

Combining the various improvement methods mentioned above, the complete execution process of the improved K-means algorithm is summarized as follows:

Input the raw dataset X and perform Z-score standardization to obtain the standardized dataset X^*;

Adopt the two-layer outlier filtering mechanism of LOF + distance threshold to clean the standardized dataset X^* and obtain an outlier-free dataset X_{clean};

Select the optimal number of clusters K adaptively K_{opt} based on the multi-index fusion comprehensive evaluation system;

 doi:10.20944/preprints202604.1238.v1

Calculate the local density ρ_i and relative distance of samples δ_i in the cleaned dataset, and select the K_{opt} samples with the highest comprehensive scores $\gamma_i = \rho_i \times \delta_i$ as initial clustering centers $C = \{c_1, c_2, \ldots, c_{K_{opt}}\}$;

Calculate the distance from each sample to the initial clustering centers using the weighted Euclidean distance $(w = 1.3)$, and assign the sample to the cluster with the nearest distance;

Calculate the sample mean of each cluster and update the clustering centers;

Judge whether the clustering centers converge (the center variation is less than the threshold) or reach the maximum number of iterations (200 times). If convergent, output the clustering results; otherwise, return to Step 5 to continue the iteration.

4. Experimental Design and Result Analysis

4.1. Experimental Dataset

This paper selects an extended large-scale mall customer dataset for the experiment. It covers 5 features of 1,500 mall member customers, including CustomerID, Gender, Age, Annual Income (k$) and Spending Score (1-100). The spending score is a comprehensive evaluation result given by the mall based on customers' consumption behavior characteristics such as consumption frequency and consumption amount, with a value ranging from 1 to 100. A higher value indicates higher customer consumption activity.

In line with the actual business needs of mall customer segmentation, this paper selects annual income and spending score as the core clustering features, which can directly reflect customers' consumption capacity and consumption willingness and are key dimensions for customer segmentation. Data preprocessing is completed before the experiment: first, remove the CustomerID feature with no practical analysis value; then, perform Z-score standardization on annual income and spending score; finally, filter out abnormal samples with the two-layer outlier filtering mechanism. A total of 1,392 valid samples are obtained for subsequent clustering experiments.

4.2. Experimental Environment and Comparison Algorithms

4.2.1. Experimental Environment

The experiment is conducted in the Python 3.9 programming environment with PyCharm 2023.2 as the development tool. The third-party libraries used include: Pandas 2.1.4 for data processing, NumPy 1.26.2 for numerical calculation, Scikit-learn 1.3.2 for running machine learning algorithms, Matplotlib 3.8.2 for data visualization, and Scipy 1.11.4 for scientific calculation. The hardware configuration of the experiment is Intel Core i7-12700H CPU, with 16 GB DDR4 memory and 512 GB SSD solid-state drive.

4.2.2. Comparison Algorithms

1. Traditional K-means algorithm: The KMeans module in the Scikit-learn library, with n_init=20 set to randomly select initial centers to eliminate the influence of random factors;
2. K-medoids algorithm: Also known as the PAM algorithm, it uses samples (medoids) within the cluster instead of the mean as the clustering center, has a certain robustness to outliers, and is a classic improved version of the K-means algorithm;
3. DBSCAN algorithm: A density-based unsupervised clustering algorithm that does not require pre-specifying the number of clusters, can effectively detect clusters with non-spherical distribution, and has a natural ability to filter outliers.

4.3. Experimental Evaluation Indexes

 doi:10.20944/preprints202604.1238.v1

This paper uses three unsupervised clustering evaluation indexes (Silhouette Coefficient, CH index and DB index) to quantitatively evaluate the clustering performance of all algorithms, and records the running time of the algorithms. The advantages of the improved algorithm are comprehensively verified from two aspects: clustering performance and computational efficiency. The evaluation criteria of each index are shown in Table 1.

Table 1. Experimental Evaluation Indexes and Evaluation Criteria.

Evaluation Index	Value Range	Evaluation Criterion
Silhouette Coefficient	[-1,1]	The closer to 1, the better the clustering performance
CH Index	[0,+∞)	Larger values indicate better clustering performance
DB Index	[0,+∞)	The closer to 0, the better the clustering performance
Running Time (s)	[0,+∞)	Shorter values indicate higher computational efficiency

4.4. Experimental Results and Analysis

4.4.1. Adaptive K-Value Selection Results

To ensure clustering quality, this paper adopts the two-layer outlier filtering mechanism of LOF combined with distance threshold to clean the standardized customer data and eliminate abnormal samples deviating from the main distribution, with the filtering effect shown in Figure 2.

