PLOS ONE

RESEARCH ARTICLE

Customer segmentation in the digital marketing using a *Q*-learning based differential evolution algorithm integrated with *K*-means clustering

Guanqun Wang[1,2]*

1 College of Accounting, Ningbo University of Finance & Economics, Ningbo, China, **2** Zhejiang Marine Development Think Tank Alliance, Ningbo, Zhejiang, China

* wangyijinqunqun@163.com

Abstract

Effective and well-structured customer segmentation enables organizations to accurately identify and comprehend the distinct characteristics and needs of various customer groups, thereby facilitating the development of more targeted marketing strategies. Contemporary artificial intelligence technologies have emerged as the predominant tools for customer segmentation, owing to their robust capabilities in analyzing complex datasets and extracting profound customer insights. This paper proposes a customer segmentation framework within the realm of digital marketing, which integrates a reinforcement learning-based differential evolution algorithm with *K*-means clustering using dimensionality reduction techniques to address challenges in the customer segmentation process. Initially, a correlation matrix is used to identify redundant noise and multicollinear features within customer feature groups, and Principal Component Analysis is applied for denoising and dimensionality reduction to enhance the ability of the model to identify potential features. Subsequently, a parameter adaptive adjustment method based on *Q*-learning is proposed, which significantly augments the clustering performance of *K*-means. Ultimately, the effectiveness of the proposed method is validated using a Kaggle dataset, and the elbow method is employed to ascertain the optimal number of clusters. Based on the cluster category centers, the typical characteristics of different customer types are analyzed. Furthermore, four widely recognized machine learning methods are employed to classify the clustering results, achieving over 95% classification accuracy on the test set. The experimental results demonstrate that the proposed model exhibits a high degree of customer characteristic identification and segmentation, which not only enhances marketing efficiency and customer satisfaction but also fosters corporate profit growth through the strategic formulation of various marketing initiatives.

Citation: Wang G (2025) Customer segmentation in the digital marketing using a *Q*-learning based differential evolution algorithm integrated with *K*-means clustering. PLoS ONE 20(2): e0318519. https://doi.org/10.1371/journal.pone.0318519

Editor: Zhengmao Li, Aalto University, FINLAND

Received: November 19, 2024

Accepted: January 16, 2025

Published: February 7, 2025

Data Availability Statement: The dataset is publicly available from the Zenodo database (DOI: 10.5281/zenodo.14614252).

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

With the rapid advancement of network information technology, the digital economy has progressively established a significant presence within the traditional economic landscape [1, 2]. This digital economy embodies a new economic paradigm primarily driven by digital knowledge and information [3, 4]. Digital marketing constitutes a fundamental and integral component of the digital economy. By leveraging internet-based operational models, enterprises can effectively reach their target markets and optimize customer experiences, thereby facilitating sales and revenue growth [5–7]. In this context, marketing capabilities within the digital economy have emerged as a crucial factor in ensuring and enhancing corporate revenue, serving as a vital strategy for companies to maintain their competitive advantages in the market [8, 9]. Understanding customer characteristics and segmenting these characteristics based on similarities is a key approach for enterprises to achieve efficient digital marketing. Within the digital economy, customer segmentation refers to the process of categorizing customer attributes —such as demographic, behavioral, and geographic—obtained through online channels into distinct sub-clusters [10]. This segmentation enables companies to assess the loyalty of various customer groups and develop tailored marketing strategies aimed at maximizing business revenue [11]. By comprehending the unique needs and preferences of each segment, businesses can optimize their marketing efforts and enhance customer engagement, ultimately driving greater profitability [12].

Customer segmentation presents a significant challenge for businesses [13–15]. Firstly, customer data is often sourced from multiple channels, which complicates the integration of this data into a cohesive and comprehensive customer profile [16]. Furthermore, the data collected is frequently characterized by inaccuracies and gaps, which can severely undermine the effectiveness of the segmentation process [17]. Secondly, customer preferences and behaviors are not static; they evolve in response to dynamic market conditions [18]. Fluctuations in the external market environment and emerging trends can significantly influence customer needs and behavioral characteristics [19]. Moreover, effective customer segmentation typically necessitates the use of sophisticated data analysis tools and methodologies [20]. While advancements in artificial intelligence, particularly through machine learning techniques, have enhanced the analytical capabilities of organizations to some extent, existing methods often exhibit various limitations in practical application. This underscores the need for a more comprehensive analytical approach to thoroughly understand customer behavioral characteristics, thereby facilitating complete and effective customer segmentation [21–24].

To address the challenges encountered in the customer segmentation process, this paper proposes a customer segmentation framework that integrates a reinforcement learning-based differential evolution algorithm with *K*-means clustering in a dimensionality reduction context. First, a correlation matrix is employed to analyze redundant noise and multicollinear features within the dataset. Subsequently, Principal Component Analysis (PCA) is applied to eliminate redundant noise from the feature data, resulting in a set of low-dimensional, uncorrelated principal component features that enable the classification model to effectively identify potential features within the data. Second, to overcome the limitations of the traditional differential evolution algorithm, which utilizes a fixed scaling factor that does not adapt dynamically to varying problems, a parameter adaptive adjustment method based on *Q*-learning is proposed. This method is effectively integrated with the *K*-means clustering algorithm, significantly enhancing the clustering performance of *K*-means. To validate the effectiveness of the proposed method, we utilize the Kaggle dataset for empirical testing, employing the elbow method to determine the optimal number of clusters for the algorithm. The typical features of different customer types are then delineated based on the cluster centers. Finally, we apply

four popular machine learning methods to classify the clustering results, thereby assessing the practicality of the clustering outcomes. The experimental results demonstrate that all four methods achieve over 95% classification accuracy on the test set. We recommend the use of Artificial Neural Networks (ANN) and Kernel Support Vector Machines (SVM) for classifying new customers.

In summary, the main contributions of this study are as follows:

- A novel customer segmentation framework is introduced that enhances segmentation quality by reducing feature dimensions and effectively integrating a reinforcement learning-based differential evolution algorithm with the *K*-means algorithm.
- A *Q*-learning-based adaptive dynamic adjustment method for the differential scaling factor is presented, allowing for improved adaptability to diverse problem environments and yielding superior search results.
- The elbow method is employed to determine the optimal number of clusters for customer segmentation. Additionally, four widely used machine learning techniques are utilized to assess the classification accuracy of the clustering results, underscoring the practical applicability of the proposed customer segmentation approach.

The remainder of the work is organized as follows. **Section 2** gives the state-of-the-art work on customer segmentation; **Section 3** proposes customer feature dimensionality reduction and combined clustering methods; The experimental results and analysis are presented in **Section 4**; **Section 5** summarizes the findings and outlines potential future work.

2. Related work

Customer segmentation plays a crucial role in marketing and data analysis [25–27]. The RFM (Recency, Frequency, Monetary) segmentation method is a widely used technique in marketing and customer relationship management [28]. The RFM model divides customers into different groups by analyzing their purchasing behavior in order to formulate more targeted marketing strategies [29, 30]. Christy et al. [31] employed RFM analysis to conduct customer segmentation and subsequently enhanced this approach by making minor modifications to the existing *K*-means clustering algorithm. Their findings indicated that the proposed algorithm outperformed other methods in terms of effectiveness. However, it is noteworthy that the study did not take into account the performance of customers within each segment. Rungruang et al. [32] proposed an innovative clustering algorithm that integrates both implicit and explicit knowledge by combining RFM analysis with formal concept analysis (FCA). Experimental results demonstrate that this proposed method can effectively generate practical marketing strategies for real-world businesses. However, the extraction and interpretation of implicit knowledge within the model may be influenced by data quality and completeness, which could limit the generalizability and applicability of the approach. The effectiveness of the RFM model is highly dependent on the quality and integrity of the data. If the data is inaccurate or incomplete, it may lead to incorrect customer classification and inappropriate marketing strategies [33, 34].

With the rapid development of computer science, AI technologies, particularly those represented by machine learning and deep learning, are increasingly being applied to the field of customer segmentation [35–37]. Yadegaridehkordi et al. [38] conducted a comprehensive study in which they first employed the *K*-means clustering algorithm to segment online reviews from travelers. Following this segmentation, they applied the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method [39] to prioritize green hotel

attributes based on their importance within each identified segment. To further enhance their analysis, they utilized the Classification and Regression Trees (CART) method [40] to investigate the relationship between environmentally friendly hotel characteristics and traveler satisfaction. This integrated approach provides managers with valuable insights for developing and implementing environmentally sustainable practices in the hospitality industry. Nilashi et al. [41] developed a new method for customer segmentation and preference prediction using text mining and predictive learning techniques and adopted clustering technology for customer segmentation. The effectiveness of the proposed method was evaluated using a restaurant dataset. The experimental results showed that the proposed method can better reveal customer satisfaction and make high-precision predictions of their preferences through their purchasing behavior. Wang et al. [10] employed the RFM model to preprocess the data and utilized an improved social spider optimization (MSSO) technique to select relevant customer features. They then applied a self-organizing neural network to identify six key features for clustering, which facilitated the formation of distinct customer segments. Finally, the Deep Neural Network (DNN) method [42] was used for customer segmentation. The experimental results indicate that the segmentation outcomes of the proposed approach surpass those of traditional methods, offering significant reference value for digital marketing strategies. Alkhayrat et al. [43] proposed a customer segmentation method that integrates Principal Component Analysis (PCA) for dimensionality reduction with an autoencoder neural network. The method reduces the feature space of original data before applying the *K*-means clustering algorithm for evaluation. Simulation experiments conducted on real telecommunications datasets demonstrate effective utilization of clustering in both reduced and latent spaces. This a dual strategy enables a deeper understanding of customer preferences and needs, ultimately leading to higher quality clustering results.

