Hindawi
Scientific Programming
Volume 2021, Article ID 7622576, 12 pages
https://doi.org/10.1155/2021/7622576

Research Article

A Clustering-based Method for Business Hall Efficiency Analysis

Tianlin Huang and Ning Wang

College of Information and Smart Electromechanical Engineering, Xiamen Huaxia University, Xiamen 361024, China

Correspondence should be addressed to Ning Wang; nwang97@163.com

Received 20 May 2021; Revised 11 August 2021; Accepted 7 September 2021; Published 1 October 2021

Academic Editor: Antonio J. Peña

Excessive or insufficient business hall resources may result in unreasonable resource allocation, adversely affecting the value of an entity business hall. Therefore, proper characteristic parameters are the key factors for analyzing the business hall, which strongly affect the final analysis results. In this study, a characteristic analysis method for the economic operation of a business hall is developed and the feature engineering is established. Because of its simplicity and versatility, the k-means algorithm has been widely used since it was first proposed around 50 years ago. However, the classical k-means algorithm has poor stability and accuracy. In particular, it is difficult to achieve a suitable balance between of the centroid initialization and the clustering number k. We propose a new initialization (LSH-k-means) algorithm for k-means clustering. This algorithms is mainly based on locality-sensitive hashing (LSH) as an index for computing the initial cluster centroids, and it reduces the range of the clustering number. Furthermore, an empirical study is conducted. According to the load intensity and time change of the business hall, an index system reflecting the optimization analysis of the business hall is established, and the LSH-k-means algorithm is used to analyze the economic operation of the business hall. The results of the empirical study show that the LSH-k-means that the clustering method outperforms the direct prediction method, provides expected analysis results as well as decision optimization recommendations for the business hall, and serves as a basis for the optimal layout of the business hall.

1. Introduction

An entity business hall is where a company directly conducts specific business activities, such as commodity trade, business handling, and service. However, owing to rapid urbanization and economic development, unreasonable resource allocation is becoming increasingly prevalent. For example, the number of entity business halls is excessive in some places and insufficient in others. Hence, the deployment of new commercial outlets (halls) or resource allocation optimization for existing retail outlets often needs to be performed manually. Therefore, how to evaluate the efficiency of business halls has emerged as a major concern for many enterprises.

To this end, many researchers have attempted to overcome the disadvantages of human judgment, which is highly subjective. Brandeau and Chiu [1] considered the transportation cost and the distance between the warehouse and the customer and used a gradient-like algorithm to study the location issue. Wang et al. [2] used nearest-neighbor clustering and the function of Ripley [3] to analyze the layout of commercial outlets and suggested that business type, land price, and traffic accessibility are the critical factors. Gerard [4] analyzed the service needs and waiting demand of customers for bank halls and attempted to shorten the perceived waiting time of customers on the basis of the customers' business types. Thus, customer satisfaction was improved. Anderson et al. [5] used the queuing model to optimize the queuing service system of banks. They determined the optimal number of service windows by acquiring and presenting a large amount of data. Lin et al. [6] studied the relationship between retail stores and street centrality and pointed out that besides the transport network, which has a strong impact on the retailer's location, the street centrality influences the type of retail store. Kang [7] analyzed the changes in warehouses from central urban areas to the urban periphery over time and studied the main factors affecting the warehouse location. Hui [8] used data mining to establish the channel analysis model for an electricity business hall and optimized the resource allocation. Based on the statistics of customer queuing time, business processing time, customer satisfaction, and so on, Yan et al. [9] established an intelligent access platform for the business data and improved the service efficiency. However, there is no unified standard for the business hall index system.

Clustering is a key technique in data mining, and its applications include pattern recognition [10, 11], image processing [12], and recommendation [13]. Clustering aims to partition data into different categories based on a measure of similarity. The k-means algorithm is widely used owing to its simplicity and effectiveness. However, the different settings of the parameters and random selection of the initial clustering centers make the classical k-means algorithm unstable.

The classical clustering algorithm involves two problems: the first problem is to classify a given dataset on the basis of the prespecified cluster number k; hence, the problem of determining the "correct cluster number" has attracted considerable interest. Although several methods have been developed for estimating the number of data clusters [14–17], it is difficult to use them in practical applications. Therefore, determining the correct number of clusters has long been an important research topic in cluster analysis. The second problem is to determine the initial clustering center, which has a significant impact on the clustering effect. Studies conducted thus far have explored several initialization methods for the *thek*-means algorithm. For example, the k-means++ algorithm [18] has been proposed to avoid this issue. This algorithm randomly selects the first centroid, and the other centroids are selected as far away as possible from the first centroid. However, random selection is still widely used in practice [19]. Erisoglu et al. [20] proposed an incremental approach for computing the initial clustering centers. In this approach, the reduced dataset is partitioned until the number of clusters equals the predefined number of clusters. However, the number of clusters must be known in advance. The compressed k-means (CKM) algorithm [21] is initialized by locality-sensitive hashing (LSH) [22], and the distance is calculated using the Hamming distance between binary codes. The LSH link [23] can rapidly find a nearby cluster to be connected through the LSH algorithm. David et al. [24] proposed a new LSH scheme adapted to the x^2 distance for approximate nearest neighbors (ANN) search in high-dimensional spaces.

In summary, there is no unified standard for the index system of business halls at present. Therefore, we establish an index system for analyzing the efficiency of a business hall. To address the problem of k-means initialization sensitivity as well as the difficulty in determining the number of clusters, we initialize the k-means centroid on the basis of LSH. Accordingly, we implement the relevant algorithms and present the optimal allocation scheme for the business hall.

The main contributions of this study are as follows:

(1) According to the average waiting time, ticketing time, and business type of a business hall, we analyze the average load rate of the business hall and use the relevant characteristic variables to describe the load of the business hall. Finally, we propose a general business hall index system.

(2) By combining the characteristics of k-means and LSH, We propose a new initialization (LSH-k-means) algorithm for k-means clustering. The model can get the load classification of each business hall by inputting the relevant index variables for the optimization of business hall distribution.

(3) The results of our empirical analysis verify the validity of the proposed LSH-k-means approach. Thus, LSH-k-means can be efficiently used for the operational analysis of a business hall.