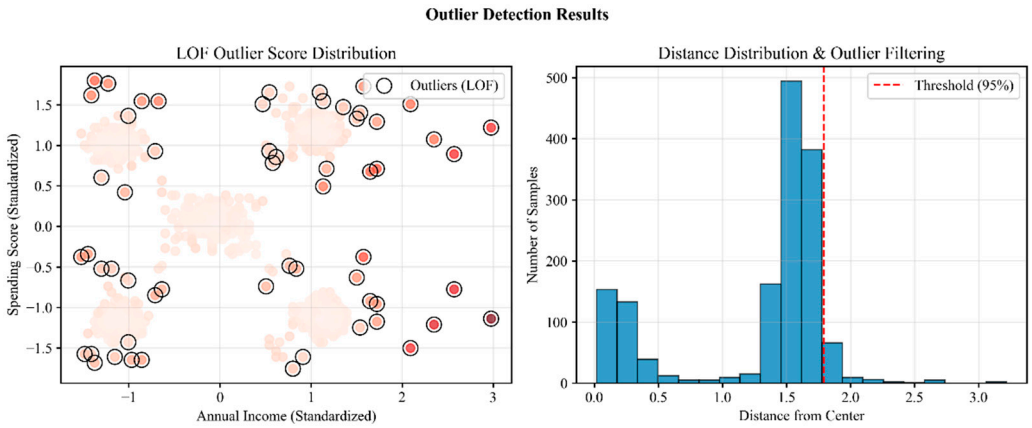


Figure 2. Visualization of the Two-layer Outlier Filtering Effect on the Mall Customer Dataset.

The LOF algorithm successfully identifies locally deviated samples in the left figure, and the distance histogram in the right figure clearly marks the filtering threshold. After screening, 1,392 high-quality samples are retained from the original 1,500 samples for subsequent clustering experiments. After cleaning the dataset, this paper uses the multi-index fusion method to determine the optimal K-value, with the relevant results shown in Figure 3.

 doi:10.20944/preprints202604.1238.v1

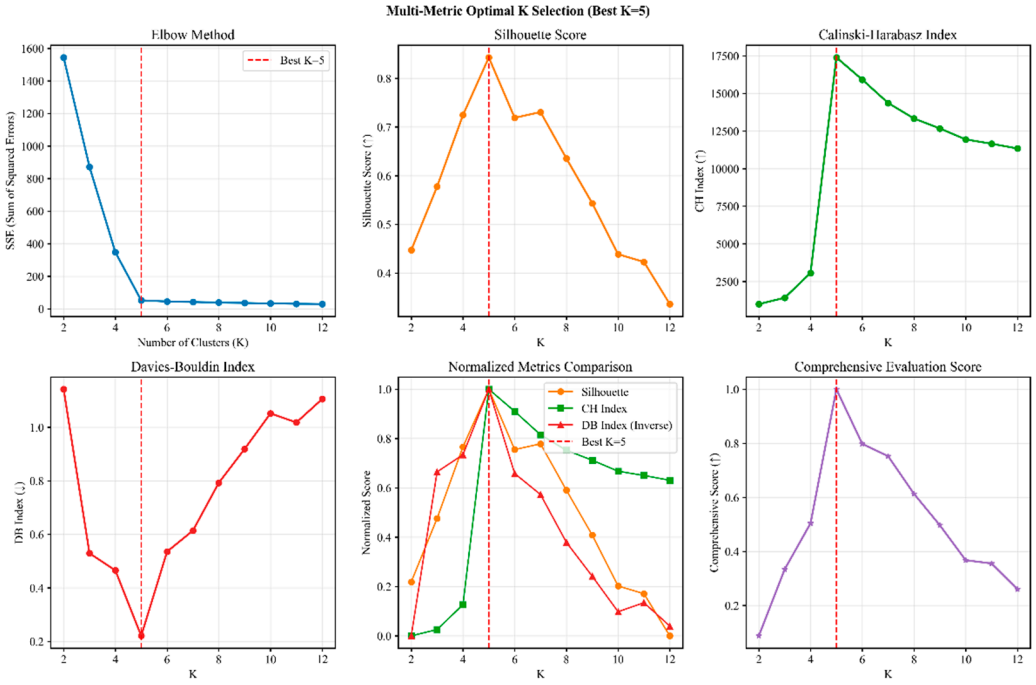


Figure 3. Optimal K-value Selection Results Based on Multi-index Fusion.