The analysis indicates that integrating traditional RFM models with AI-based technologies can effectively address the dynamic requirements of customer segmentation, reducing the risks associated with incorrect segmentation and decision-making that may arise from reliance on conventional methods. Despite the availability of numerous technical approaches for customer segmentation and the achievements realized through these methods, the increasing complexity of data sources and formats poses significant challenges, as existing technologies often encounter difficulties in processing and analyzing this diverse data effectively. Combining heuristic algorithms with AI classification models has proven to be a reasonable and effective approach [44–47]. Consequently, the adoption of customer segmentation models that utilize heuristic algorithms and artificial intelligence technologies is recommended, as these methodologies can provide deeper insights into customer behavior and preferences, thereby enabling organizations to make more informed decisions and adapt strategies to meet the evolving demands of the market. This approach not only enhances the accuracy of customer segmentation but also aligns with the necessity for businesses to remain agile in a rapidly changing environment.

3. Proposed methodology

3.1 Feature dimension reduction strategy based on PCA

Customer segmentation data often exhibits high correlations among features, along with the presence of irrelevant features, which significantly increases the difficulty of precise identification and classification by modeling techniques [48]. Principal Component Analysis (PCA) [49], illustrated in Fig 1, is a widely used unsupervised dimensionality reduction technique that employs orthogonal transformation to convert multiple linearly correlated variables into a smaller number of linearly independent principal components.

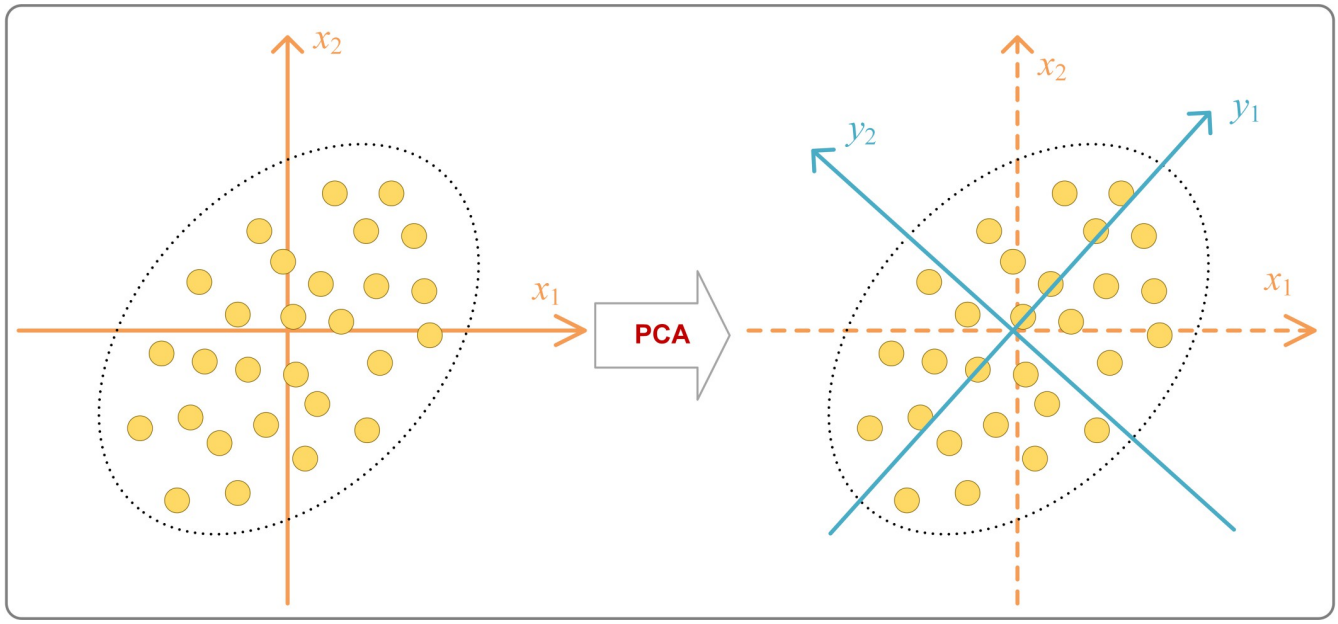

Fig 1. Schematic diagram of PCA principle.

https://doi.org/10.1371/journal.pone.0318519.g001

In the dimensionality reduction process of principal component analysis (PCA), one of the key steps is to select the first few principal components that can explain the largest possible variance in the data [50]. In practical implementation, PCA begins with the normalization of the data, ensuring that each feature contributes equally to the analysis. Following this, the covariance matrix of the dataset is computed. The next step involves performing an eigenvalue decomposition of the covariance matrix to extract the eigenvectors and eigenvalues. Here, the eigenvectors represent the basis of the new feature space, while the eigenvalues indicate the importance of each principal component in terms of the variance it explains. Finally, the number of principal components to retain is determined by examining the cumulative variance explained by the principal components. A common practice is to select enough components to reach a specified threshold of explained variance, ensuring that the selected components capture a significant portion of the data's variability. Through these steps, PCA effectively reduces the dimensionality of the data while preserving as much relevant information as possible, facilitating subsequent data analysis and modeling tasks.

3.2 The basic *K*-means clustering algorithm

The *K*-means clustering algorithm [51], illustrated in Fig 2, is an unsupervised hard clustering method that partitions a sample set. Given a dataset $\boldsymbol{X}_o = \{x_1, x_2, x_3, \ldots, x_N\}$ consisting of N samples, where each sample is characterized by d features, the distance between any two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be mathematically represented using a distance metric, defined as follows:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{l=1}^{d} (x_{il} - x_{kl})^2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \tag{1}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ represents the squared Euclidean distance; x_{il} represents the value of the i-th sample with respect to the l-th feature.

The *K*-means clustering method typically involves identifying K cluster centers C (where $C \in \{m_1, m_2, \ldots, m_k\}$) that partition the samples in the dataset $\boldsymbol{X}_o$ so that the sum of squared

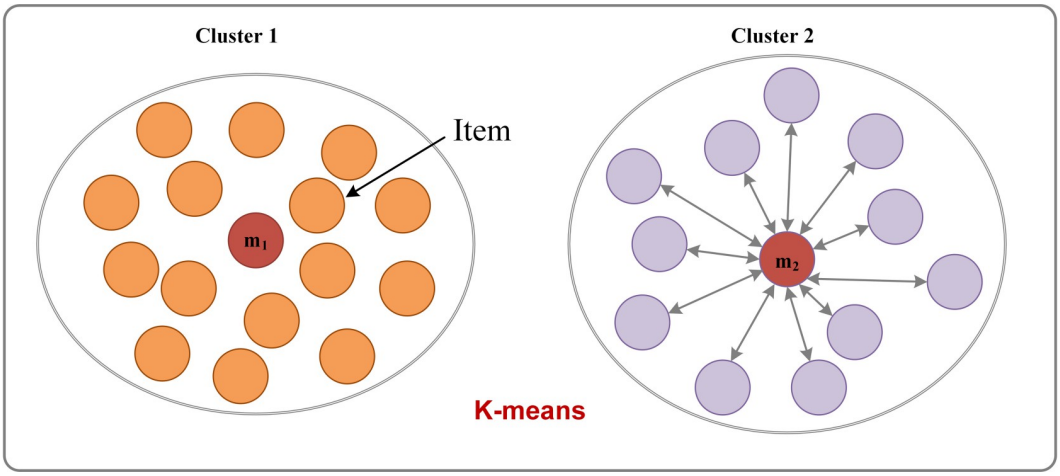

Fig 2. Schematic diagram of *K*-means algorithm.

https://doi.org/10.1371/journal.pone.0318519.g002

Euclidean distance (SSE) [52] is minimized:

$$\min SSE = \sum_{l=1}^{K} \sum_{C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{m}_l\|^2 \tag{2}$$

where $\boldsymbol{m}_l$ represents the center position vector of the l-th cluster, defined as follows:

$$\boldsymbol{m}_l = \frac{1}{n_l} \sum_{C(i)=l} \boldsymbol{x}_i, \quad l \in \{1, \cdots, k\} \tag{3}$$

where n_l represents the number of samples in the l-th cluster partition. In addition, the optimal solution for *K*-means clustering is an NP-hard problem and is typically addressed using an iterative method.