The remainder of this paper is organized as follows: Section 2 introduces the required preliminaries, definitions, and models. Section 3 describes the proposed initialization methodology. Section 4 presents, compares, and discusses the experimental results. Finally, Section 5 concludes the paper.

2. Preliminaries

2.1. k-means Algorithm. The notations used in this paper are defined in Table 1. The k-means [25] method is the most well-known clustering method because of its simplicity. It has been identified as one of the top 10 algorithms in data mining [26]. Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$, k-means aims to partition it into k different clusters $C = \{C_1, C_2, \ldots, C_k\}$, where $k \leq n$ is a predefined number. The objective of the k-means clustering algorithm is to minimize the sum of squared errors (SSE) [27] over all k clusters. The SSE is defined as follows:

$$\mathrm{SSE} = \sum_j \sum_{x \in C_j} \left\| x - \Omega_j \right\|_2^2, \tag{1}$$

where Ω_j denotes the j-th cluster centroid, which is computed as the mean of points in C_j, and $x \in C_j$ is the data object in the j-th cluster.

$$\Omega_j = \frac{1}{\left|C_j\right|} \sum_{x_i \in C_j} x_i, \tag{2}$$

where $|C_j|$ denotes the number of data points in the j-th cluster.

To solve equation (1), an expectation–maximization (EM)-like optimization method is adopted by updating $I(x)$ or Ω and simultaneously fixing the other [28]. In general, the clustering procedure involves three steps: (1) initialize k cluster centroids; (2) assign each sample to its closest centroid; and (3) recompute the cluster centroids with the assignments produced in Step 2 and go back to Step 2 until convergence. This is known as the Lloyd iteration procedure [29]. Such an iterative optimization approach has several drawbacks. First, it is sensitive to the initialization, which may lead to an inferior result for a given poorly initialized Ω. Many methods have been proposed to obtain a stable solution, including the k-means++ algorithm [18]. Second,

Table 1: The notations used throughout this paper.

Notation	Representation
D	The training set with N points and d dimensions
x_i	The data point
C_j	The j-th clusters
SSE	The sum of squared errors
Ω_j	Denotes the j-th cluster centroid
$\lvert C_j\rvert$	The number of data points in the j-th cluster
$O(tkn\ d)$	Denotes the complexity of k-means
H	A family of hash functions
Pr	The probability
$\mathrm{sim}(x, y)$	The similarity between x and y
M	The maximum load of business hall
p_i	The proportion of a specific business
w_i	The average time to process the business
n	The number of business types
K	The ratio of actual daily load to maximum load
S_w	The working time
S	The number of staffs in the business hall
A_i	Represents the actual load during peak period of the business hall
MS	The ratio of average load to the maximum load
W	The number of working days
AT	The actual load trend
$f(l)$	Denotes the actual load curve fitting function
b_0	A constant
b_1	The regression coefficient
BH	The proportion of high-value business
HBV_i	The high-value business volume
TBV_i	The total business volume
FH	The high-frequency load
h	The high threshold
l	The low threshold
T_{now}	Represent the current time
T_{latest}	Represent the latest time
FL	The low-frequency load
RH	The latest high-load interval
RL	The latest low-load interval
U	The copy training set
N	The size of training dataset
T	The minimum number data point in one cluster
L	The maximum distance in one cluster
B_n	A two-dimensional array
Q	The nearest neighbor data
$\mathrm{query}(x_i)$	Calculate the similarity or distance between x_i
k	The number of clusters
$d(x_i, x_j)$	Distance between x_i and x_j
KM	Conduct k-means algorithm
d_i	The i-th dataset
R^+	Represents the sum of rank, which better than the other
R^-	Opposite to R^+
$S(U, k)$	The tight and separative indicator
XB	The XB index
$\mathrm{Obj}_{\min}$	Objective function
F	The quantitative value of factors
h_i	The weight

finding the optimal solution to k-means is an NP-hard problem. Some variants of k-means have been proposed, such as various parametric k-means, including fuzzy c-means [30, 31]. Third, k-means cannot handle new data, which requires the entire dataset to be observed. The complexity is $O(t \cdot k \cdot n \cdot d)$, where t, n, k, and d denote the number of iterations, size of the dataset, number of clusters, and dimensionality, respectively. This complexity is considerably higher than that of other well-known clustering algorithms such as DBSCAN [32] and mean shift [33].

2.2. LSH. LSH is a well-known solution for the approximate nearest neighbor problem in high-dimensional spaces. LSH was first introduced for the Hamming metric by Indyk and Motwani [34]. Data points are assigned to individual hash buckets in each hash function. The idea of LSH is that closer data points are mapped to the same hash bucket with high probability. LSH has been shown to be effective even for high-dimensional data, both theoretically and experimentally [35]. H are a family of hash functions. Each hash function H must satisfy the LSH property: $\Pr[H(x) = H(y)] = \mathrm{sim}(x, y)$, where $\mathrm{sim}(x, y)$ is the similarity between x and y. These hash functions must meet the following two conditions:

(1) If $d(x, y) \le r$, then $\Pr[H(x) = H(y)] \ge p_1$

(2) If $d(x, y) \ge cr$, then $\Pr[H(x) = H(y)] \le p_2$

where $d(x, y)$ represents the distance measure between x and y, $r > 1$, and $c < 1$. The definition implies that x and y are hashed into the same bucket in the projection with a very high probability $\ge p_1$. Regardless of whether they are close to each other, they will be hashed into the same bucket $\le p_2$ with a low probability. A (r, cr, p_1, p_2)-sensitive family of hash functions is useful when the collision probabilities p_1, p_1 satisfy $p_1 > p_2$. Figure 1 shows an example of hashing key space.

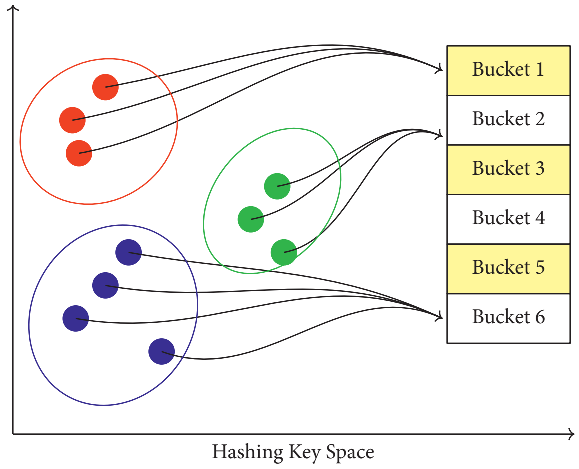


Figure 1: Illustration of hashing key space.