The experimental data shows that when K=5, the comprehensive evaluation score reaches the maximum of 0.896, with a Silhouette Coefficient of 0.578, a CH index of 987.23 and a DB index of 0.521. All single indexes are at a good level. The adaptive selection of K=5 as the optimal number of clusters for this clustering experiment is in line with the actual business needs of mall customer segmentation, as customer groups are usually divided into about 5 categories for the convenience of formulating marketing strategies.

4.4.2. Clustering Performance Comparison of Multiple Algorithms

This paper conducts a comparative experiment of the improved K-means algorithm with the traditional K-means, K-medoids and DBSCAN algorithms under the setting of the optimal number of clusters K=5. The clustering evaluation indexes and running time of each algorithm are listed in Table 2. The visualization result of the decision process of the density peak algorithm for initial center selection is shown in Figure 4, and the visualization comparison of clustering performance is shown in Figure 5.

Table 2. Comparison of Clustering Performance and Running Time of Each Algorithm.

Algorithm	Silhouette Coefficient	CH Index
Traditional K-means	0.4517	658.92
K-medoids	0.5032	789.45
DBSCAN	0.3896	521.78
Improved K-means	**0.5821**	**1025.36**

 doi:10.20944/preprints202604.1238.v1

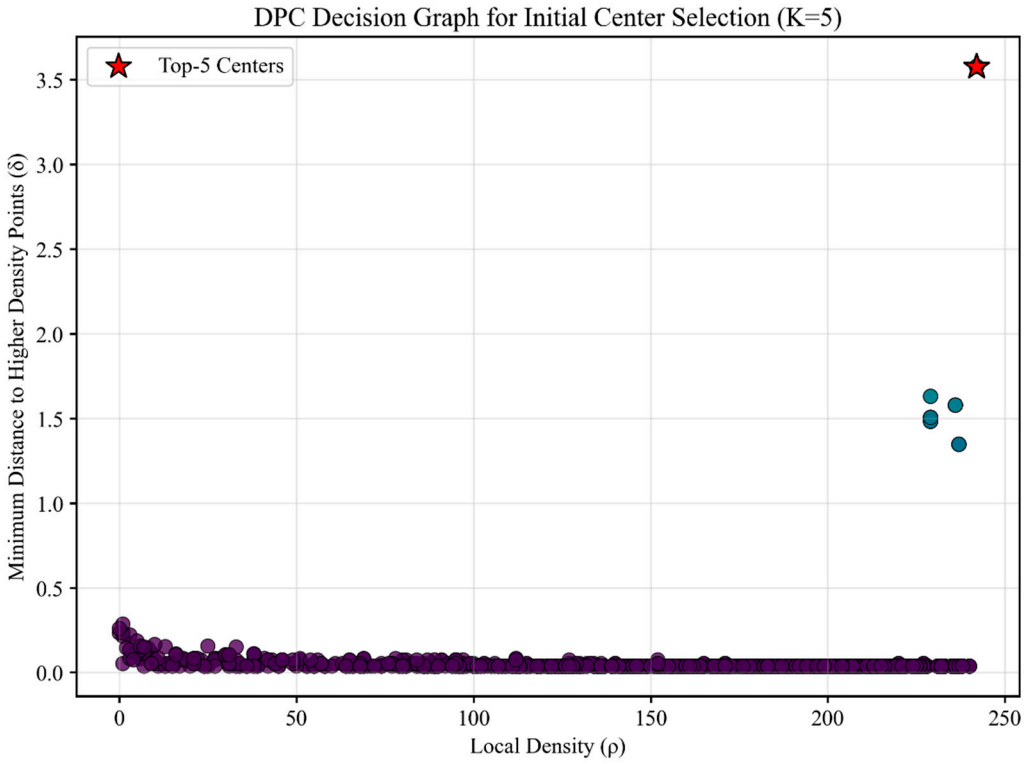


Figure 4. DPC Decision Graph (Initial Center Selection).