3.3 The differential evolution algorithm based *Q*-Learning (QLDE)

A. The differential evolution algorithm. The Differential Evolution (DE) algorithm [53], is a well-known heuristic optimization technique inspired by the principles of population evolution. It has been successfully applied across various fields, including engineering [54], machine learning [55], and operations research [56], yielding compelling results. The DE algorithm primarily consists of four key processes: initialization, mutation, crossover, and selection. While the mutation and crossover processes are illustrated in Fig 3. And four key processes are described as follows:

1. Initialization process
   The initial generation method of the population will affect the performance of the algorithm. To ensure the diversity of the algorithm in the early stage, Logistic mapping [57] is used to generate chaotic sequences and generate the initial population, Logistic mapping is showed as follows:

$$\phi_{j+1} = \mu \cdot \phi_j \cdot (1 - \phi_j) \tag{4}$$

   where ϕ_j a represents the chaotic random number generated in the j-th dimension of the problem, when $j = 0$, ϕ_j is a random number in [0, 1]; μ is the control parameter, typically chosen within the range [3.57, 4], then the generation method of the population can be

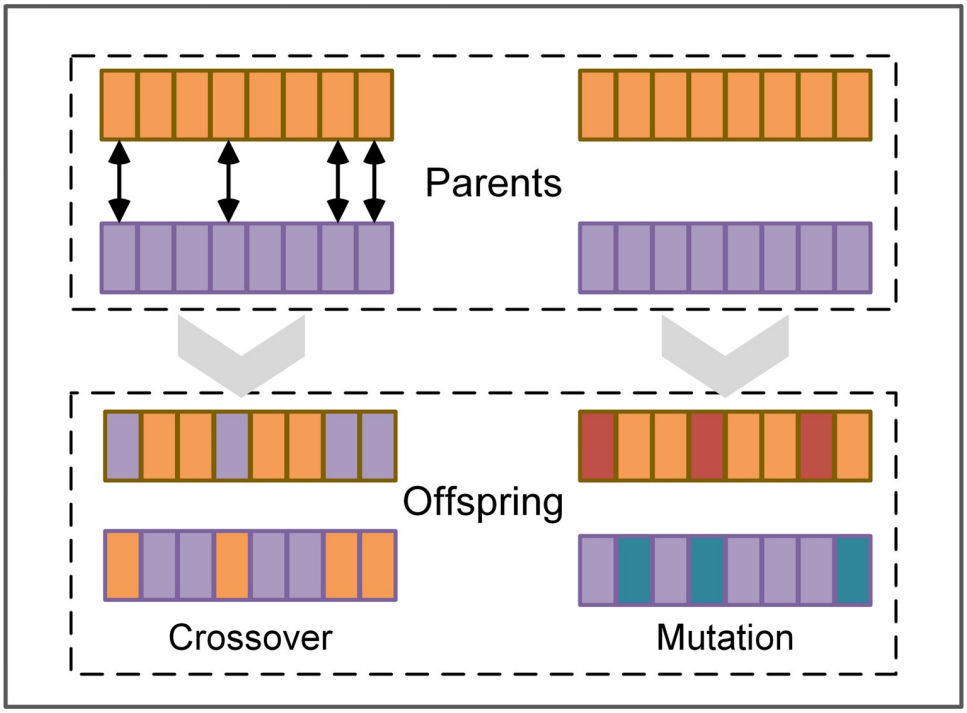

Fig 3. Crossover and mutation process of DE.

https://doi.org/10.1371/journal.pone.0318519.g003

expressed as:

$$z_{ij}^{k} = LB_j + \phi_j \cdot (UB_j - LB_j) \tag{5}$$

where z_{ij}^{k} represents the position of the j-th dimension of i-th individual at k-th iteration; LB_j and UB_j are represent the lower and upper boundary limits of the problem in the j-th dimension, respectively.

2. Cluster-guided mutation process

The mutation process generates a new parameter vector by adding a weighted difference vector, derived from any two individuals in the population, to a third individual. However, traditional mutation operations primarily ensure the randomness of the algorithm and do not provide a reliable guiding direction. To address this limitation, this paper introduces a clustering strategy into the population. Specifically, the top 50% of individuals with better fitness values in each generation are considered to have a better guiding direction, and then they are combined with differential vectors to form new mutated individuals. This approach facilitates the exploration of new solution spaces and enhances the ability of population to adapt to various characteristics of objective functions by incorporating randomness into the search process. The details are shown as follows:

$$\boldsymbol{v}_i^k = \boldsymbol{z}_{Lk}^k + F \cdot (\boldsymbol{z}_{i1}^k - \boldsymbol{z}_{i2}^k) \tag{6}$$

where $\boldsymbol{v}_i^k$ represents the mutation position vector corresponding to the i-th individual at the k-th iteration; F is the scaling factor, which controls the range of mutation; Lk denotes the position index randomly selected from the top 50% of the entire population with better performance at the k-th iteration. $i1$ and $i2$ are the index of a position randomly selected from the entire population at the k-th iteration.

3. Crossover and selection process

The crossover process primarily involves combining the mutation vector with the current individual to generate a trial vector. This process is characterized by the crossover factor C_r, which indicates the degree of information exchange. A larger C_r value suggests that the algorithm can explore a broader range of positions, while a smaller C_r value may lead to premature convergence, which is detrimental to global optimization. The selection process evaluates the fitness of the current individual and the trial vector. Following the principle of survival of the fittest, individuals with better fitness performance are carried forward into the next iteration of the cycle, as follows:

$$u_{ij}^{k} = \begin{cases} z_{ij}^{k}, if[r_2(0,1) \le C_r] or [j = j_{rand}] \\ v_{ij}^{k}, else \end{cases}, j = \{1, 2, \ldots, J\} \tag{7}$$

$$\boldsymbol{z}_{i}^{k+1} = \begin{cases} \boldsymbol{z}_{i}^{k}, if[f(\boldsymbol{z}_{i}^{k}) < f(\boldsymbol{u}_{i}^{k})] \\ \boldsymbol{u}_{i}^{k}, else \end{cases} \tag{8}$$

where u_{ij}^{k} represents the position of the j-th dimension of i-th trial individual at k-th iteration; $r_2(0,1)$ is the random number in the range [0, 1]; C_r is the crossover factor; j_{rand} represents the dimension index randomly selected from the total dimension J of the individual; $f(\boldsymbol{z}_i^k)$ and $f(\boldsymbol{u}_i^k)$ are the fitness value of the i-th position vector $\boldsymbol{z}_i^k$ and trial vector $\boldsymbol{u}_i^k$, respectively.

B. The *Q*-learning algorithm. *Q*-learning [58] is a popular value-based reinforcement learning method that has numerous applications in areas such as supply chain management [59], time series forecasting [60], control systems [61], and computer games [62]. The *Q*-learning algorithm enables an agent to interact dynamically with its environment, allowing it to make specific actions based on the current state. This approach constructs a *Q*-table to select actions that can yield the maximum reward and continuously explores through an iterative process to discover potential global reward paths. As shown in Fig 4, *Q*-learning, like other reinforcement learning algorithms, consists of four primary components: the agent, states, rewards, and actions. After each interaction with the environment, the agent evaluates whether to change its state and whether to receive rewards. Subsequently, it updates the *Q*-table to adapt to the demands of the environment. Specifically, the update process of the *Q*-table can

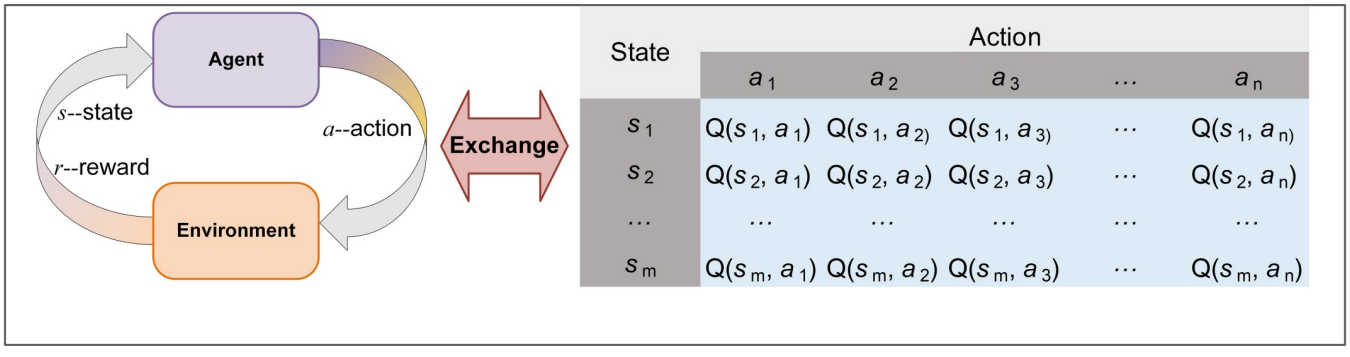

Fig 4. The *Q*-learning update process.

https://doi.org/10.1371/journal.pone.0318519.g004

be expressed by the following formula:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max\ Q(s_{t+1}, a') - Q(s_t, a_t)] \tag{9}$$

where α represents the learning rate, which determines the extent to which new knowledge covers old knowledge, $\alpha \in [0,1]$; γ represents the discount factor, which represents the importance of future rewards compared to immediate rewards, $\gamma \in [0,1]$; r_{t+1} represents the reward received at the next time step; s_{t+1} represents the state of the environment at the next time step; a' is the action that yields the highest Q value for a given state s_{t+1}.

C. The proposed QLDE algorithm. The exploration of parameter settings for Differential Evolution (DE) has been ongoing for many years; however, the relationship between these parameters and performance remains inadequately defined. The fixed nature of the scaling factor F in traditional DE limits its ability to adapt to challenges encountered during the search process, such as stagnation. To address this issue, a hybrid approach that combines Q-learning with the Differential Evolution algorithm is proposed, enabling the algorithm to autonomously adapt the scaling factor without the need for predefined parameter settings throughout the search process.

In addition, traditional Q-learning often faces the risk of becoming trapped in local optima during the search process. To address this challenge and effectively integrate with the differential evolution algorithm, a dynamic ε-greedy strategy [63] is introduced to improve the expected selection process in Q-learning. By incorporating randomness, this approach helps prevent the agent from converging to a local optimum during the learning phase, thereby increasing the likelihood of discovering the global optimum. The updated process for the Q-table can be expressed as follows:

$$\exp\ Q = \begin{cases} \max\ Q(s_{t+1}, a'),\ \ if\left(rand < 1 - \boldsymbol{\varepsilon} \cdot \dfrac{k}{k_{\max}}\right) \\ \text{rand\_choice}(),\ \ else \end{cases} \tag{10}$$

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \exp Q - Q(s_t, a_t)] \tag{11}$$

where $\boldsymbol{\varepsilon}$ is the greedy threshold, where a larger value indicates greater randomness. rand_-choice() refers to randomly selecting an action from a predefined set of actions.