3. Proposed LSH-Based Initialization Algorithm

The proposed framework involves three steps: (1) an index system for the efficiency analysis of a business hall is established in Section 3.1. (2) To overcome the problems of poor stability and low accuracy of the classical algorithm, a boost k-means algorithm based on LSH initialization is proposed in Section 3.2. (3) The k-means algorithm is implemented to obtain the clustering results. The details of these three steps are illustrated in Figure 2.

3.1. Establishment of the Index System. Through the load analysis of the business hall, we can determine the high and low loads and optimize the business hall. The average utilization rate of each business hall is analyzed according to multiple indicators (including average waiting time, ticketing time for business, and business type). Thus, we can use the relevant characteristic variables to describe the load of the business hall. By applying the clustering algorithms, we can obtain the load categories of different business halls, which provides a basis for planning the locations of the business halls. First, the following two essential features are extracted: the maximum load of the business hall (**M**) and the ratio of the actual daily load to the maximum load (**K**).

(1) The maximum load of the business hall (**M**) is given by

$$M = \frac{S \times S_w}{\sum_{i=1}^{n} p_i w_i}, \tag{3}$$

where p_i is the proportion of a specific business, w_i is the average time for the clerk to handle the business, n is the number of business types, S_w is the working time of the clerk, and S is the number of clerks in the business hall. The variables are taken from the peak period. This value represents the maximum business volume that a business hall can withstand during the peak period. The peak period can be obtained by measuring the historical data of each business hall.

(2) The ratio of the actual daily load to the maximum load (**K**) is given by

$$K = \frac{A}{M}, \tag{4}$$

where A represents the actual daily load of the business hall and M is the maximum load in one day.

By combining the essential characteristics of the business hall and based on the analysis of historical data, we can obtain the calculation indicators of the business hall to prepare for the subsequent model input. Therefore, the feature engineering for the business hall efficiency analysis is established, and the critical indexes extracted are as follows:

(1) The ratio of the average load to the maximum load (**MS**) is given by

$$\mathrm{MS} = \frac{\sum_{i=1}^{W} A_i}{(W \times M)}, \tag{5}$$

where W is the number of days, and A and M are the same as above. This index denotes the ratio of the average actual load to the maximum load over some time.

(2) The actual load trend (**AT**) is given by

$$f(l) = b_0 + b_1 t,$$
$$\mathrm{AT} = b_1 + \frac{\sum f(l) \times t - n \times f(l) \times t}{\sum t_i^2 - n \times t^2}, \tag{6}$$

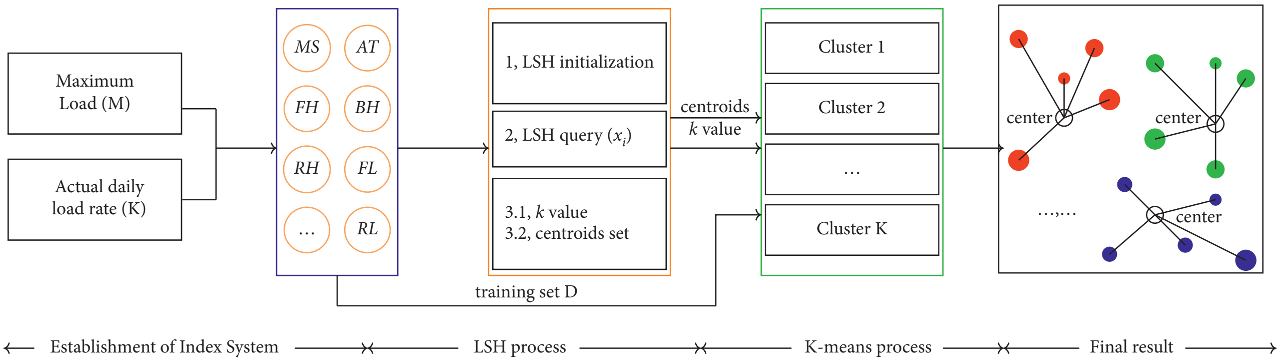


Figure 2: Architecture of the proposed LSH k-means model.

where $f(l)$ denotes the actual load curve fitting, b_0 is a constant, b_1 is the regression coefficient, t is a time-independent variable, and n is the number of statistical data. This index indicates whether the load trend of the business hall will be rising, flat, or declining for some time. A fitting curve $f(l)$ can be used to characterize the load trend of the business hall over some time, and the slope b_1 represents the trend state. Our method includes a commercial center, residential center, new urban area, and other factors.

(3) The proportion of high-value business (**BH**) is given by

$$\mathrm{BH} = \frac{\sum_{i=1}^{W} HBV_i}{\sum_{i=1}^{W} TBV_i}, \tag{7}$$

where HBV_i is the high-value business volume and TBV_i is the total business volume. Thus, this index denotes the proportion of high-value business to total business in the peak period.

(4) The high-frequency load (**FH**) is given by

$$\mathrm{FH} = \frac{\operatorname{count}(K_i > h)}{W}, \quad i = 1, 2, \ldots, W, \tag{8}$$

where h is a high threshold and FH represents the load of K exceeding h within a period. Furthermore, h can be obtained by statistical analysis of the historical data of the business hall.

(5) The low-frequency load (**FL**) is given by

$$\mathrm{FL} = \frac{\operatorname{count}(K_i < l)}{W}, \quad i = 1, 2, \ldots, W, \tag{9}$$

where l is a low threshold and FL denotes the frequency that K is less than l for some time. Furthermore, l can be obtained by statistical analysis the of historical data of the business hall.

(6) The latest high-load interval (**RH**) is given by

$$\mathrm{RH} = T_{\mathrm{now}} - T_{\mathrm{latest}\,(K_i > h)}, \quad i = 1, 2, \ldots, W, \tag{10}$$

where T_{now} represents the current time, h denotes a high threshold, and $T_{\mathrm{latest}\,(K_i>h)}$ refers to the time when the latest K is greater than h. Furthermore, RH denotes the interval from T_{now} to $T_{\mathrm{latest}\,(K_i>h)}$.