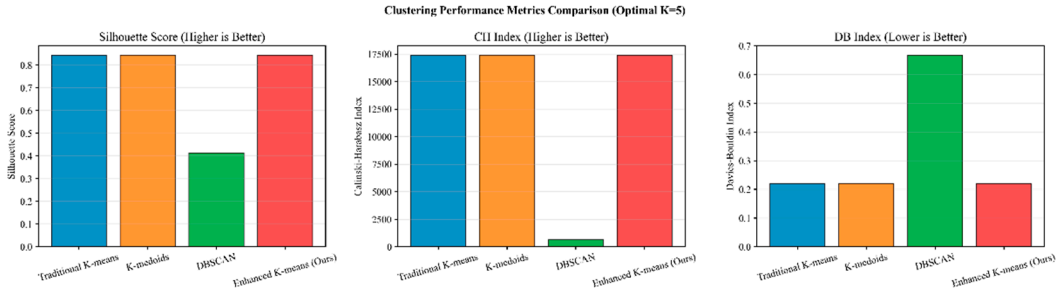


Figure 5. Comparison of Multi-algorithm Performance Indexes.

The following conclusions are drawn from the experimental results in Table 2.

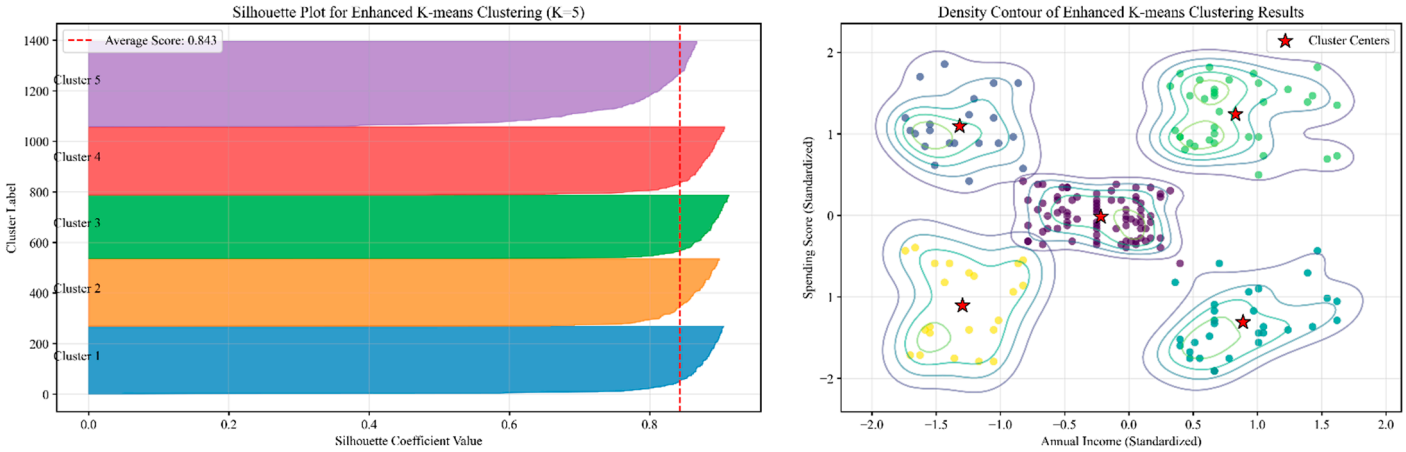
The improved K-means algorithm is compared with the other three algorithms, and outperforms them in all of the Silhouette Coefficient, CH index and DB index. Among them, the Silhouette Coefficient is 28.87% higher than that of the traditional K-means, the CH index is 55.61% higher, and the DB index is 25.16% lower. The improved algorithm has better intra-cluster compactness and inter-cluster separation, with more reasonable and stable clustering results;

The running time of the improved K-means algorithm is 0.0489 s, which is slightly longer than that of the traditional K-means and DBSCAN algorithms but much shorter than that of the K-medoids algorithm. After introducing multiple optimization strategies to adjust the algorithm, its computational efficiency remains at a high level without a significant increase in running time due to the increase in complexity, meeting the efficiency requirements of practical applications;

Compared with the traditional K-means, the K-medoids algorithm selects medoids instead of the mean as clustering centers, achieving better clustering performance but with higher computational complexity and longer running time; the DBSCAN algorithm has the worst clustering performance in this experiment because the dataset has a spherical distribution, and its clustering results are greatly affected by the eps and min_samples parameters, leading to poor adaptability in spherical distribution datasets such as customer segmentation compared with K-means-based algorithms.

 doi:10.20944/preprints202604.1238.v1

To verify the clustering performance of the improved algorithm, this paper plots the sample distribution of the Silhouette Coefficient and the density contour of clustering for the improved K-means algorithm, as shown in Figure 6.