In the proposed method of QLDE, the probability-based Softmax strategy [64] is employed for action selection to determine the most likely action that Q-learning would execute in a specific state, as follows:

$$\pi(s_i, a_j) = \text{Softmax}[Q(s_i, a_j)] = \frac{\exp[Q(s_i, a_j)]}{\sum_{j=1}^{D} \exp[Q(s_i, a_j)]} \tag{12}$$

where $\pi(s_i, a_j)$ represents the probability of taking action a_j in state s_i; D represents the total number of possible actions of the agent.

Finally, each individual is assigned a separate scaling factor, with three specified behaviors: λ = -0.01, λ = 0, and λ = 0.01, ensuring that the scaling factor can be adaptively adjusted. Furthermore, a reward of R = 1 and a state of S = 1 are granted to the agent only when the fitness of the trial individual exceeds that of the parent generation; otherwise, the reward is set to R = 0 and the state to S = 2. The detailed process of QLDE is shown in Fig 5. The adjustment method for the i-th individual through Q-learning can be expressed as follows:

$$F_i = F_i + \lambda_i \tag{13}$$

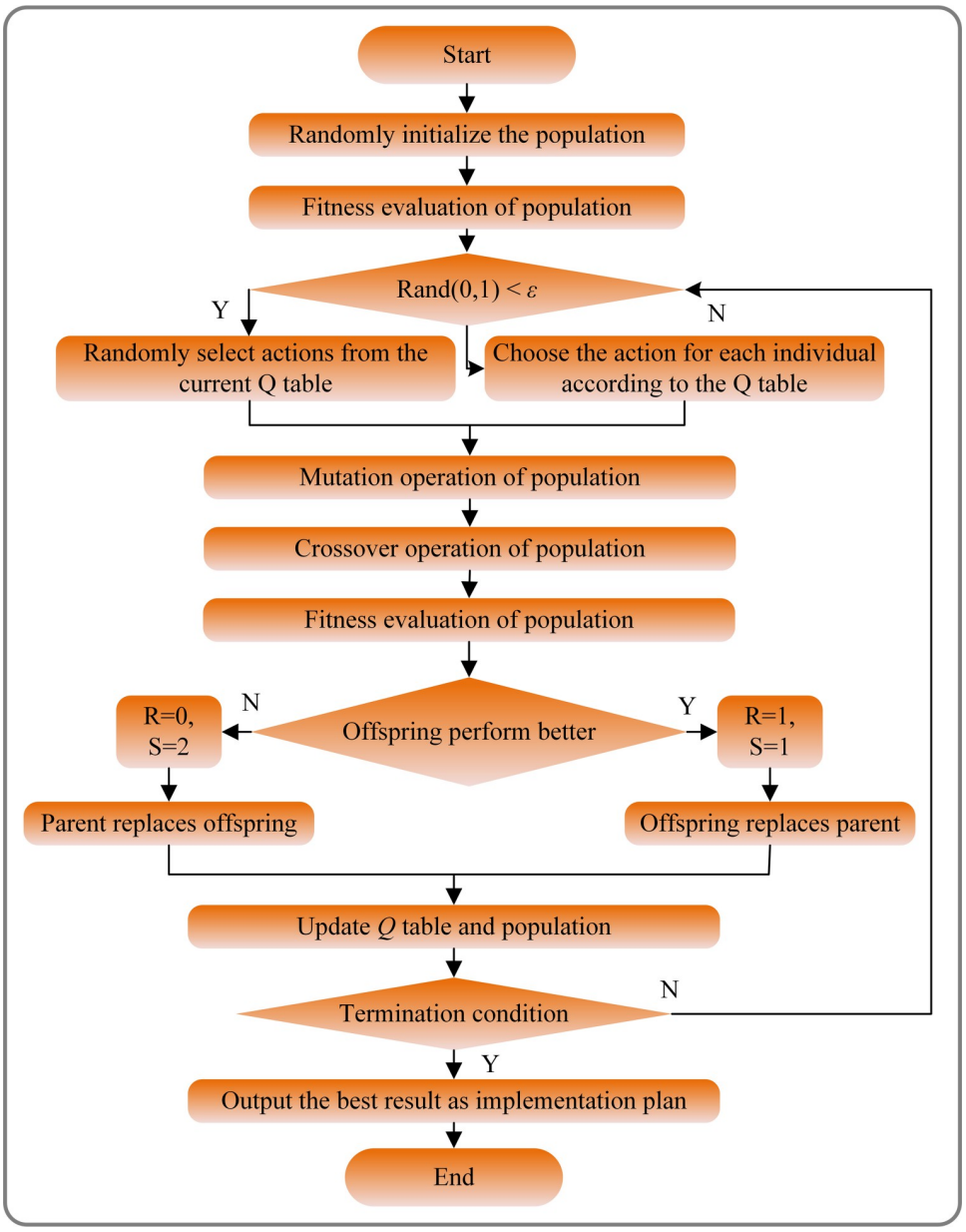

Fig 5. The flowchart of QLDE algorithm.

https://doi.org/10.1371/journal.pone.0318519.g005

where F_i and λ_i are the scaling factor and adjustment factor corresponding to the i-th individual.

3.4 Customer segmentation based on *K*-means-QLDE

Although the *K*-means algorithm is simple and efficient, its accuracy is highly dependent on the selection of the initial cluster centers. Different initializations can lead to varying clustering results, and poorly chosen initial centers may even result in significantly inferior outcomes.

To address the challenge of setting initial cluster centers in the traditional *K*-means algorithm, we propose a *K*-means-QLDE method based on a dimensionality reduction approach. This method first employs PCA to linearly transform high-dimensional data into a lower-

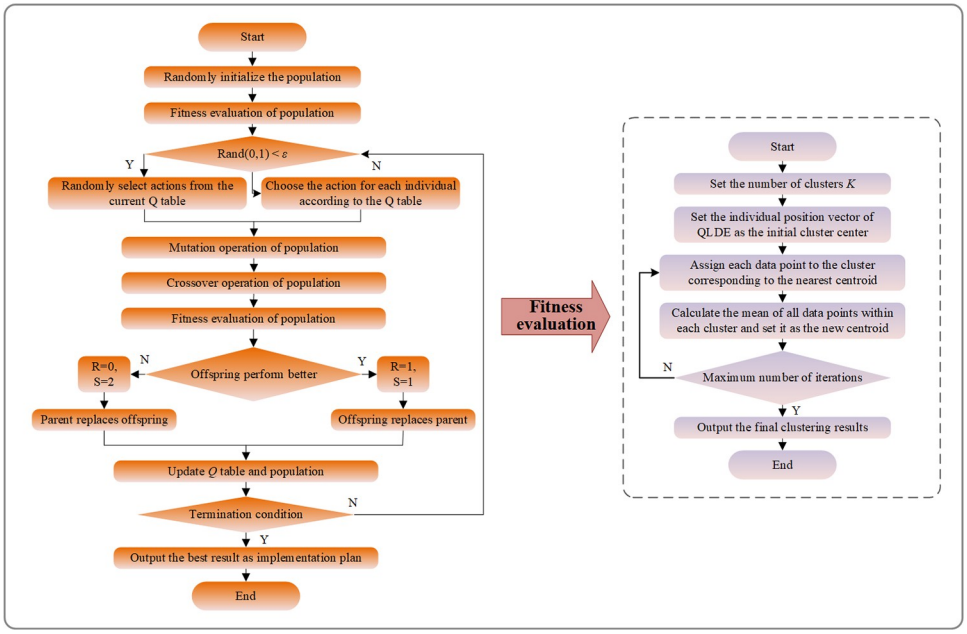

Fig 6. The implementation process of *K*-means-QLDE.

https://doi.org/10.1371/journal.pone.0318519.g006

dimensional space, utilizing cumulative variance to select the principal components, which represent the features of the lower-dimensional space. Then, we combine the QLDE algorithm with *K*-means for adaptive clustering. Specifically, the number of clusters and the corresponding low-dimensional features serve as inputs to the QLDE algorithm, optimizing with the SSE of *K*-means as the objective to obtain the final distance results. The implementation process of *K*-means-QLDE is illustrated in Fig 6.

4. Results and discussions

To validate the effectiveness of the method, we used the highly credible Kaggle customer segmentation dataset [65] to verify the proposed digital marketing customer segmentation method. The original data consists of eight features, namely "InvoiceNo", "StockCode", "Description", "Quantity", "InvoiceDate", "UnitPrice", "CustomerID" and "Country". To better capture the potential patterns and relationships within the data and enhance the predictive capability of our machine learning model, we employed the RFM (Recency, Frequency, Monetary) method to transform the original features into more meaningful information, as illustrated in Table 1.

4.1 Data preprocessing and feature dimension reduction

Data normalization is a crucial step in data preprocessing, aimed at bringing features of different dimensions onto the same scale. This process helps reduce the instability of numerical calculations and enhances model performance. In this paper, we employ Z-score normalization to process the transformed features. The formula for Z-score normalization [66] is expressed as follows:

$$\boldsymbol{X} = \frac{\boldsymbol{X}_o - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{14}$$

where $\boldsymbol{X}$ represents standardized dataset; $\boldsymbol{X}_o$ represents the original dataset, while $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ represent the mean and standard deviation vectors for each feature in $\boldsymbol{X}_o$, respectively.

Table 1. Description of the new features after transform.

