

Article

Gms-Afkmc2: A New Customer Segmentation Framework Based on the Gaussian Mixture Model and ASSUMPTION-FREE K-MC2

Liqun Xiao  and Jiashu Zhang *

Sichuan Province Key Lab of Signal and Information Processing, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China; xiaoliqun@my.swjtu.edu.cn

* Correspondence: jszhang@home.swjtu.edu.cn

Abstract: In this paper, the impact of initial clusters on the stability of customer segmentation methods based on K-means is investigated. We propose a novel customer segmentation framework, Gms-Afkmc2, based on the Gaussian mixture model and ASSUMPTION-FREE K-MC2, a better cluster-based K-means method, to obtain greater customer segmentation by generating better initial clusters. Firstly, a dataset sampling method based on the Gaussian mixture model is designed to generate a sample dataset of custom size. Secondly, a data clustering approach based on ASSUMPTION-FREE K-MC2 is presented to produce initialized clusters with the proposed dataset. Thirdly, the enhanced ASSUMPTION-FREE K-MC2 is utilized to obtain the final customer segmentation on the original dataset with the initialized clusters from the previous stage. In addition, we conduct a series of experiments, and the result shows the effectiveness of Gms-Afkmc2.

Keywords: customer segmentation; Gaussian mixture model; K-MC2; K-means



Citation: Xiao, L.; Zhang, J.

Gms-Afkmc2: A New Customer Segmentation Framework Based on the Gaussian Mixture Model and ASSUMPTION-FREE K-MC2.

Electronics **2024**, *13*, 3523. <https://doi.org/10.3390/electronics13173523>

Academic Editor: Domenico Ursino

Received: 6 August 2024

Revised: 28 August 2024

Accepted: 29 August 2024

Published: 5 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Customer segmentation is the process of dividing a customer base into groups of individuals who have similar characteristics or behaviors. This allows businesses to tailor their marketing efforts and product offerings to specific segments, increasing the effectiveness of their strategies [1]. The methods based on K-means and the RFM model in [2–4] are a typical class of the customer segmentation task, which constructs a dataset of customer segmentation with the attributes of Recency, Frequency and Monetary and then generates a customer segmentation result by K-means-based clustering methods. Compared with the traditional customer segmentation methods, this could deal with larger data due to its automatic optimization feature. In addition, those algorithms based on K-means have advantages in terms of algorithm complexity and algorithm time consumption compared to current clustering algorithms such as spectral cluster [5], DBSCAN [6] and deep learning [7]. For example, using spectral clustering methods, it could be time-consuming to compute the Laplace matrix and feature vectors for large-scale datasets. DBSCAN-based clustering algorithms are more sensitive to the choice of neighborhood radius and minimum number of samples. And it is difficult to obtain the optimal domain radius and minimum number of samples. Moreover, the DBSCAN-based clustering algorithm may need more time to compute the domain of each point on large-scale datasets. However, since the quality of the initial clustering centers has a large impact on the performance of K-means, how to enhance the quality of initial clusters is an interesting topic.

Lloyd et al. in [8] proposed a Least Squares Quantization in 1982, usually referred to simply as K-means. The method is to find the best clusters to fix a preassigned set of quanta. Inaba et al. [9] were the first to give an exact algorithm for the K-means problem with the running time of ($O(n^{kd})$) by minimizing the Sum of Squared Errors within clusters. Those

methods are flawed in terms of the randomized seeding technique, which causes the low quality of the initial cluster and then makes the K-means unstable.

David Arthur proposed K-means++ in [10], which has been widely adopted. It improves the selection of initial clusters by using a specific probability distribution and proves its complexity is more competitive than that of K-means. In addition, Makarychev et al. provided theoretical analysis and evidence of the effectiveness of the K-means++ algorithm in [11]. But K-means++ seeding does not scale well to massive datasets as it is inherently sequential and requires a full pass through the data.

Bachem et al. in [12] proposed ASSUMPTION-FREE K-MC2, the seeding technique of which is enhanced by a Markov chain-based initial cluster center generation algorithm. Compared to K-means++, it exhibits stronger competitiveness in terms of complexity and the quality of initial clusters, particularly performing better on large-scale datasets.