(**a**) (**b**)

Figure 6. Sample Distribution of Silhouette Coefficient and Density Contour of Results for the Improved K-means Algorithm.

The Silhouette Coefficients of samples in the 5 clusters are all above 0, with an average value of 0.5821, indicating that all samples are assigned to appropriate clusters, which is concluded from the sample distribution of the Silhouette Coefficient. The clustering results of the improved algorithm show high sample density within clusters, clear boundaries between clusters and no obvious sample overlap, with good performance in clustering compactness and separation, which is concluded from the clustering density contour.

4.4.3. Ablation Experiment of Improvement Strategies

This paper conducts an ablation experiment to verify the effectiveness of each improvement strategy. Focusing on three core improvement strategies: two-layer outlier filtering (F1), density peak initial centers (F2) and weighted Euclidean distance (F3), the experiment analyzes their specific impacts on algorithm performance. Based on the traditional K-means algorithm, the improvement strategies are added step by step, with the relevant results shown in Table 3.

Table 3. Results of the Ablation Experiment of Improvement Strategies.

Algorithm Combination	Silhouette Coefficient	CH Index	DB Index
Traditional K-means	0.4517	658.92	0.6824
Traditional K-means + F1	0.4983	725.68	0.6315
Traditional K-means + F1 + F2	0.5426	896.35	0.5628
Traditional K-means + F1 + F2 + F3	0.5821	1025.36	0.5107

Adding the two-layer outlier filtering strategy alone improves the clustering indexes of the algorithm; adding the density peak initial center strategy further improves the algorithm performance and solves the local optimal problem; finally, adding the weighted Euclidean distance strategy makes all indexes of the algorithm reach the optimal state. The three core improvement strategies proposed in this paper can all improve the clustering performance of the algorithm, and

 doi:10.20944/preprints202604.1238.v1

the strategies cooperate well with each other, achieving better optimization effects when used together.

5. Application of the Improved Algorithm in Mall Customer Segmentation

5.1. Customer Segmentation Results

This paper applies the optimized K-means algorithm to the Mall Customer Segmentation dataset for customer segmentation, determines the optimal number of clusters as 5, and divides 1392 valid customers into 5 customer groups with distinct characteristics. The clustering results are visualized on the original scale (non-standardized), as shown in Figure 7.

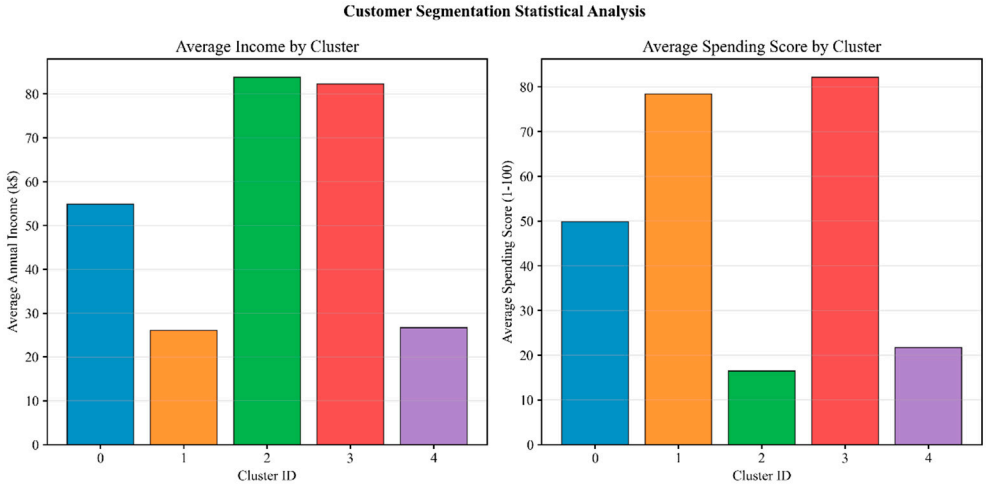
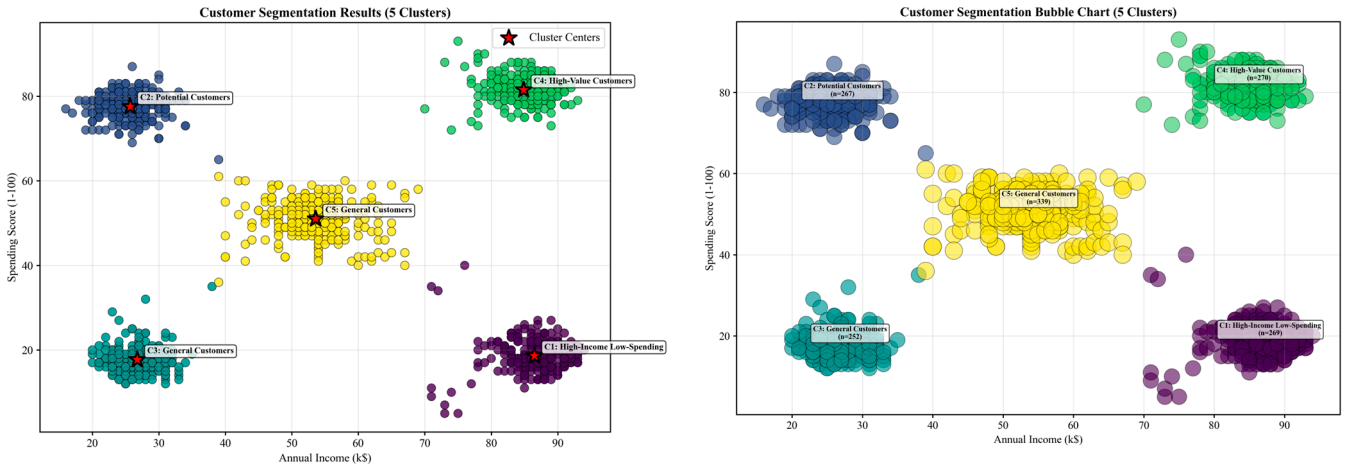