No.	Feature	Description
1	Var 1	The number of days since last purchase of the customer
2	Var 2	Total number of transactions
3	Var 3	Total number of products by customer
4	Var 4	Total expenditure on purchased items
5	Var 5	Average transaction cost
6	Var 6	Number of product types purchased
7	Var 7	Average number of days to purchase
8	Var 8	Expected purchase days
9	Var 9	From the country of UK
10	Var 10	Transaction cancellation frequency
11	Var 11	Average monthly expenditure

https://doi.org/10.1371/journal.pone.0318519.t001

To assess whether the transformed features provide distinct characteristic information, Fig 7 presents the results of the correlation analysis among various feature variables. In the figure, a darker yellow block indicates a stronger positive correlation, while a darker blue block signifies a stronger negative correlation. From the analysis, it is evident that there are significant correlations among the variables. For instance, feature var 1 exhibits a strong positive correlation with feature variables 2, 3, 5, and 9.

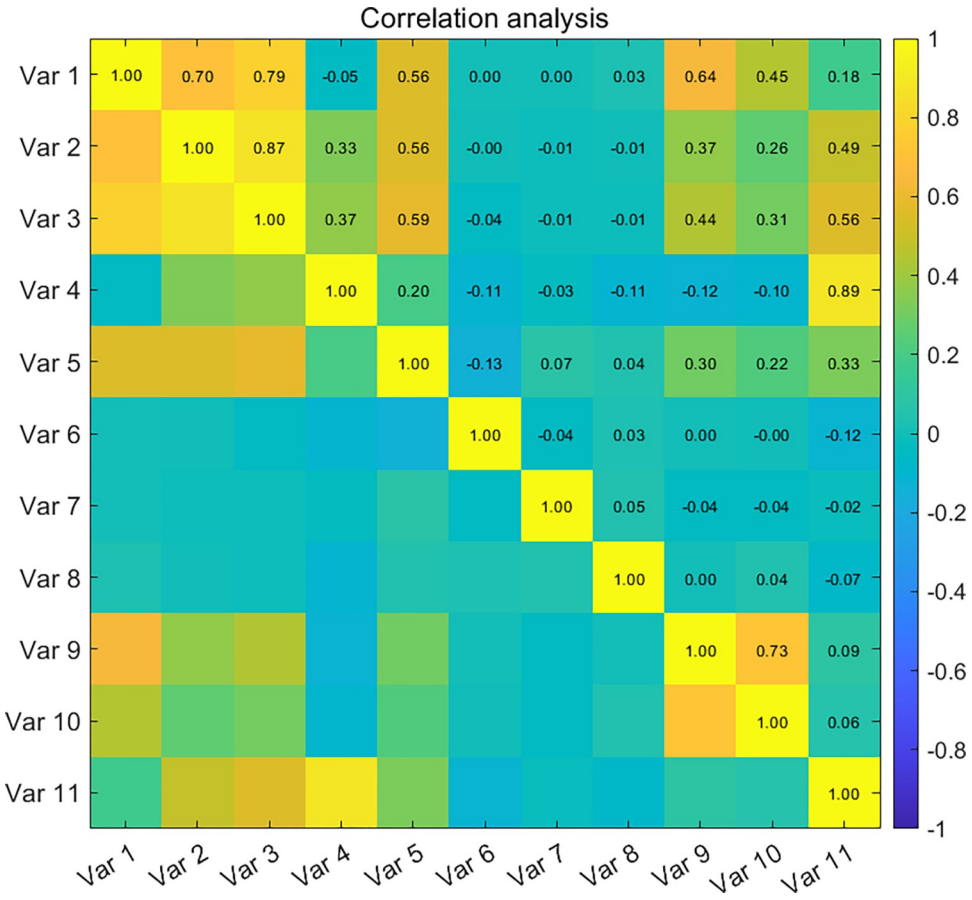

Fig 7. Correlations between different features.

https://doi.org/10.1371/journal.pone.0318519.g007

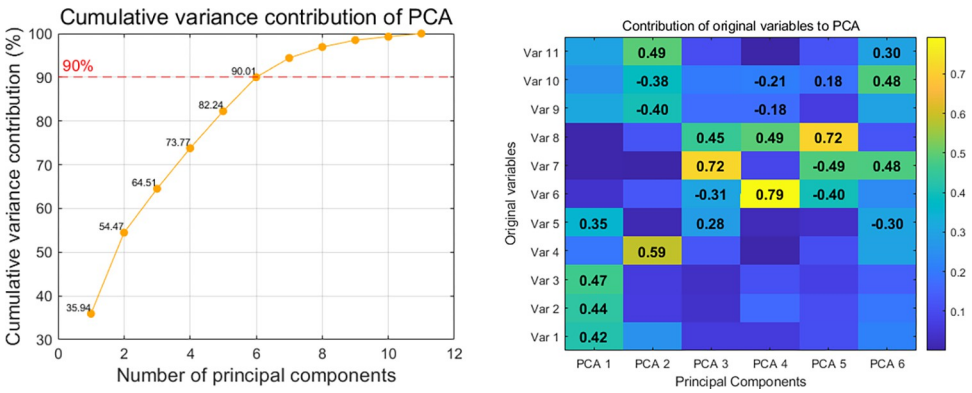

Fig 8. Principal component variance and features contribution. (a) Cumulative variance of PCA, (b) The contribution of the variables to PCA.

https://doi.org/10.1371/journal.pone.0318519.g008

Taking into account the varying degrees of correlation between variables and aiming to enhance the quality of customer segmentation, Fig 8(A) and 8(B) illustrate the results of the different principal components obtained through the PCA method, including the cumulative explained variance contribution and the importance of the top four original features. From Fig 8(A), it is evident that as the number of principal components increases, the growth rate of cumulative explained variance gradually decreases. To ensure effective clustering and minimize data noise, the first 6 principal components, which account for more than 90% of the explained variance, are selected as input. Furthermore, Fig 8(B) demonstrates that after PCA processing, the principal components effectively represent the various characteristics of the original customers, confirming that this approach serves as an effective dimensionality reduction strategy.

4.2 Clustering applications of *K*-means-QLDE

To achieve effective clustering with the *K*-means algorithm, careful selection of the number of clusters *K* is essential. To this end, we employed the widely used elbow method [67], which identifies an "elbow" point as the optimal number of clusters by calculating the SSE for various values of *K*, as illustrated in Fig 9 From the figure, it is evident that as the value of *K* increases, the SSE decreases, indicating that customer segmentation data can achieve a lower SSE with more clusters. It is important to highlight that at $K = 6$, a clear inflection point indicates a significant improvement in the clustering performance of the *K*-means method, marking it as the optimal number of clusters for achieving the best results in our analysis.

Fig 10 shows the segmentation results of different customer segments obtained through the *K*-means-QLDE algorithm. From the figure, it can be seen that Cluster 2 accommodates 47.9% of the customers, while Cluster 5 contains 21.3% of the customers. These two clusters account for the vast majority of all customers, allowing business managers can focus on analyzing the characteristics of this potential customer group and formulate corresponding business strategies. In the remaining clusters (1, 3, 4, and 6), the distribution of customers is relatively balanced. Although each cluster occupies a relatively small proportion, the overall level accounts for 30.8% of all customers. Therefore, it is essential to gain a deeper understanding of the characteristics of these customer groups to adapt to the needs of future business development.

Fig 11 further illustrates the customer segmentation results obtained using the *K*-means-QLDE algorithm. The figure clearly demonstrates that different customers form distinct clusters in the space defined by the first three principal components. Each cluster represents a

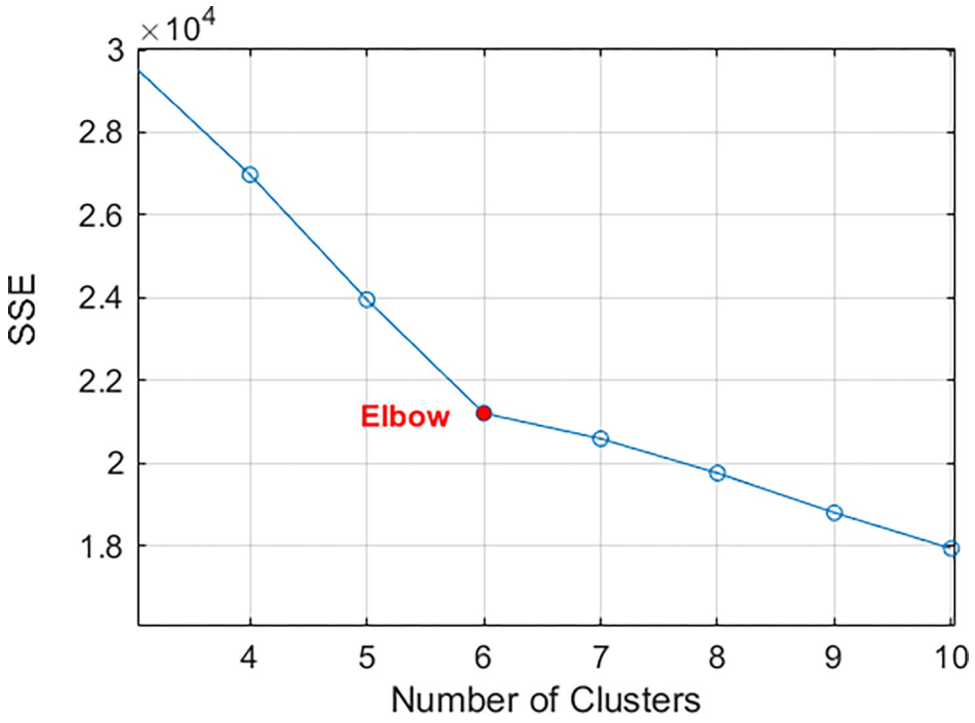

Fig 9. Relationship between the number of clusters and SSE.