(7) The latest low-load interval (**RL**) is given by

$$\mathrm{RL} = T_{\mathrm{now}} - T_{\mathrm{latest}\,(K_i < l)}, \quad i = 1, 2, \ldots, W, \tag{11}$$

where T_{now} represents the current time, l denotes a low threshold, and $T_{\mathrm{latest}\,(K_i>h)}$ refers to the time when the latest K is greater than h. Furthermore, RL denotes the interval from T_{latest} to $T_{\mathrm{latest}\,(K_i>h)}$.

3.2. LSH-k-Means. The main purpose of clustering is to divide data into clusters in which objects in the same cluster are close to one another, whereas objects in different clusters are far from one another. Two factors affect the quality of k-means clustering. Before applying the algorithm, we need to specify the number of clusters k and select the initial cluster centroid. Selecting an appropriate initial cluster centroid can improve the quality of clustering. To this end, a critical study was conducted by Vassilvitskii et al. [18, 36]. If the initial cluster centroid is selected carefully, the k-means algorithm converges to a better local optimal solution. Furthermore, careful selection of the initial cluster centroid makes the k-means iteration converge faster [18]. However, to make the initial centroid adapt to the data distribution, it is necessary to scan k rounds. Therefore, although the number of scanning wheels in [36] has been reduced to a small value, the additional computing cost is still inevitable. Our algorithm exploits LSH. The algorithm minimizes the path by adding the nearest neighbor, and LSH can effectively search for the nearest group features in the path. The average time complexity of the hash-based search is $O(1)$. LSH scans the data records and finds the nearest points; the average values are computed after the nearest points are classified as a category. Algorithm 1 describes the process of obtaining the initialization centroids in our proposed LSH-k-means scheme. The main steps are as follows:

(1) Suppose that we have a set of points $D \in \mathbb{R}^d$ via the index system in Section 3.1. We use LSH to index the feature vectors extracted from the dataset D to reduce the search time for the nearest neighbor of each query. This is based on the hash mapping function, hash functions, and hash table L [37]. Constructing

an effective LSH index structure for approximate nearest neighbor search depends on the number of hash tables L and the number of bits V of the hash codes.

(2) To facilitate the statistics of nonclustered data points, in Algorithm 1, we copy a dataset U from D. Randomly select one data point x_i from U as the centroid. Then, x_i is merged into the set A_n and removed from the dataset U, where n is the n-th cluster. After obtaining x_i points, query the corresponding bucket number according to the hash table L in Step 1 and take out the data in bucket number V. Calculate the similarity or distance between x_i and the data points x_i in the bucket and return the nearest neighbor data $Q = \text{query}\,(x_i)$.

(3) Take data point x_j from Q, whose distance to x_i does not exceed L. Put x_j merged into B_n, that is, $B_n = B_n \cup x_j$, and remove it from the dataset U.

(4) Repeat Step 3 until the other data point x_j in Q reaches a certain threshold; the threshold can be computed as follows:

$$L \geq d\left(x_i, x_j\right) \tag{12}$$

(5) Repeat Steps 2-3 until the length of the U dataset is less than the threshold L. As shown in Algorithm 1, $\text{count}\,(U) \leq T$.

(6) The arithmetic mean values for the final k sets of samples are computed; then, we can obtain the clustering centers for all the categories in this way:

$$C_n = \frac{\text{Sum}\,(B_n)}{|B_n|}, \quad n = 1, 2, \ldots, k. \tag{13}$$

Therefore, based on the aforementioned steps, we will have two algorithms to choose from: "best" movement and "fast" movement [38]. For the "best" movement, we can use equation (13) and the k value in Algorithm 1 as the initial clustering center of the classical k-means input $\mathbf{KM}\,(k, c_i)$, and run the algorithm; the result is the final result. For "fast" movement, the divided categories can be regarded as approximate clustering results and directly used as the classification results. Because the initial clustering center is determined and the initial category is obtained, the result of the algorithm is more stable and accurate, and it requires a relatively short running time.

4. Experimental Results

First, we use the UCI https://archive-beta.ics.uci.edu/ml/datasets datasets [39] to verify the performance of the proposed algorithm, and we state the verification criteria. In addition, we use the Mall-Customers dataset https://www.kaggle.com/shwetabh123/mall-customers for the value range of the number of clusters k of the proposed LSH-k-means model. Our experimental results demonstrate the effectiveness and superiority of the proposed LSH-k-means. Then, we compare it with the actual business hall dataset and present an example to optimize the business hall operation.

4.1. Experimental Design. To verify the aforementioned points and evaluate the effectiveness of the proposed LSH-k-means model, numerous experiments were conducted on the UCI datasets, which consist of Balance, Wine, Breast, Diabet, Iris, Hayes-roth, Tic-tac-toe, and Bupa. We followed the experiments conducted in a previous study [40]. We briefly review the existing baselines as follows:

(1) k-means [25] is derived from the classical k-means.

(2) Enhanced k-means [38] enhances the classical k-means algorithm. The initial cluster centers are determined in advance instead of random selection.