Customer segmentation methods based on K-means and K-means++ have been widely investigated. Maryani et al. introduced a customer analysis algorithm based on the K-means algorithm and the RFM model and successfully applied it to 82,648 transaction records of 102 customers in a specific company [13]. Similarly, Dzulhaq et al. also proposed an improved customer segmentation method based on K-means and RFM, by optimizing the algorithm's performance through techniques such as elbow plots [14]. Other relative customer segmentation methods can be seen in the literature [15,16]. Anitha et al. presented a customer segmentation method in [15] through Silhouette analysis and validated its effectiveness. The aforementioned studies only utilized K-means-based clustering techniques and had relatively limited customer scenarios. Khalili-Damghani et al. introduced a practical composite customer analysis algorithm [16] based on K-means, aiming to meet a broader range of application scenarios. This algorithm, based on the K-means clustering algorithm, effectively integrates decision tree technology and rule extraction techniques.

However, the above-mentioned references mainly discuss the issue of improving the quality of customer segmentation, which, so far, has not been solved. Specifically, the existing customer segmentation methods based on K-means still have the following challenges. The quality of the initial clusters is still unstable, which may potentially affect the customer segmentation results. In addition, the current methods may pose a challenge to the performance and time cost when the size of the customer segmentation dataset is more large and more complicated. Meanwhile, the topic motivates this work. This paper proposes Gms-Afkmc2, which is more stable in performance and consumes less time than the existing customer segmentation methods based on K-means by generating better initial clustering centers. In general, the main contributions are as follows. First, we propose a novel customer segmentation framework (Gms-Afkmc2) based on the Gaussian mixture model, Olivier Bachem's K-means algorithm and the RFM model. It not only generates high-quality initial clusters but also reduces the following iteration costs. Second, we analyze the impact of the parameter of sampling rate on the subsequent number of iterations and the quality of customer segmentation. And then, we obtain the optimal sampling rate for Gms-Afkmc2 in the dataset. Third, Gms-Afkmc2 shows the ability of customer segmentation in a comparison with the methods of K-means and K-means++. The proposed Gms-Afkmc2 not only has the advantage of handling large-scale customer data but also achieves better clustering results. It helps enterprises grasp customers' consumption patterns and ultimately enhances the competitiveness of their products.

The rest of this paper is organized as follows. In Section 2, the related theory is described. A novel Gms-Afkmc2 is designed in Section 3. Section 4 analyzes the efficiency of the proposed Gms-Afkmc2 through a series of comparative experiments on public datasets and details the application of customer segmentation results.

2. Related Work

2.1. RFM Model

The RFM model is a classic customer segmentation method. In this model, R (Recency) represents the time period since the customer's most recent purchase, F (Frequency)

represents the number of purchases made by the customer within a certain period of time, and M (Monetary) represents the amount spent by the customer within a certain period of time. Yao Zhang et al. proposed the RFMC model by using Clumpiness [17]. Heldt et al. proposed the RFM/P model based on the RFM model, which further enhances the performance of the RFM model by adding the new attribute of product price [18].

2.2. K-Means

K-means has been widely applied in customer segmentation for its high performance and lower complexity. There are two main types, Lloyd-based K-means [8] and Elkan-based K-means [19]. Lloyd-based K-means divides all elements of the dataset into K clusters so that the distance between all points in each cluster and the cluster center is the shortest, which can be represented in Equation (1).

$$\min \left\{ \sum_{j=1}^k Var^{\alpha}(S_j) \mid k - clustering(S_1, \dots, S_k) \in S \right\} \quad (1)$$

where S is a dataset and S_j represents the clustering results of dataset. Var^{α} refers to variance-based intra-cluster criterion, which is presented in Equation (2).

$$Var^{\alpha}(S) = |S|^{\alpha-1} \sum_{X_i \in S} \|X_i - \tilde{x}(s)\|^2 \quad (2)$$

where $\tilde{x}(s)$ is the centroid of S , which is defined by Equation (3).