https://doi.org/10.1371/journal.pone.0318519.g009

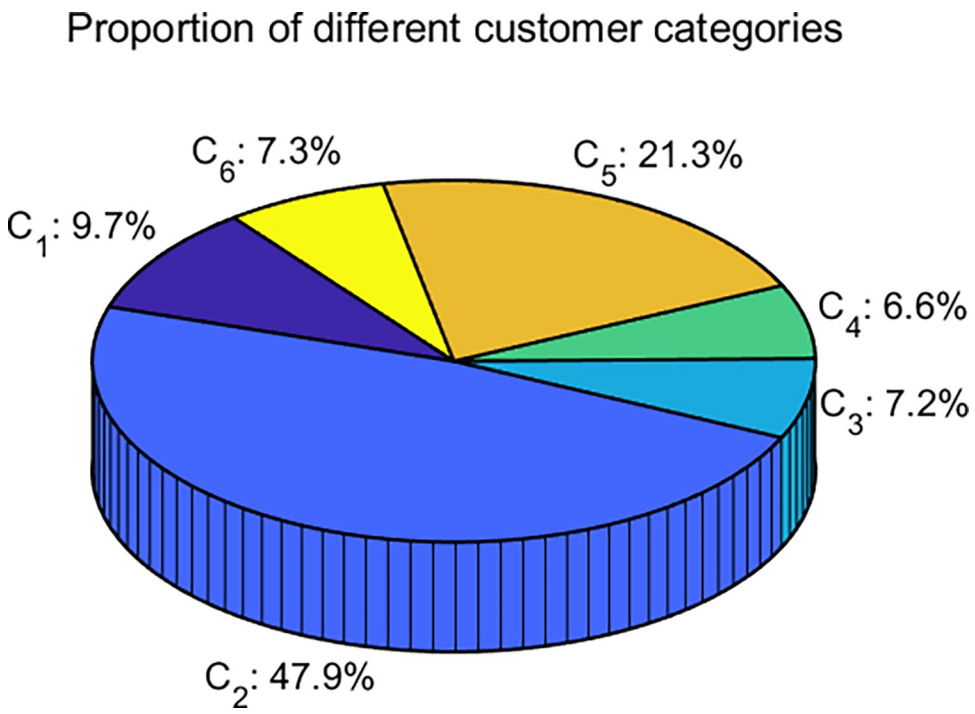

Fig 10. The proportion of different customer categories.

https://doi.org/10.1371/journal.pone.0318519.g010

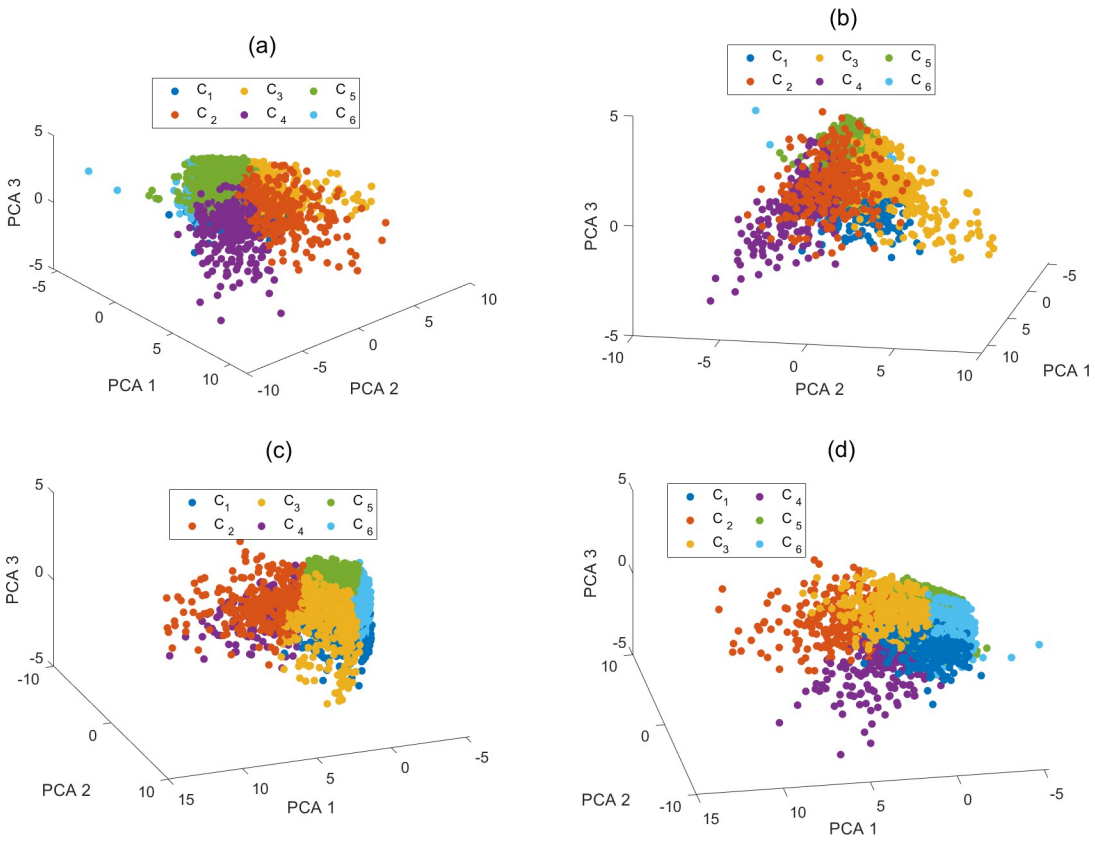

Fig 11. The clustering results in the first three principal components.

https://doi.org/10.1371/journal.pone.0318519.g011

group of customers with similar characteristics, and there is no overlap between the groups in the principal component space, underscoring the effectiveness and validity of the clustering results. Furthermore, analyzing these clusters allows for the identification of behavioral patterns and demand characteristics among various customer segments, providing valuable data support for the development of subsequent marketing strategies.

Fig 12 illustrates the normalized statistical ranking results of the six cluster centers obtained through the 11 extracted features. This figure intuitively shows the differences between different groups in different dimensions. For example, for the customers in Fig 12(A), they have the characteristics of high average monthly expenditure, the most expenditure products, and low transaction cancellation frequency. Such customers are usually the most important customer groups for enterprises because they contribute the most to revenue and purchase frequently.

The customers represented in Fig 12(B) constitute 47.9% of the total customer base. They exhibit the highest volume of product purchases; however, they demonstrate a lower average purchase frequency and monthly expenditure. This suggests that they are inclined to buy low-priced or discounted items, or that they approach their purchasing decisions with greater rationality. Such customers can be categorized as price-sensitive consumers. To effectively engage this group, the industry should develop targeted marketing strategies aimed at stimulating their consumption, thereby fostering the long-term growth of the enterprise.

Fig 12(C) and 12(D) mainly illustrate the characteristics of high-expectation customers, especially focusing on the relationship between transaction cancellation rate and consumer origin, especially consumers from the UK. As can be seen from the figure, these consumers

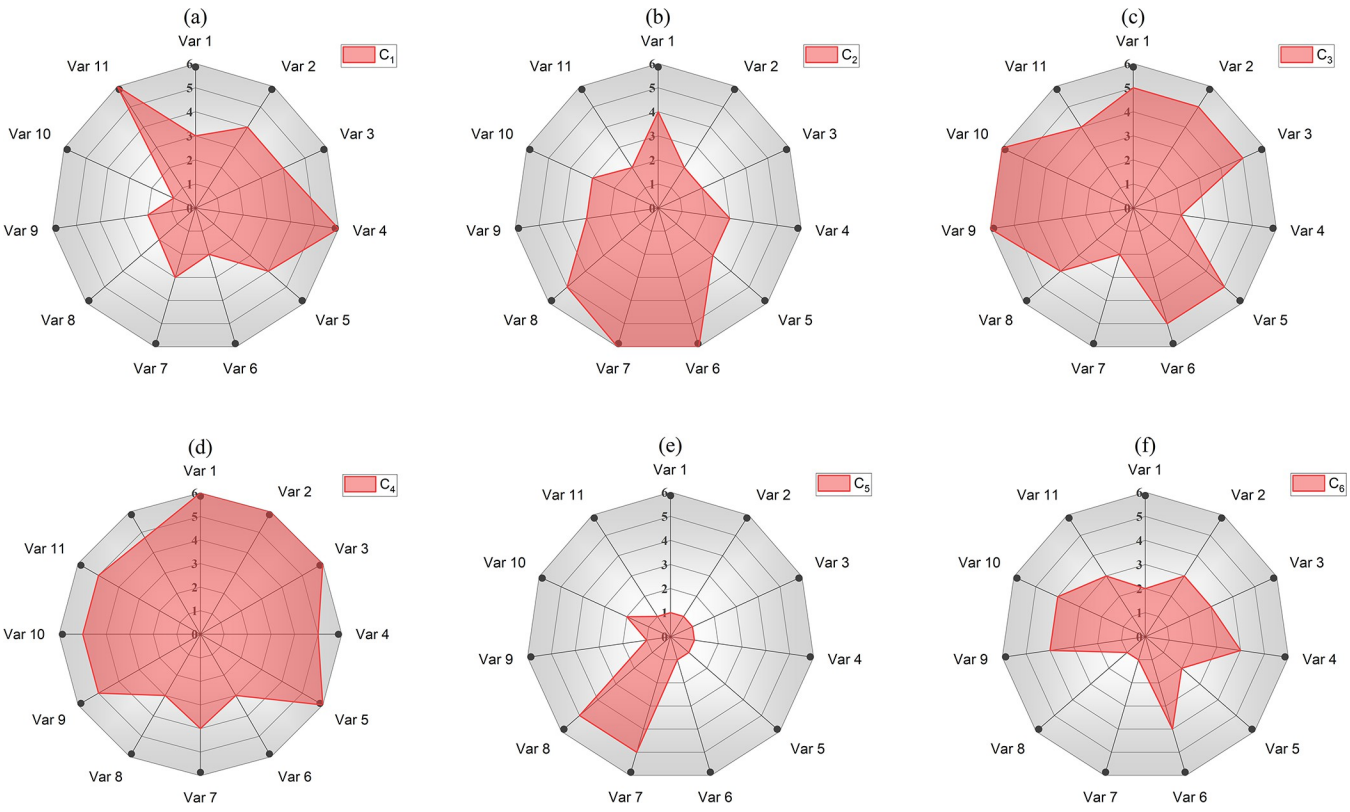

Fig 12. Ranking of cluster centers in different features.

https://doi.org/10.1371/journal.pone.0318519.g012

show a certain degree of uncertainty or indecision in their purchase decisions. In addition, British consumers may have higher expectations for product quality, which makes them more inclined to re-evaluate orders after purchase. Companies should consider implementing strategies to provide clearer product information and optimize the payment process to reduce transaction cancellation rates and improve customer satisfaction.