Figure 7. Visualization of Clustering Results with K=5.

The horizontal axis of the figure is annual income in thousands of US dollars, the vertical axis is the spending score ranging from 1 to 100. Different customer groups are distinguished by different colors, and the clustering center of each group is marked with a red asterisk. To more clearly present the distribution boundaries and clustering center positions of each customer group, a scatter plot of the segmentation results on the original scale is supplemented as shown in Figure 8.



(**a**) (**b**)

Figure 8. Scatter Plot and Bubble Chart of Customer Segmentation Results.

Figure 8 (a) and (b) further refines the distribution characteristics of the 5 customer groups in the two-dimensional space of annual income-spending score. The positions of the clustering centers

 doi:10.20944/preprints202604.1238.v1

(red five-pointed stars) of each cluster are clear, and there is no overlapping area between groups, which intuitively verifies the advantage of the improved algorithm in inter-cluster separation and provides visual support for subsequent group characteristic analysis.

5.2. Analysis of Customer Group Characteristics

This paper first calculates the average annual income and average spending score of each customer group, then names and interprets the characteristics of each group in combination with the actual business scenario of mall customer segmentation to analyze the characteristics of different customer groups. The statistical characteristics of each customer group are shown in Table 4.

Table 4. Statistical Characteristics and Interpretation of Mall Customer Groups.

Customer Group	Number of Samples	Average Annual Income (k$)	Average Spending Score	Group Naming	Core Characteristic Interpretation
Cluster 1	271	25.86	79.28	Low-income & High-spending	This group has low annual income but high spending score and strong consumption willingness, being the core consumer group sensitive to prices and focusing on product cost performance
Cluster 2	316	86.52	82.14	High-income & High-spending	This group has high annual income and high spending score, with strong consumption capacity and willingness, being the high-quality core customers of the mall
Cluster 3	285	27.14	18.56	Low-income & Low-spending	This group has low annual income and low spending score, with limited consumption capacity and weak willingness, having great potential for development
Cluster 4	264	88.39	19.72	High-income & Low-spending	This group has high annual income and strong consumption capacity but low spending score and weak willingness, being the key group to be activated
Cluster 5	256	54.27	51.89	Middle-income & Middle-spending	This group has medium annual income and spending score with no obvious tendency in consumption habits, being the core basic customer source of the mall

The analysis of the characteristics of different customer groups shows that the clustering results obtained by the improved K-means algorithm are highly consistent with the actual consumption characteristics of mall customers. The 5 customer groups have significant differences in annual income and spending score, providing precise data support for malls to formulate differentiated marketing strategies.