$$\tilde{x}(s) = \frac{1}{|s|} \sum_{X_i \in S} X_i \quad (3)$$

The main process of the K-means algorithm is as follows. Firstly, we set the number of clusters, denoted as K , and generate the initial clusters of K , and then we assign the class of each point by finding the nearest center. Then, we optimize the clusters by calculating the distance between clusters and other elements of the dataset and reassign the classes of all points based on new centers. Finally, we iterate the second step until the termination condition. The termination condition is usually set to a specific number of running K-means, or the cluster results are sufficiently stable. The main factors affecting the performance of the K-means include the quality of initial clusters, optimization function and distance metrics.

However, Lloyd-based K-means calculates the distance from all points to every center in every iteration, which could cost a lot of time. Elkan-based K-means accelerates K-means by using the Triangle Inequality [19]. Specially, Elkan utilizes Theorem 1, when $\frac{1}{2}d(c, \hat{c}) \geq d(x, c)$ is true, to avoid calculating $d(x, c)$. And Elkan-based K-means applies Theorem 2, when it is supposed that in the previous iteration a lower bound $\hat{l}$ is known such that $d(x, \hat{b}) \geq \hat{l}$ is true, to calculate $d(x, \hat{b})$ by $\max\{0, 1 - d(b, \hat{b})\}$ faster.

Theorem 1. Let x be a point and let c be a center. If $d(b, c) \geq 2d(x, b)$, then $d(x, c) \geq d(x, b)$.

Theorem 2. Let x be a point and let b and c be centers. Then, $d(x, c) \geq \max\{0, d(x, b) - d(b, c)\}$.

In summary, the RFM model is widely and successfully applied to customer segmentation. Inspired by this, we propose a Gms-Afkmc2 framework for customer segmentation based on the RFM model. In addition, since Elkan-based K-means [19] is more efficient than Lloyd-based K-means in terms of solving clusters, the proposed Gms-Afkmc2 adopts Elkan-based K-means to optimize clusters.

3. The Proposed Gms-Afkmc2

In this section, we will detail the implementation of the proposed customer segmentation framework Gms-Afkmc2. Firstly, the proposed dataset sampling method Gms is

introduced in Section 3.1. And then, the details of Gms-Afkmc2 are presented on the basis of Gms in Section 3.2.

3.1. Gaussian Mixture Model Dataset Sampling Algorithm (Gms)

The Gaussian mixture model is a powerful non-linear model for accurately fitting various data distributions. Vemuri et al. proposed a Bayesian sampling framework-based Gaussian mixture model in [20]. In this work, a generalized Gaussian mixture model was used to learn the data distribution. Feng Wang et al. [21] used the Gaussian mixture model to obtain the existing data distribution for a more accurate prediction. Inspired by those works, the Gaussian mixture model in this paper is utilized to fit the dataset distribution.

The proposed Gms method ensures similar distributions on each relative data component between the subset and the original dataset. The Gaussian distribution function corresponding to each component is represented via Equation (4).

$$g(x | \mu_j) = \frac{1}{(2\pi)^{1/2} |\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_j) \Sigma_j^{-1} (x - \mu_j) \right\} \quad (4)$$

where $g(x | \mu_j)$ denotes the probability of x on the component μ_j . And μ_j, Σ_j are the expectation vector and covariance matrices on the component j , respectively. And the Gaussian mixture model, representing the final probability distribution in the subsampled dataset, is a combination of the independent Gaussian distribution functions. The distribution function $p\theta(x)$ of the Gaussian mixture model is shown in Equation (5).

$$p\theta(x) = \sum_{j=1}^J \omega_j g(x | \mu_j, \Sigma_j) \quad (5)$$

where θ is a manual parameter and ω_j represents the weights of the Gaussian function of the i th component, respectively. The Gaussian mixture model $p\theta(x)$ of the data in this paper is computed by maximizing the log-likelihood $\mathbb{E}_{q(x)} \log p\theta(x)$ in Equation (6) [22,23].