Fig 12(E) illustrates the characteristics of cautious consumers. Such customers exhibit a higher level of caution in making purchase decisions, often engaging in multiple comparisons and research. Their purchasing behavior is primarily concentrated on weekends, and they tend to be more rational and planned in their buying processes. Therefore, merchants can implement targeted marketing and promotional activities during the weekend. Fig 12(F) depicts the characteristics of balanced customers. These customers demonstrate relatively balanced performance across multiple purchasing indicators, without obvious preferences or extreme behaviors. Merchants can maintain and enhance customer relationships through diversified marketing strategies.

4.3 Classification accuracy verification based on machine learning

To further illustrate the practicality of the proposed refined customer segmentation method, the obtained clustering results were classified using several algorithms: Kernel Support Vector Machine (KSVM) [68], Decision Tree (DT) [69], AdaBoost [70], and Artificial Neural Network (ANN) [71]. 80% of the data was utilized as the training set, while 20% served for testing. Additionally, 5-fold cross-validation [72] was employed to adjust hyperparameters of the model during training, enhancing generalization ability and performance of the model. The

Table 2. Statistical results of classification using different methods.

Methods	Validation mean accuracy (%)	test set accuracy (%)	Time(Seconds)
KSVM	97.11	97.17	7.63
DT	95.88	95.82	10.15
AdaBoost	96.40	95.69	2.84
ANN	98.99	99.38	221.14

https://doi.org/10.1371/journal.pone.0318519.t002

first six principal components, derived from the input factors using the PCA method described in the previous section, were used as the input of the models.

The implementation was carried out using Statistics and Machine Learning Toolbox of MATLAB software. The analysis results are shown in Table 2, and the process of different classification accuracy is shown in Fig 13. The table and figure indicate that all methods achieved a classification accuracy exceeding 95% on the test set, demonstrating that the proposed customer segmentation method effectively ensures high classification accuracy. Notably, the ANN method achieved the highest classification accuracy, followed closely by KSVM. However, the time cost associated with the ANN is approximately 29 times greater than that of KSVM. Therefore, if merchants do not prioritize time efficiency, the ANN method may be preferred for customer classification; otherwise, KSVM is recommended. In summary, the proposed *K*-means-QLDE method effectively identifies key information among customers, significantly enhancing the marketing capabilities of merchants.

5. Conclusions and future works

This paper presents a novel method for customer segmentation in the digital marketing process, termed *K*-means-QLDE, which integrates heuristic algorithms with machine learning techniques. Initially, the PCA method is employed to extract principal component features from customer variables. Subsequently, a differential evolution algorithm based on *Q*-Learning (QLDE) is introduced to address the clustering problem, while the elbow method is utilized to ascertain the optimal number of clusters for the *K*-means algorithm. Utilizing a customer dataset sourced from Kaggle, the proposed method effectively segments customers into six distinct

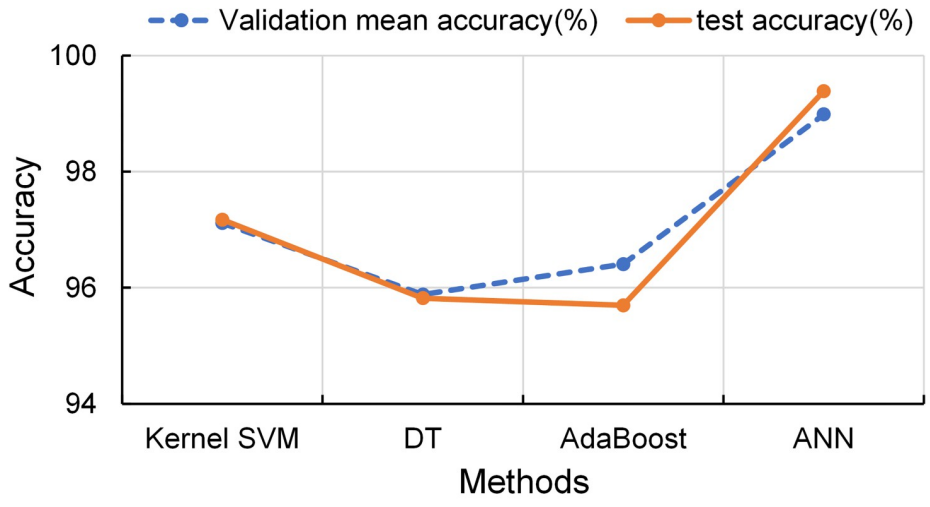

Fig 13. Classification accuracy of different methods.

https://doi.org/10.1371/journal.pone.0318519.g013

clusters, each representing varying levels of importance. To validate the practicality of this customer segmentation approach, four widely recognized methods are employed to assess the classification accuracy of the segmentation results. The findings indicate that the proposed method achieves a segmentation accuracy exceeding 95% on the test set. These results underscore the practical applicability of the *K*-means-QLDE segmentation method, empowering marketers to devise tailored digital marketing strategies for diverse customer groups, thereby enhancing the company's marketing revenue.

Nevertheless, this study acknowledges several limitations. Firstly, while the proposed QLDE algorithm demonstrates effective clustering performance, it simultaneously complicates model interpretation. Secondly, the inherent complexity of the QLDE algorithm may present challenges when applied to large datasets. In future research, we aim to address these limitations by investigating more transparent dimensionality reduction techniques and optimization algorithms. Additionally, we plan to broaden the scope of our research to encompass high-dimensional datasets across various industries.

Acknowledgments

We would like to express our sincere gratitude to the reviewers for their valuable comments and suggestions, which significantly improved the quality of this manuscript.

Author Contributions

Conceptualization: Guanqun Wang.

Data curation: Guanqun Wang.

Formal analysis: Guanqun Wang.

Funding acquisition: Guanqun Wang.

Investigation: Guanqun Wang.

Methodology: Guanqun Wang.

Project administration: Guanqun Wang.

Resources: Guanqun Wang.

Software: Guanqun Wang.

Supervision: Guanqun Wang.

Validation: Guanqun Wang.

Visualization: Guanqun Wang.

Writing – original draft: Guanqun Wang.

Writing – review & editing: Guanqun Wang.

References

1. Malecki EJ, Moriset B. The digital economy: Business organization, production processes and regional developments.: Routledge; 2007.
2. Litvinenko VS. Digital economy as a factor in the technological development of the mineral sector. Nat Resour Res. 2020; 29(3):1521–41.
3. Teece DJ. Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world. Res Policy. 2018; 47(8):1367–87.
4. Pan W, Xie T, Wang Z, Ma L. Digital economy: An innovation driver for total factor productivity. J Bus Res. 2022; 139:303–11.