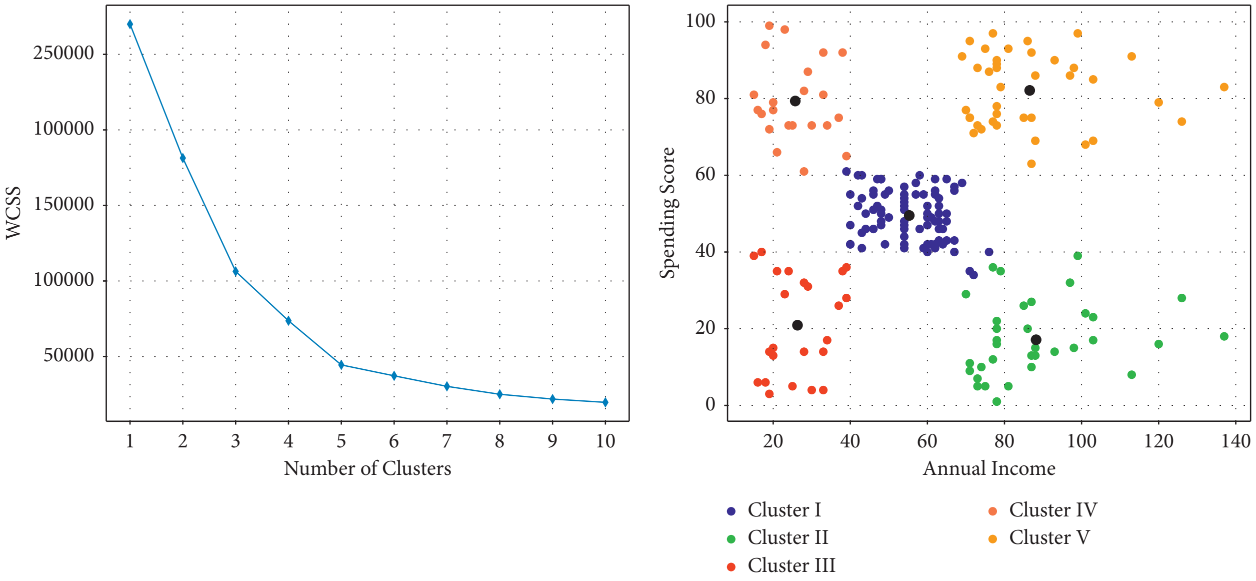
(3) The AC algorithm [41] for clustering can assume each sample as a pattern; by computing the similarity between patterns, the more similar patterns are grouped into one class, and the less similar patterns are classified into different classes. The difference between two patterns in AC clustering is usually measured by the distance function, including the Euclidean distance or Hamming distance. In the experiment, the AC algorithm is implemented by the KnowledgeMiner Software [41].

There are 200 samples in the Mall-Customers dataset. It includes gender, customer ID, age, annual income, and expenditure scores. In addition, it collects insights from the data and groups them according to their behaviors. The elbow method [42] is a well-known method for determining the optimal value of k. As shown in Figure 3(a), the optimum number of clusters of the Mall-Customers dataset is 5. According to Algorithm 1, we set the minimum number $T = 20$ and the maximum distance $L = 50.0$. Owing to the small amount of data, we set the number of buckets to 1. After 10 LSH-based initializations, we get the value of k between $[4 - 6]$. Figure 3(b) shows the results of LSH k-means clustering. The black dots represent the centroids.

There are 525 samples in the Balance dataset. For the classical k-means algorithm, the number of clustering categories that match the real categories is 271, and the matching rate is 51.62%. The corresponding values of the LSH k-means algorithm are 288 and 54.87%, respectively. Similarly, the results of the other UCI datasets are listed in Table 2. To determine whether there are significant differences between algorithms, we use the Wilcoxon signed-rank test [43]. It is a nonparametric statistical test. The Wilcoxon

```
**Required:** Training dataset $D \in \mathbb{R}^d$; the size of dataset $N$; the minimum number $T$ of data points in one cluster; the maximum distance
    in one cluster $L$; the closest set $B_n$, where $n \in \{1, 2, \ldots, k\}$ and $A$ is a two-dimensional array; $k$ is the number of clusters.
**Output:** The final clustering result.
Initialize LSH.
Index dataset $D$ via LSH.
Let $0 \leftarrow k$.
Let $U = D$.//copy $D$
**while** count$(U) \leq T$ **do**
    $k \leftarrow k + 1$.
    Randomly select one point $x_i$ from dataset $U$.
    $B_n = B_n \cup x_i$.
    $x_i$ is removed from $U$.
    $Q =$ query $(x_i)$.
    Let $0 \leftarrow q$
    **for** $q \leq$ count$(Q)$ **do**
        $q \leftarrow q + 1$.
        **if** $L \geq d(Q_q)$ **then**
        break;
        **end if**
        **if** $Q_q \in U$ **then**
        $Q_q$ removed from $U$.
        $B_k = B_k \cup Q_q$.
        **end if**
    **end for**
**end while**
Compute the centeroids $c_i$ via $B_n$.
Function **KM**$(k, c_i)$
The final clustering result. $C = \{C_1, C_2, \ldots, C_k\}$.
```

Algorithm 1: LSH-based k-means.



(a)

(b)

Figure 3: Results on the Mall-Customers dataset. (a) Elbow point graph, (b) Results of LSH k-means clustering.

test has been widely used in many fields, especially in algorithm comparison and analysis [40]. It is expressed as follows:

$$R^+ = \sum_{d_j>0} \operatorname{rank}(d_j) + \frac{1}{2} \sum_{d_j=0} \operatorname{rank}(d_j),$$
$$R^- = \sum_{d_j<0} \operatorname{rank}(d_j) + \frac{1}{2} \sum_{d_j=0} \operatorname{rank}(d_j), \tag{14}$$

where d_j is the difference in clustering performance between the two algorithms on the j-th dataset, and the absolute values of their difference are arranged in the ascending order. If the rank is the same, we take the average value. R^+ implies that the sum of ranks for the algorithm is better than the other, and R^- implies the opposite.

The calculations for the eight aforementioned datasets are presented below.

$$R^+ = 7 + 8 + 2 + 5 + 6 = 28,$$
$$R^- = 4 + 1 + 3 = 8. \tag{15}$$

Let $T = \min(R^+, R^-) = 8$; we get $T = 8$. According to the critical value table of the Wilcoxon test, we can judge that the difference between algorithms is significant under the condition $a = 0.05$. Furthermore, as shown in Table 2, there are five datasets for the LSH-based k-means, which is hence better than the enhanced k-means; thus, in terms of quantity, the LSH-based k-means algorithm outperforms the enhanced k-means algorithm. Therefore, we can judge that the efficiency of LSH-based k-means is significant.

In addition to comparison with actual categories, we further distinguish the clustering effects of k-means clustering and the AC algorithm. A tight and separative indicator is used to evaluate the clustering results [44], which is defined as follows:

$$S(U,k) = \frac{1/n \sum_{i=1}^{n} \sum_{j=1}^{k} (u_{ij}) \left| x_i - c_j \right|^2}{\min\limits_{p,q=1,2,\ldots,k} \left| c_p - c_q \right|}, \tag{16}$$

where c_p, c_q, and c_j denote the cluster centers, x_i is any sample in the dataset, k is the number of clusters, and U is the sample set. The Xie–Beni (XB) index [45, 46] is based on intracluster and intercluster distances; it is formulated in terms of the cluster compactness and separation between the clusters. We use the XB index for the evaluation of the cluster effects, and it is defined as follows:

$$\mathrm{XB} = \max\{S(U,k)\}, \quad k = 2, 3, \ldots, n-1, \tag{17}$$

where $S(U,k)$ is the ratio of the average distance between data objects and their corresponding clustering centers to the minimum distance of the cluster centers. The smaller the value of $S(U,k)$, the higher is the clustering quality. The results are summarized in Table 3.