$$\mathbb{E}_{q(x)} \log p\theta(x) = \mathbb{E}_{q(x)} \left[\mathbb{E}_{q(z|x)} \log \frac{p\theta(x | \mu)}{q(z | x)} \right] + KL[q(z | x) \| p\theta(z | x)] \quad (6)$$

where $q(x)$ is the variational data distribution and $\mathbb{E}_{q(x)}[\cdot]$ is the expectation of data, which is estimated with data samples from $q(x)$. KL is a function of the Kullback–Leibler divergence in Equation (7), which is used to measure the difference between two probability distributions.

$$KL[q(z | x) \| p\theta(z | x)] = \sum_x p\theta(z | x) \log \left(\frac{p\theta(z | x)}{q(z | x)} \right) \quad (7)$$

In order to obtain optimal $p\theta(x)$ in Equation (6), the approximate posterior $q(z | x)$ needs to be iteratively updated with Equations (8) and (9).

$$q(z | x) = p\theta(z | x) = \frac{\mathbb{E}_{q(x)}[q(z | x)]g(x | \mu_z, \Sigma_z)}{\sum_j^K \omega_j g(x | \mu_z, \Sigma_z)} \quad (8)$$

$$KL[q(z | x) \| p\theta(z | x)] = \sum_x p\theta(z | x) \log \left(\frac{p\theta(z | x)}{q(z | x)} \right) \quad (9)$$

The data sampling algorithm based on the Gaussian model designed in this paper has three main steps. In the first, we configure the size of the goal subset and the corresponding weights of each component in the dataset. In the second, we train the Gaussian mixture model by the original dataset. Moreover, we generate a new sample dataset with the trained Gaussian mixture model. The parameter W , the weight of each component in the

dataset, also refers to work [15]. The specific implementation of the algorithm is shown in Algorithm 1.

Algorithm 1 Gaussian mixture model dataset sampling (*Gms*)

Input: DS_{raw} (Original dataset), ω (Weight of each comonent), k (Num of cluster), n_{sample} (Num of subset)
Output: DS_{sample} (Subsample dataset)
 # Train Gaussian mixture model with DS_{raw}
 # The fit function is a python library function belonging to sklearn, a machine learning library
 $Model_{GMM} = GMM.fit(DS_{raw})$
 # Obtain the samples
 # The sample function is a python library function belonging to sklearn, a machine learning library
 $Sample = model_{GMM}.sample()$
 # Find the nodes of the dataset closest to the samples as DS_{sample}
for sample in samples
 for index < $len(DS_{norm})$
 $dis = norm((DS_{norm} - sample) \times \omega)$
 if $dis < dis_{min}$
 $dis_{min} = dis$
 $index_{min} = index$
 $DS_{sample} = DS_{sample} \cup DS_{raw}[index_{min}]$
return DS_{sample}

3.2. Assumption-Free K-MC2 (*Afkmc2*)

The *afkmc2*, based on a Markov chain and the Monte Carlo algorithm, is for generating initial clusters with a distribution of a dataset. Specifically, the distribution is shown in Equation (10), and the function for generating a Markov chain is presented in Equation (11).

$$q(x | c_1) = \frac{d(x, c_1)^2}{2 \sum_{x \in X} d(x, c_1)^2} + \frac{1}{2n} \quad (10)$$

$$\pi(x_{j-1}, y_j) = \min \left(\frac{e(y_j, c)^2}{e(x_{j-1}, c)^2}, 1 \right) \quad (11)$$

where $q(x | c_1)$ refers to the probability that x belongs to c_1 , $d(x, c_1)$ denotes the distance between x and c_1 and n is the number of datasets, with probability $\pi(x_{j-1}, y_j)$ in Equation (11). We then set state x_{j-1} to y_j and probability $1 - \pi$ to x_{j-1} . $e(x, c)$ can be denoted via Equation (12).

$$e(x, c) = \min_{c \in C} \|x - c\|_2^2 \quad (12)$$

Algorithm 2 details the proposed *Gms-Afkmc2*. The main process of *Gms-Afkmc2* consists of three steps. Firstly, a subset is generated by the proposed *Gms* from the raw dataset. Secondly, we utilize the proposed *Afkmc2* algorithm to make stable clusters from the previous subset. Thirdly, we make the final customer segmentations with *Afkmc2* and the initial centers on the original dataset.