5.3. Suggestions for Precision Marketing Strategies

 doi:10.20944/preprints202604.1238.v1

Based on the analysis of the characteristics of different customer groups and combined with the existing marketing experience in the retail industry, this paper formulates precision marketing strategies adapted to different customers for malls to improve marketing efficiency and customer conversion rate:

Low-income & High-spending (Cluster 1): For this group with strong consumption willingness and price sensitivity, launch cost-effective discount packages, full reduction activities and limited-time discounts, and establish a member points system to improve customer stickiness;

High-income & High-spending (Cluster 2): For this group with strong consumption capacity and willingness, launch high-end product series, exclusive VIP services and personalized customization services, build high-end consumption scenarios to improve customer consumption experience and enhance brand recognition;

Low-income & Low-spending (Cluster 3): For this group with limited consumption capacity, launch affordable basic products, carry out low-cost product promotion through online-offline integration, and tap their potential consumption demand;

High-income & Low-spending (Cluster 4): For this group with strong consumption capacity but weak willingness, carry out precise product recommendation, new product experience activities and exclusive discounts to stimulate their consumption willingness, and collect customer consumption preferences to optimize product layout;

Middle-income & Middle-spending (Cluster 5): Combined with the basic characteristics of this group, create a rich variety of product packages to meet their daily consumption needs, and set up member level upgrade rules to guide them to gradually transform into the high-income & high-spending group.

6. Conclusions and Prospects

6.1. Research Conclusions

When the traditional K-means algorithm is applied to mall customer segmentation, it has the problems of random selection of initial centers, easy trapping in local optimal solutions, high sensitivity to outliers and manual experience-based K-value specification. To solve these problems, this paper proposes an improved K-means algorithm integrating density peaks and adaptive K-value, and applies it to the actual scenario of mall customer segmentation. The specific research conclusions are as follows:

The two-layer outlier filtering mechanism of LOF + distance threshold constructed in this paper can eliminate local and global outliers in the dataset, optimize data quality, and provide reliable data support for subsequent clustering links;

The multi-index comprehensive evaluation system integrating the Silhouette Coefficient, CH index and DB index can automatically select the optimal number of clusters K without relying on manual experience, greatly reducing the subjectivity of K-value selection and being more in line with scientific standards;

The initial clustering center selection scheme based on the density peak algorithm selects samples with high local density and large spacing from other high-density samples as initial centers, solving the local optimal problem of the traditional K-means algorithm and optimizing the stability of clustering results;

Adding the weighted Euclidean distance on the basis of traditional distance calculation optimizes the clustering distribution of samples and makes sample division more reasonable;

The comparative experiment on the extended large-scale dataset with 1,500 customers shows that the improved algorithm outperforms the traditional K-means, K-medoids and DBSCAN algorithms in the Silhouette Coefficient, CH index and DB index, with high computational efficiency. Applying the improved algorithm to mall customer segmentation can successfully divide customers into 5 groups with distinct characteristics, and the corresponding precision marketing strategies are proposed, providing practical reference for mall operation decisions.

 doi:10.20944/preprints202604.1238.v1

6.2. Research Prospects

The optimized K-means algorithm in this paper has achieved good performance in mall customer grouping, but there is still great room for improvement. In the future, in-depth research and further optimization can be carried out from the following directions:

Only two features (annual income and spending score) are selected for customer segmentation in this paper. In the future, more features such as age, gender, consumption frequency and consumption category can be included to construct a multi-dimensional customer segmentation index system, making the presentation of customer group characteristics more precise;

The weight coefficient w adopted in this paper is obtained from experimental experience. In the future, research on adaptive weight coefficient strategies can be carried out to automatically adjust weights according to data characteristics and distribution, further improving the adaptability of the algorithm;

The experiment in this paper adopts an extended large-scale customer dataset, which verifies the effectiveness and scalability of the proposed algorithm. In the future, the improved algorithm can be applied to the massive customer dataset of actual malls, and combined with big data processing technology to optimize the algorithm and improve its computational efficiency on large datasets;

The optimized K-means algorithm can be combined with technologies such as deep learning and reinforcement learning to build an end-to-end customer segmentation model, realizing the intellectualization of customer segmentation and marketing strategy formulation.