From the XB value calculated in Table 3, we can conclude that the difference between the algorithms is significant. The XB value of the AC algorithm is the largest, while that of the LSH-based k-means algorithm is the smallest, which implies that the LSH k-means algorithm outperforms the other algorithms in the experiment. Thus, the experimental results verify the effectiveness and superiority of the proposed method. Therefore, it can finally be applied to the empirical analysis. In the next section, we describe the application of LSH k-means to business hall analysis.

4.2. Business Hall Analysis. In reality, business hall resource allocation may be unreasonable. For example, some business halls may be busy, while others may be idle. This may be caused by overlapping user coverage in different business halls, unreasonable location of the business halls, and a large proportion of low-value businesses. In this section, we experimentally verify the effectiveness of our index system and analyze the results of the proposed LSH-k-means model.

4.2.1. Business Hall Clustering. When the index system is established as described in Section 3.1, we get the characteristic information of the business hall. After data preprocessing, the number of clusters is determined subjectively. Consider the load intensity and time change information for the business hall. The load intensity can be categorized into High, Medium, and Low, and the load trend can be categorized into No change, Slow grouth, Fast grouth. Thus, a nine-square grid (Figure 4(b)) map can be obtained. At the same time, by referring to the knowledge of field experts, the number of clusters can be defined as 9 for the subjective clustering methods. After the clusters are determined, the extracted feature indicators can be taken as the input, and the clustering model is implemented. For the LSH k-means algorithm, the distance parameter was selected as the Euclidean distance, the maximum number of iterations was set to 500, the number of seeds was set to 10, and the number of the clusters was set to 9. Then, the outcomes were obtained, as shown in Table 4 and Figure 4(a).

Meanwhile, in the case of different predetermined cluster numbers for the subjective clustering methods, the AC algorithm determined the clustering number automatically, which was computed on the basis of the similarity between the samples. Here, the similarity was set at 95%, and the algorithm was implemented at the same time. Thus, the result was exactly consistent with that of the LSH k-means algorithm. The details are presented in Table 5. For example, the first and sixth samples, the second sample, and the third sample were clustered into the same category.

The final classification results obtained from the model can provide the load grades and decision-making suggestions, which can serve as a basis for site planning optimization of the business halls. In addition, the results of the two algorithms were consistent, which indicate that the LSH k-means algorithm is effective and a stable result was obtained. Accordingly, further optimization action can be implemented.

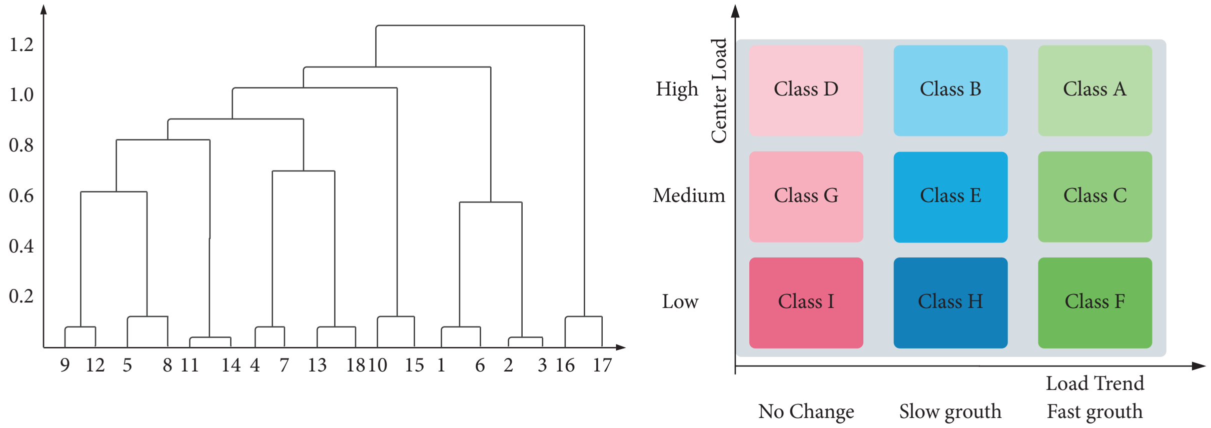
4.2.2. Optimization Analysis. As shown in Figure 5(a), the 16th and 17th business halls are both in Class I. This category indicates that the current load is Low, and the load trend remains unchanged, implying that the business hall

Table 2: Experimental results on UCI datasets.

Datasets	k-means [25] (%)	Enhanced k-means [38] (%)	LSH k-means (this work) (%)	Difference (%)	Rank
Balance	51.62	53.16	54.87	1.71	7
Wine	57.30	65.18	70.22	5.04	8
Breast	93.85	93.99	94.43	0.44	2
Diabet	66.28	67.76	66.72	−1.04	4
Iris	89.33	90.67	89.77	−0.90	3
Hayes-roth	34.09	35.19	35.19	0.00	1
Tic-tac-toe	54.02	56.01	57.17	1.16	5
Bupa	44.64	44.93	46.43	1.50	6

Table 3: XB evaluation value.