Algorithm 2 Gms-Afkmc2

Input: DS_{raw} (Original dataset), ω (Weight of each component), k (Num of cluster), n_{sample} (Num of subset)
Output: CC_F (The final clusters)
Produce a subsample dataset DS_{sample} by the function Gms in Algorithm 1
 $DS_{sample} = Gms(DS_{raw}, \omega, n_{sample})$
Generate initial centers of subdataset CC_{SI}
 $c_1 = \text{Point uniformly sampled}(DS_{sample})$
for x in DS_{sample} **do**
 $q(x) = \text{Compute data distribution}(x, DS_{sample})$
 $CC_{SI} = \{c_1\}$
 for i in $range(2, k)$ **do**
 $x = \text{Point uniformly sampled}(DS_{sample})$
 $c_i = \text{comput}(x, DS_{sample})$
 $CC_{SI} = CC_{SI} \cup c_i$
Make initial clusters of dataset CC_{SF}
The AFkmc2 function is a python library function belonging to kmc2-0.1, a library proposed by Oliver Bachem
 $CC_{SF} = AFkmc2(DS_{sample}, CC_{SI})$
Obtain final segmentations CC_F
 $CC_F = AFkmc2(DS_{raw}, CC_{SF})$
return CC_F

4. Experiments**4.1. Dataset and Evaluation Metrics**

In this work, two public e-commerce consumption datasets are used to evaluate the performance of the proposed Gms-Afkmc2. These two datasets contain four attributes: Customer ID, Recently Purchased Time, Frequency of Purchase and Purchase Amount. Among them, Customer ID is used to uniquely identify the customer. And the terms Recently Purchased Time, Frequency of Purchase and Purchase Amount correspond to the Recency, Frequency and Monetary of the RFM model, respectively. The first dataset (named Dataset 1) is a raw online retail dataset named online retail II dataset from kaggle, which has 5878 pieces of customer data and consists of more than 1.04 million consumption records. And the steps of preprocessing are as follows. Firstly, the raw data are cleaned, de-duplicated and filled with missing values to ensure the accuracy and usability of the data. Secondly, we generate the final Dataset 1 based on the RFM model for the following segmentation tasks. The second dataset (named Dataset 2) is a public customer segmentation dataset from kaggle, which contains 10,000 pieces of customer data.

In order to measure the quality of segmentation results belonging to the proposed method and other comparison algorithms more comprehensively, three popular metrics for evaluating the quality of clusters are applied for assessment. These are the Sum of Squared Errors, Silhouette Coefficient [24] and Davies–Bouldin [25]. Among them, Sum of Squared Errors, named SSE in Equation (13), quantifies the variance within each cluster. The index of Sum of Squared Errors, where a smaller index indicates a higher quality of clusters, is calculated as the sum of the squared distances between each point and the centroid of the cluster it belongs to. It is easy to understand and compute, but it does not provide information about the separation between clusters. The Silhouette Coefficient, named SH in Equations (14) and (15), measures how similar an object is to its own cluster compared to other clusters, where a high value indicates that an object is well matched to its own cluster and poorly matched to neighboring clusters. The index of Davies–Bouldin, named DB in Equations (16) and (17), refers to the relation between distances of inner clusters and distances of external clusters. When the index is larger, the clusters are better. The Davies–Bouldin considers both within-cluster scatter and between-cluster separation

and can be used with any clustering algorithm, not just those based on centroids. But it may not be sensitive to the addition of new clusters if the new clusters are small.

$$SSE = \sum_{i=1}^n \sum_{c \in C} \|x_i - c\|_2^2 \quad (13)$$

$$SH = \frac{\min(d(i, C))_{C \neq A} - d(i, A)}{\max\{\min(d(i, C))_{C \neq A}, d(i, A)\}} \quad (14)$$

$$d(i, A) = \frac{1}{n} \sum_{j=1}^n \|x_i - x_j\|_2^2 \quad (15)$$

$$DBI = \frac{1}{n} \sum_{i=1}^k \max \left(\frac{S_i + S_j}{d(c_i, c_j)} \right)_{i \neq j} \quad (16)$$