5. Kannan PK. Digital marketing: A framework, review and research agenda. Int J Res Mark. 2017; 34 (1):22–45.
6. Gupta S, Justy T, Kamboj S, Kumar A, Kristoffersen E. Big data and firm marketing performance: Findings from knowledge-based view. Technol Forecast Soc. 2021; 171:120986.
7. Tolstoy D, Nordman ER, Vu U. The indirect effect of online marketing capabilities on the international performance of e-commerce SMEs. Int Bus Rev. 2022; 31(3):101946.
8. Brenner B. Transformative sustainable business models in the light of the digital imperative—A global business economics perspective. Sustainability-Basel. 2018; 10(12):4428.
9. Hojeghan SB, Esfangareh AN. Digital economy and tourism impacts, influences and challenges. Procedia-Social and Behavioral Sciences. 2011; 19:308–16.
10. Wang C. Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. Inform Process Manag. 2022; 59(6):103085. https://doi.org/10.1016/j.ipm.2022.103085
11. Floh A, Zauner A, Koller M, Rusch T. Customer segmentation using unobserved heterogeneity in the perceived-value–loyalty–intentions link. J Bus Res. 2014; 67(5):974–82. https://doi.org/10.1016/j.jbusres.2013.08.003
12. Kim S, Jung T, Suh E, Hwang H. Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Syst Appl. 2006; 31(1):101–7. https://doi.org/10.1016/j.eswa.2005.09.004
13. Sarvari PA, Ustundag A, Takci H. Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes. 2016; 45(7):1129–57.
14. Kasem MS, Hamada M, Taj-Eddin I. Customer profiling, segmentation, and sales prediction using AI in direct marketing. Neural Computing and Applications. 2024; 36(9):4995–5005.
15. Tang YE, Mantrala MK. Incorporating direct customers' customer needs in a multi-dimensional B2B market segmentation approach. Ind Market Manag. 2024; 119:252–63.
16. Williams DS. Connected CRM: implementing a data-driven, customer-centric business strategy.: John Wiley & Sons; 2014.
17. Gangale F, Mengolini A, Onyeji I. Consumer engagement: An insight from smart grid projects in Europe. Energ Policy. 2013; 60:621–8.
18. Zhang JZ, Chang C. Consumer dynamics: Theories, methods, and emerging directions. J Acad Market Sci. 2021; 49:166–96.
19. Sheth JN. Impact of emerging markets on marketing: Rethinking existing perspectives and practices. J Marketing. 2011; 75(4):166–82.
20. Dolnicar S, Kaiser S, Lazarevski K, Leisch F. Biclustering: Overcoming data dimensionality problems in market segmentation. J Travel Res. 2012; 51(1):41–9.
21. Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, et al. Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. Mathematics-Basel. 2022; 10(15):2552.
22. Qin SJ, Chiang LH. Advances and opportunities in machine learning for process data analytics. Comput Chem Eng. 2019; 126:465–73.
23. Miklosik A, Kuchta M, Evans N, Zak S. Towards the adoption of machine learning-based analytical tools in digital marketing. Ieee Access. 2019; 7:85705–18.
24. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2022:1–66.
25. Talaat FM, Aljadani A, Alharthi B, Farsi MA, Badawy M, Elhosseini M. A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing. Mathematics-Basel. 2023; 11(18):3930.
26. Chen H, Zhang L, Chu X, Yan B. Smartphone customer segmentation based on the usage pattern. Adv Eng Inform. 2019; 42:101000.
27. Wang B, Miao Y, Zhao H, Jin J, Chen Y. A biclustering-based method for market segmentation using customer pain points. Eng Appl Artif Intel. 2016; 47:101–9.
28. Joung J, Kim H. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. Int J Inform Manage. 2023;70. http://doi.org/10.1016/j.ijinfomgt.2023.102641
29. Chang H, Tsai H. Group RFM analysis as a novel framework to discover better customer consumption behavior. Expert Syst Appl. 2011; 38(12):14499–513.

30. Cheng C, Chen Y. Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst Appl. 2009; 36(3):4176–84.

31. Christy AJ, Umamakeswari A, Priyatharsini L, Neyaa A. RFM ranking–An effective approach to customer segmentation. Journal of King Saud University—Computer and Information Sciences. 2021; 33 (10):1251–7. https://doi.org/10.1016/j.jksuci.2018.09.004

32. Rungruang C, Riyapan P, Intarasit A, Chuarkham K, Muangprathub J. RFM model customer segmentation based on hierarchical approach using FCA. Expert Syst Appl. 2024; 237:121449. https://doi.org/10.1016/j.eswa.2023.121449

33. Dursun A, Caber M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tour Manag Perspect. 2016; 18:153–60.

34. Barrera F, Segura M, Maroto C. Multiple criteria decision support system for customer segmentation using a sorting outranking method. Expert Syst Appl. 2024; 238:122310. https://doi.org/10.1016/j.eswa.2023.122310

35. Lemley J, Bazrafkan S, Corcoran P. Deep learning for consumer devices and services: pushing the limits for machine learning, artificial intelligence, and computer vision. Ieee Consum Electr M. 2017; 6 (2):48–56.

36. Dong S, Wang P, Abbas K. A survey on deep learning and its applications. Comput Sci Rev. 2021; 40:100379.

37. Nguyen SP. Deep customer segmentation with applications to a Vietnamese supermarkets' data. Soft Comput. 2021; 25(12):7785–93.

38. Yadegaridehkordi E, Nilashi M, Nizam Bin Md Nasir MH, Momtazi S, Samad S, Supriyanto E, et al. Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. Technol Soc. 2021; 65:101528. https://doi.org/10.1016/j.techsoc.2021.101528

39. Chen P. Effects of normalization on the entropy-based TOPSIS method. Expert Syst Appl. 2019; 136:33–41.

40. Timofeev R. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin. 2004; 54:48.

41. Nilashi M, Ahmadi H, Arji G, Alsalem KO, Samad S, Ghabban F, et al. Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. J Retail Consum Serv. 2021; 62:102630. https://doi.org/10.1016/j.jretconser.2021.102630

42. Canziani A, Paszke A, Culurciello E. An Analysis of Deep Neural Network Models for Practical Applications. Arxiv. 2016;abs/1605.07678.

43. Alkhayrat M, Aljnidi M, Aljoumaa K. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data-Ger. 2020; 7(1):9. http://doi.org/10.1186/s40537-020-0286-0

44. Thamer MK, Algamal ZY, Zine R. Enhancement of Kernel Clustering Based on Pigeon Optimization Algorithm. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2023; 31:121–33. http://doi.org/10.1142/S021848852340007X

45. Al-Kababchee SGM, Algamal ZY, Qasim OS. Enhancement of K-means clustering in big data based on equilibrium optimizer algorithm. J Intell Syst. 2023; 32(1). http://doi.org/10.1515/jisys-2022-0230

46. Al Radhwani AMN, Algamal ZY, ^editors.Improving K-means clustering based on firefly algorithm. Journal of Physics: Conference Series; 2021 星期五 2021/1/1. Pub Place: IOP Publishing Ltd; Year Published.

47. Al-Kababchee SGM, Algamal ZY, Qasim OS. Improving penalized-based clustering model in big fusion data by hybrid black hole algorithm. Fusion: Practice and Applications. 2023; 11(1):70–6. http://doi.org/10.54216/FPA.110105

48. Böckenholt U. Market segmentation:: Conceptual and methodological foundations. J Classif. 2000; 17 (1):143–5.

49. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2(4):433–59. http://doi.org/10.1002/wics.101

50. Gewers FL, Ferreira GR, De Arruda HF, Silva FN, Comin CH, Amancio DR, et al. Principal Component Analysis: A Natural Approach to Data Exploration. Acm Comput Surv. 2021; 54(4):70. http://doi.org/10.1145/3447755

51. Likas A, Vlassis N, J. Verbeek J. The global k-means clustering algorithm. Pattern Recogn. 2003; 36 (2):451–61. https://doi.org/10.1016/S0031-3203(02)00060-2

52. M. R, A. BS. A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm. Ieee Access. 2022; 10:86286–97. http://doi.org/10.1109/ACCESS.2022.3198992

53. Pant M, Zaheer H, Garcia-Hernandez L, Abraham A. Differential Evolution: A review of more than two decades of research. Eng Appl Artif Intel. 2020; 90:103479.
54. Liao TW. Two hybrid differential evolution algorithms for engineering design optimization. Appl Soft Comput. 2010; 10(4):1188–99.
55. Juwono FH, Wong WK, Pek HT, Sivakumar S, Acula DD. Ovarian cancer detection using optimized machine learning models with adaptive differential evolution. Biomed Signal Proces. 2022; 77:103785.
56. Basu M. Economic environmental dispatch using multi-objective differential evolution. Appl Soft Comput. 2011; 11(2):2845–53.
57. Yang D, Liu Z, Zhou J. Chaos optimization algorithms based on chaotic maps with different probability distribution and search speed for global optimization. Commun Nonlinear Sci. 2014; 19(4):1229–46.
58. Jang B, Kim M, Harerimana G, Kim JW. Q-learning algorithms: A comprehensive classification and applications. Ieee Access. 2019; 7:133653–67.
59. Yan Y, Chow AH, Ho CP, Kuo Y, Wu Q, Ying C. Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. Transportation Research Part E: Logistics and Transportation Review. 2022; 162:102712.
60. Dabbaghjamanesh M, Moeini A, Kavousi-Fard A. Reinforcement learning-based load forecasting of electric vehicle charging station using Q-learning technique. Ieee T Ind Inform. 2020; 17(6):4229–37.
61. Kiumarsi B, Lewis FL, Modares H, Karimpour A, Naghibi-Sistani M. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. Automatica. 2014; 50 (4):1167–75.
62. Clifton J, Laber E. Q-learning: Theory and applications. Annu Rev Stat Appl. 2020; 7(1):279–301.
63. Liu X, Zhang P, Fang H, Zhou Y. Multi-objective reactive power optimization based on improved particle swarm optimization with ε-greedy strategy and pareto archive algorithm. Ieee Access. 2021; 9:65650–9.
64. Vamplew P, Dazeley R, Foale C. Softmax exploration strategies for multiobjective reinforcement learning. Neurocomputing. 2017; 263:74–86.
65. F. D. "Customer segmentation." https://www.kaggle.com/code/fabiendaniel/customer-segmentation/notebook.; 2017.
66. Komatsu H, Kimura O. Customer segmentation based on smart meter data analytics: Behavioral similarities with manual categorization for building types. Energ Buildings. 2023; 283:112831. https://doi.org/10.1016/j.enbuild.2023.112831
67. Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. Iop Conference Series: Materials Science and Engineering. 2018; 336(1):12017. http://doi.org/10.1088/1757-899X/336/1/012017
68. Howley T, Madden MG. The Genetic Kernel Support Vector Machine: Description and Evaluation. Artif Intell Rev. 2005; 24(3):379–95. http://doi.org/10.1007/s10462-005-9009-3
69. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27(2):130–5. https://doi.org/10.11919/j.issn.1002-0829.215044 PMID: 26120265
70. Schapire RE. Explaining AdaBoost. In: Schölkopf B, Luo Z, Vovk V, ^editors. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 37–52.
71. Grnholdt L, Martensen A. Analysing Customer Satisfaction Data: A Comparison of Regression and Artificial Neural Networks. Int J Market Res. 2005; 47(2):121–30. http://doi.org/10.1177/147078530504700201
72. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 2011; 21 (2):137–46. http://doi.org/10.1007/s11222-009-9153-8