Algorithms	Iterations similarity	Category			XB
		C_1	C_2	C_2	
k-means [25]	8	27	102	49	0.7067
Enhanced k-means [38]	4	47	69	62	0.7045
AC algorithm [41]	95%	27	126	25	2.1694
LSH k-means (this work)	3	48	68	62	0.7033



(a)

(b)

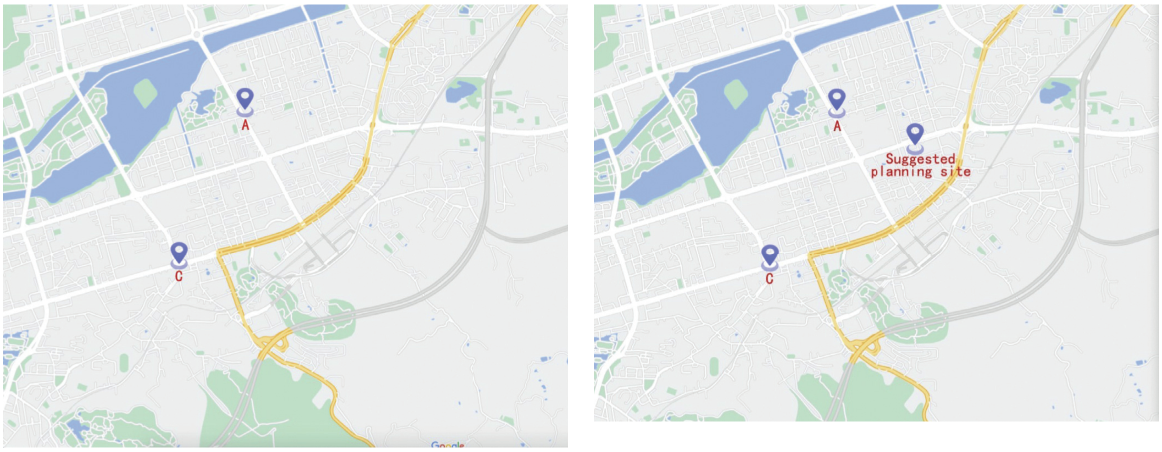
Figure 4: Results of LSH k-means clustering on business hall dataset. (a) Diagram of clustering results, (b) Nine-square grid map.

Table 4: Clustering result of LSH-k-means.

No.	Full data	1	2	3	4	5	6	7	8	9
Attributes						Cluster centroids				
S	0.55	0.25	0.60	0.20	0.10	0.96	0.50	0.95	0.50	0.90
Q	0.00	0.51	−0.51	0.01	−0.50	0.50	0.53	−0.53	0.01	0.01
V	0.62	0.75	0.50	0.45	0.40	0.80	0.60	0.80	0.65	0.70
FH	0.32	0.05	0.08	0.03	0.01	0.95	0.25	0.60	0.08	0.09
FL	0.35	0.85	0.10	0.80	0.95	0.02	0.25	0.10	0.06	0.02
RH	0.37	0.10	0.70	0.80	0.01	0.02	0.50	0.60	0.65	0.03
RL.	0.51	0.30	0.10	0.04	0.98	0.90	0.50	0.08	0.80	0.90

resources are redundant in this area. The business halls in this class are idle, and the site may be unreasonable. In addition, the merger of business halls, relocation, and reduction of resource input in this area should be considered.

By contrast, for Class A (the red part of Figure 5(b)), it can be seen that the current load and load trend are both High, which implies that the business volumes of the business halls are large, and the load trend change is still on the rise. Currently, the first and sixth business halls belong to this category. The future trend is still likely to be growing, and the business volumes will keep increasing. Therefore, this area is where more business hall resources need to be input, and the optimal site planning of business halls should be considered accordingly. We can define the objective



(a)

(b)

Figure 5: Illustration of business hall location optimization. (a) Location of business halls in class *A*; (b) Suggested location for supplementary business hall.

Table 5: Clustering result of the AC algorithm.

Result	S	Q	V	FH	FL	RH	RL
{C1}	0.95	0.5	0.8	0.95	0.01	0.02	0.9
{C2}	0.89	0.01	0.69	0.89	0.02	0.02	0.89
{C2}	0.9	0	0.7	0.9	0.01	0.03	0.9
{C3}	0.94	−0.56	0.79	0.59	0.09	0.59	0.9
{C4}	0.5	0.5	0.6	0.25	0.25	0.5	0.5
{C1}	0.96	0.49	0.79	0.94	0.02	0.01	0.89
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
{C9}	0.09	−0.49	0.39	0.01	0.94	0.01	0.97
{C9}	0.1	−0.5	0.4	0	0.95	0	0.98
{C8}	0.59	−0.51	0.49	0.07	0.09	0.69	0.09

function of the optimal site for the input business hall resources as follows:

$$\mathrm{Obj}_{\min} = \frac{\sum_{k=1}^{n} F_k}{n} + \sum_{i=l} h_i d\left(v_i, x\right), \tag{18}$$

where F_k is the quantitative value of factors that affect the rationality of business hall location, n is the number of factors, $d\left(v_i, x\right)$ is the distance function, h_i is the weight, and x is the target point to be solved. Thus, according to the objective function and relevant coordinate information of key units in the area, we can compute the optimal planning location of the business hall using the optimization algorithm. Here, the optimal location of the business hall was computed as [100.0644, 128.8199], and the optimal solution was 527.0368. The details are shown in Figure 5.

The 9th and 12th business halls belong to the Class *E*, which shows that the current load and load trends are both normal, and the status is stable. Therefore, the business halls in this class are not the current focus of optimization. In addition, the other classes are similar to this category, which is also not the current focus. The main objects are Class *I* and Class *A*, that is, excessive or insufficient business hall resources are mainly concentrated in these two classes, which are the focus of our optimization analysis.

5. Conclusion

Excessive or insufficient business hall resources may result in unreasonable resource allocation, which adversely affects the value of an entity business hall. Therefore, proper characteristic parameters are the key factors for analyzing the business hall, which strongly affect the final analysis results. According to the time change and load trend, multiple variables such as average load rate, actual load trend, and high-frequency load are extracted as the characteristic indexes of the business hall. In this study, a characteristic analysis method for the economic operation of a business hall was developed, and the specific calculation process was presented; accordingly, the feature engineering was established. Moreover, based on the load intensity and time change information of business halls, we built an index system and performed further optimization analysis. The key characteristic indicators extracted were the average waiting time, ticket handling time, and business type, and a model for evaluating business hall efficiency was established. The model obtained the load grading of each business hall by the relevant variable input, which provided a basis for optimal site planning of the business halls.

An empirical study showed that the LSH-*k*-means clustering method outperforms the direct prediction method, provides expected analysis results and decision optimization suggestions for business halls, and serves as a basis for the optimal layout of business halls. In addition, by considering the load intensity and time change information, the cluster number was determined according to the characteristic analysis results, with a certain theoretical and practical significance. In the future, we will explore and develop a general method to automatically determine the parameters and use it in practical applications.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. L. Brandeau and S. S. Chiu, "An overview of representative problems in location research," *Management Science*, vol. 35, no. 6, pp. 645–674, 1989.