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} d(x, C_i) \quad (17)$$

4.2. Experiment Analysis

The optimal number of clusters is crucial for the quality of clusters. The elbow, which measures the relationship between the number of clusters and the distance within clusters, is a popular approach for finding the optimal cluster number in clustering methods. In this work, two elbow diagrams from Dataset 1 and Dataset 2 are used to find the optimal clustering number, as shown in Figures 1 and 2. According to these two diagrams, the optimal number for the largest inflection point is $k = 4$.

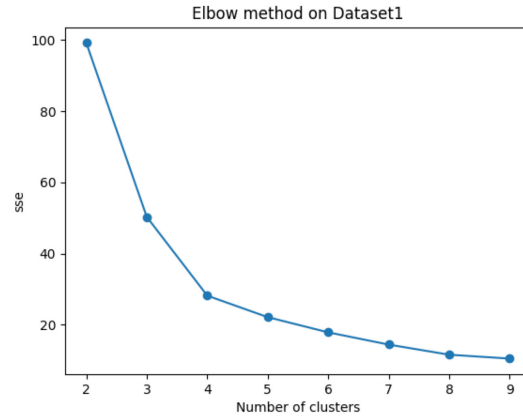


Figure 1. Elbow chart on Dataset 1.

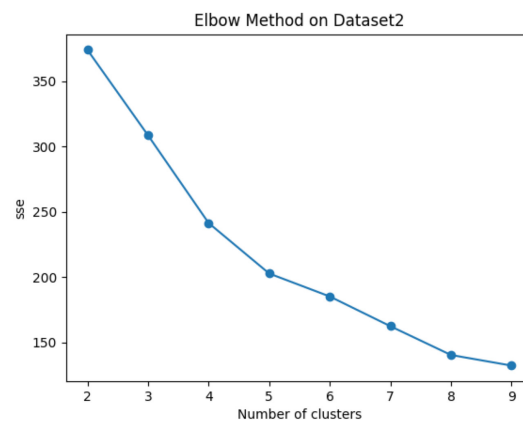


Figure 2. Elbow chart on Dataset 2.

In order to obtain high-quality initial centers with smaller subsets, we research the impact of sample rates on runtime, the number of iterations and the quality of clusters, shown in Tables 1 and 2. The two tables show trends that the sample rate is higher, the quality of clusters is better and the following iteration costs after seeding is fewer, but the cost time of the proposed method is higher. Therefore, we should find the optimal trade off among these factors. And when the sample rate is 0.5, Gms-Afkmc2 achieves a better balance.

Table 1. The index of different sample rates on Dataset 1.

Rate	Iterations	Runtime	SH ↑	DBI ↓
1	6	0.024	0.916	0.415
0.8	6	0.024	0.916	0.415
0.6	8	0.032	0.916	0.415
0.5	8	0.032	0.916	0.415
0.4	9	0.036	0.913	0.420
0.2	11	0.044	0.910	0.423
0.1	12	0.048	0.910	0.423
0.05	12	0.048	0.910	0.423

Table 2. The index of different sample rates on Dataset 2.

Rate	Iterations	Runtime	SH ↑	DBI ↓
1	9	0.027	0.355	1.065
0.8	10	0.030	0.355	1.065
0.6	12	0.036	0.355	1.065
0.5	13	0.039	0.355	1.065
0.4	13	0.039	0.354	1.066
0.2	13	0.039	0.352	1.067
0.1	15	0.045	0.352	1.067
0.05	15	0.045	0.350	1.068

In this paper, we compare our algorithm with K-means++ and K-means algorithms in terms of Sum of Squared Errors, Silhouette Coefficient and Davies–Bouldin metrics as well as the running time of the algorithm in Dataset 1 and Dataset 2, as shown in Tables 3 and 4. The optimal values of each index of these tables are in bold. According to the results in Tables 3 and 4, it is obvious that the proposed Gms-Afkmc2 is better than other algorithms in terms of the indicators of clustering, which further validates the effectiveness of the algorithm. Specifically, the results demonstrate the effectiveness of using Gaussian mixture models to fit the customer consumption dataset and the performance of the proposed sampling method because of its obtaining better initial clusters. Moreover, the results of this experiment prove that the performance of the proposed Afkmc2 based on Markov chains and Monte Carlo is better than that of the comparative methods such as K-means, K-means++, etc.