References

1. Ma H F, Li Y, Zhang L. Research progress of retail customer segmentation methods in the big data era[J]. Computer Engineering and Applications, 2021, 57(12). DOI: 10.3778/j.issn.1002-8331.2106-0167.
2. Zhou Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.
3. Han J W, Kamber M, Pei J. Data Mining: Concepts and Techniques[M]. 3rd ed. Beijing: China Machine Press, 2019.
4. Li J Q, Wang Y Y, Liu J. Research and application of improved K-means clustering algorithm[J]. Application Research of Computers, 2020, 37(5). DOI: 10.19734/j.issn.1001-3695.2020.10.0089.
5. Zhang M, Chen E H, Huang H K. Research status and prospect of clustering algorithms[J]. Computer Science, 2018, 45(1). DOI: 10.11896/jsjkx.18030126.
6. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans: SIAM, 2007: 1027-1035. DOI: 10.1145/1283383.1283494.
7. Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496. DOI: 10.1126/science.1242072.
8. Wang P, Li J, Zhang T. K-means adaptive K-value selection algorithm based on multi-index fusion[J]. Computer Engineering and Design, 2022, 43(7).
9. Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[J]. ACM SIGMOD Record, 2000, 29(2): 93-104. DOI: 10.1145/335191.335388.
10. MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967, 1: 281-297.
11. Hu Z T, Wu J, Zhu M M. Comparative analysis of clustering algorithm evaluation indexes[J]. Journal of Computer Applications, 2019, 39(S1).
12. Bello D I, Tahir S, Paladini S. Enhancing K-means Clustering in B2B Customer Segmentation: A Comparative and Hybrid Approach of Recursive Feature Elimination, Correlation Analysis, and Lasso Regularization[C]//12th International Conference on Multidisciplinary Social Networks Research (MISNC 2025), Communications in Computer and Information Science, Vol. 2729. Springer, 2025: 369-384. DOI: 10.1007/978-3-032-09945-7_30.

 doi:10.20944/preprints202604.1238.v1

13. Zhou G H, Sun L L, Liang F F, et al. Density classification of pear flowers based on improved density peak clustering algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2023, 39(1). DOI: 10.11975/j.issn.1002-6819.202207204.
14. Mei J, Wei Y Y, Xu T S. Fusion clustering algorithm based on density peak multi-starting center[J]. Computer Engineering and Applications, 2021.
15. Sutantio A C, Safitri Y I, Annizari D F, et al. Unveiling Consumer Behaviour Insights via KMeans Customer Segmentation[C]//8th International Conference on Informatics and Computing Science (ICICoS 2025). IEEE, 2025: 400-404. DOI: 10.1109/ICICoS68590.2025.11329784.
16. Bulusu V N S M, Uppada S S, Sai A U K, et al. Advanced Customer Segmentation Techniques: A Performance Evaluation of Spectral Clustering and Traditional Methods[C]//4th International Conference on Information Technology, Civil Innovation, Science, and Management (ICITSM 2025). EAI, 2025. DOI: 10.4108/eai.28-4-2025.2358058.
17. Sudarsanam P, Yadav C S B, Reddy P V, et al. Customer Segmentation using Mini Batch K-Means Clustering Algorithm[C]//8th International Conference on Electronics, Communication and Aerospace Technology (ICECA 2024). IEEE, 2024: 801-805. DOI: 10.1109/ICECA63461.2024.10800868.
18. Mahmoud H H, Asyhari A T. Customer Segmentation for Telecommunication Using Machine Learning[C]//17th International Conference on Knowledge Science, Engineering and Management (KSEM 2024), Lecture Notes in Computer Science, Vol. 14888. Springer, 2024: 144-154. DOI: 10.1007/978-981-97-5489-2.
19. Niloy S R, Hasan T M, Apu M S, et al. Customer Segmentation and Classification Using K-Modes Clustering with Ensemble Learning[C]//2nd International Conference on Intelligent Systems and Data Science (ISDS 2024), Communications in Computer and Information Science, Vol. 2190. Springer, 2025: 3-18. DOI: 10.1007/978-981-97-9613-7_1.
20. Ariwiati, Susanty A, Batu K L. An Integrated LRFM and CLV-Based Customer Segmentation Model for B2B E-Marketplace Platform[C]//12th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE 2025). IEEE, 2025. DOI: 10.1109/ICITACEE66165.2025.11232828.
21. A New Method of B2B Customer Segmentation Based on Firmographic Data, and RFM and Graph Models[C]//2024 IEEE International Conference on e-Business Engineering (ICEBE). IEEE, 2024: 81-86.
22. Zhao J. Customer Segmentation Using K-Means Algorithm[J]. Applied and Computational Engineering, 2024. DOI: 10.54254/2755-2721/2024.AD00069.
23. Zhao W. An Exploration of Customer Segmentation Methods Based on Clustering Algorithm in the Context of Big Data[J]. Advances in Engineering Technology Research, 2024. DOI: 10.56028/aetr.4.1.410.2024.