[2] W. Shijun, H. Feilong, and J. Lili, "Locations and their determinants of large-scale commercial sites in changchun," *China*, vol. 70, no. 6, pp. 893–905, 2015.

[3] F. Goreaud and R. Pélissier, "On explicit formulas of edge effect correction for Ripley's K -function," *Journal of Vegetation Science*, vol. 10, no. 3, pp. 433–438, 1999.

[4] J. W. Cohen and O. J. Boxma, *Boundary Value Problems in Queueing System Analysis*, Elsevier, Amsterdam, Netherlands, 2000.

[5] D. R. Anderson, D. J. Sweeney, T. A. Williams, J. D. Camm, and J. J. Cochran, *An Introduction to Management Science: Quantitative Approach*, Cengage learning, Boston, MA, USA, 2018.

[6] G. Lin, X. Chen, and Y. Liang, "The location of retail stores and street centrality in guangzhou, China," *Applied Geography*, vol. 100, pp. 12–20, 2018.

[7] S. Kang, "Warehouse location choice: a case study in los angeles, ca," *Journal of Transport Geography*, vol. 88, Article ID 102297, 2020.

[8] X. Hui, "Optimization model and algorithm research of business hall service channel power," *Electronic Test*, vol. 1, pp. 20-21, 2014.

[9] X. T. Yan, Y. Zhang, Y. J. Huang, and W. U. Ying-Chun, "Management application and service data integration of the electricity supply business hall," *Power Demand Side Management*, vol. 37, pp. 50–52, 2017.

[10] B. Baraldi, "A survey of fuzzy clustering algorithms for pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, vol. 29, 1999.

[11] J. Lu, W. Gang, W. Deng, and K. Jia, "Reconstruction-based metric learning for unconstrained face verification," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 79–89, 2014.

[12] M. Yambal and H. Gupta, "Image segmentation using fuzzy c means clustering: a survey," in *Proceedings of the 2010 6th International Conference on Emerging Technologies (ICET)*, Islamabad, Pakistan, October 2010.

[13] H. Zhang, T. W. Chow, and Q. M. Wu, "Organizing books and authors by multilayer som," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 27, no. 12, p. 2537, 2015.

[14] R. C. De Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Information Sciences*, vol. 324, pp. 126–145, 2015.

[15] C.-W. Tsai, W.-L. Chen, and M.-C. Chiang, "A modified multiobjective ea-based clustering algorithm with automatic determination of the number of clusters," in *Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2833–2838, IEEE, Seoul, Korea, October 2012.

[16] C. Hennig and T. F. Liao, "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 62, no. 3, pp. 309–369, 2013.

[17] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," *Journal of Computational & Graphical Statistics*, vol. 29, no. 1, pp. 162–173, 2020.

[18] S. Vassilvitskii and D. Arthur, "K-means++: the advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, New Orleans LA, USA, January 2006.

[19] M. A. Masud, J. Z. Huang, C. Wei et al., "I-nice: a new approach for identifying the number of clusters and initial cluster centres," *Information Sciences*, vol. 466, pp. 129–151, 2018.

[20] M. Erisoglu, N. Calis, and S. Sakallioglu, "A new algorithm for initial cluster centers in k-means algorithm," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1701–1705, 2011.

[21] X. Shen, W. Liu, I. Tsang, F. Shen, and Q.-S. Sun, "Compressed k-means for large-scale clustering," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco CA USA, February 2017.

[22] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Proceedings of the 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 459–468, IEEE, Berkeley, CA, USA, October 2006.

[23] H. Koga, T. Ishibashi, and T. Watanabe, "Fast agglomerative hierarchical clustering algorithm using locality-sensitive hashing," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 25–53, 2007.

[24] D. Gorisse, M. Cord, and F. Precioso, "Locality-sensitive hashing for chi2 distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 402–409, 2011.

[25] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, July 1967.

[26] X. Wu, V. Kumar, J. Ross Quinlan et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[27] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[28] X. Peng, I. W. Tsang, J. T. Zhou, and H. Zhu, "K-meansnet: when k-means meets differentiable programming," https://arxiv.org/abs/1808.07292.

[29] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3.

[31] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer Science & Business Media, Berlin, Germany, 2013.

[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*, vol. 96, pp. 226–231, 1996.

[33] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[34] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the 13th Annual ACM Symposium on Theory of Computing*, pp. 604–613, Dallas TX USA, May 1998.

[35] A. Gionis, P. Indyk, and R. Motwani, “Similarity search in high dimensions via hashing,” *Vldb*, vol. 99, pp. 518–529, 1999.

[36] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, “Scalable k-means++,” https://arxiv.org/abs/1203.6402.

[37] W. Hu, Y. Fan, J. Xing, L. Sun, Z. Cai, and S. Maybank, “Deep constrained siamese hash coding network and load-balanced locality-sensitive hashing for near duplicate image detection,” *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4452–4464, 2018.

[38] J. Chen, D. Zhang, and Y. Nanehkaran, “Research of power load prediction based on boost clustering,” *Soft Computing*, vol. 25, no. 8, pp. 6401–6413, 2021.

[39] D. J. Newman, “Uci repository of machine learning database,” http://www.ics.uci.edu/mlearn/MLRepository.html.

[40] J. Chen, D. Zhang, and Y. A. Nanehkaran, “An economic operation analysis method of transformer based on clustering,” *IEEE Access*, vol. 7, pp. 127956–127966, 2019.

[41] F. Lemke and J.-A. Müller, “Self-organising data mining,” *Systems Analysis Modelling Simulation*, vol. 43, no. 2, pp. 231–240, 2003.

[42] R. L. Thorndike, “Who belongs in the family?” *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[43] L. Deng, J. Pei, J. Ma, and D. L. Lee, “A rank sum test method for informative gene discovery,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 410–419, Seattle, WA, USA, August 2004.

[44] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107–145, 2001.

[45] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[46] M. Singh, R. Bhattacharjee, N. Sharma, and A. Verma, “An improved xie-beni index for cluster validity measure,” in *Proceedings of the 2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–5, IEEE, Shimla, India, December 2017.