Table 3. The evaluation results on Dataset 1.

Method	SSE ↓	SH ↑	DBI ↑	Runtime
Gms-Afkmc2	28.19	0.92	0.42	0.018
K-means++	29.57	0.90	0.44	0.008
K-means	31.25	0.89	0.52	0.051

Table 4. The evaluation results on Dataset 2.

Method	SSE ↓	SH ↑	DBI ↑	Runtime
Gms-Afkmc2	241.423	0.355	1.065	0.026
K-means++	245.745	0.352	1.067	0.013
K-means	248.131	0.347	1.071	0.0750

4.3. Customer Segmentation and Application

Customer segmentation divides all customers into several clusters based on their consumption behavior, and customers within a cluster have similar consumption habits. According to the segmentation results, enterprises could further analyze the proportion of customers in each category and consumption habits and other specific information to master product marketing better. Therefore, they can carry out more targeted research to improve product quality and develop a more comprehensive sales strategy for enhancing the competitiveness of products in the market. In this work, the segmentation results on Dataset 1 and Dataset 2 will be analyzed to summarize the consumption capacity and habits of customers and then provide reference for enterprises to formulate marketing strategies.

The segmentation distribution of Gms-Afkmc2 is shown in Figures 3 and 4. We analyze the meaning of each cluster by mean and median statistics in terms of Recency, Frequency and Monetary in Tables 5 and 6. Cluster 3 and cluster 4 perform best in Recency, Frequency and Monetary indexes, these belong to high-quality customers. So, enterprises can adopt a relative marketing strategy to stabilize these two clusters. However, customers in cluster 1 perform the lowest in every index, so enterprises can appropriately reduce the marketing investment. In addition, the size of cluster 2 is large, accounting for 25.43%. And customers in this cluster have higher consumption potential than those in cluster 1. Therefore, this category of users is a potential growth point for enterprise revenue, and enterprises can moderately increase marketing efforts.

Table 5. The mean value in terms of RFM on Dataset 1.

	Rate	Recency ↓	Frequency ↑	Monetary ↑
Cluster 1	13.46%	597.76	30.24	657.41
Cluster 2	25.43%	348.1	56.62	1099.67
Cluster 3	60.85%	54.09	169.44	3451.62
Cluster 4	0.26%	3.86	4629.43	204,593.38

Table 6. The median value in terms of RFM on Dataset 1.

	Rate	Recency ↓	Frequency ↑	Monetary ↑
Cluster 1	13.46%	593	18	284.85
Cluster 2	25.43%	375	31.5	515.82
Cluster 3	60.85%	36	87	1428
Cluster 4	0.26%	2.5	3844	143,587.09

The segmentation distribution of Gms-Afkmc2 is shown in Figures 5 and 6. We analyze the meaning of each cluster using mean and median statistics in terms of Recency, Frequency and Monetary in Tables 7 and 8. The customers of cluster 4 are much better than other customers in Monetary and Frequency, which belong to high-quality customers. The enterprise should design a targeted strategy to maintain this type of customer. In addition, from the data in Tables 7 and 8, the Recency of customers in cluster 1 is significantly slower than that of other types, and the Frequency and Monetary of cluster 1 are higher than the average value. This indicates that the customers of cluster 1 are the most active among all clusters and are potential customers with high consumption capacity, so the company's marketing investment on these customers of cluster 1 may earn better benefits.

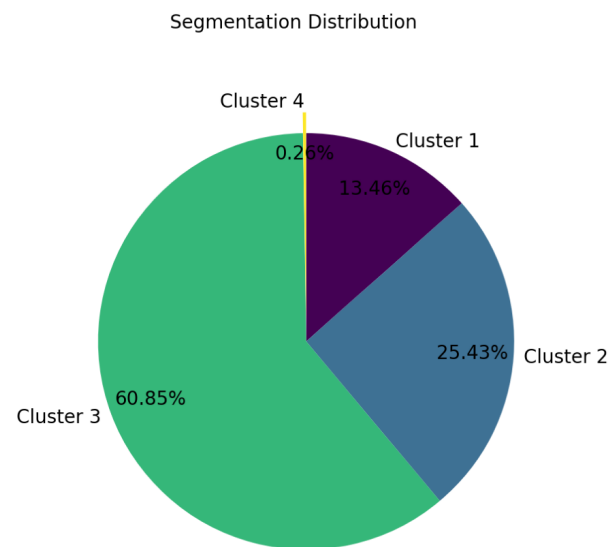


Figure 3. Distribution on Dataset 1.

With the high development and popularization of the Internet and the increasingly large amount of customer consumption data, customer segmentation plays a more important role in helping enterprises to grasp customer consumption behavior more accurately. Moreover, customer segmentation could improve the marketing efficiency and reduce the marketing cost of enterprises by analyzing the consumption behavior of customers. In addition, it can also support enterprises to better designate their marketing strategy, improve the competitiveness of the products of the company and enhance service for customers.

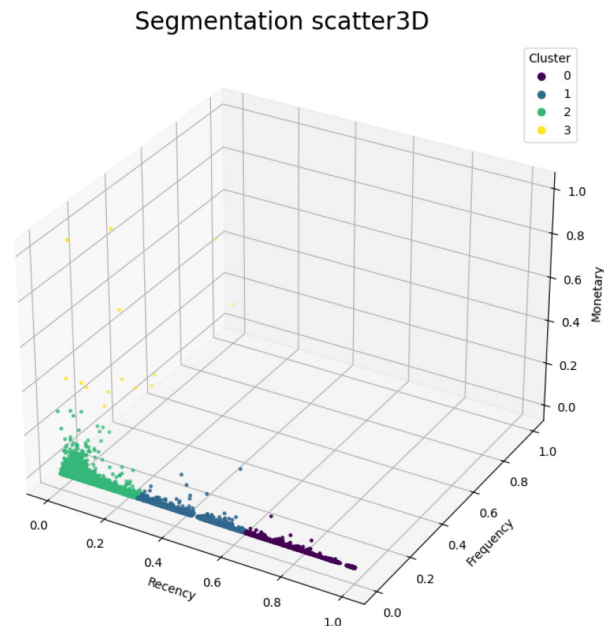


Figure 4. Scatter3D on Dataset 1.

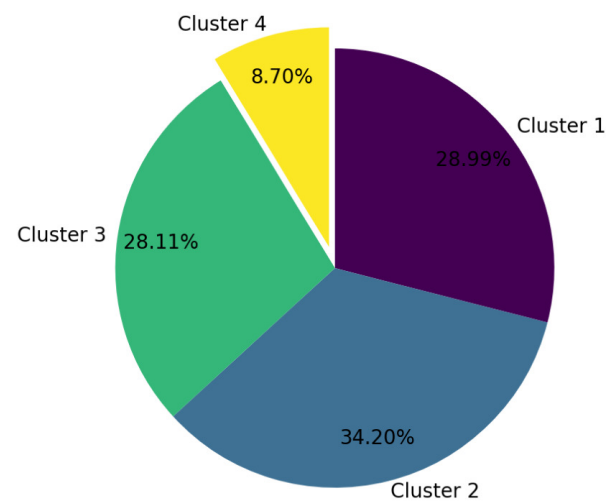
Table 7. The mean value in terms of RFM on Dataset 2.

	Rate	Recency ↓	Frequency ↑	Monetary ↑
Cluster 1	28.99%	101.69	93.57	202,034.79
Cluster 2	34.20%	314.14	76	169,933.59
Cluster 3	28.11%	210.99	73.24	156,509.03
Cluster 4	8.7%	203.36	334.14	756,760.6

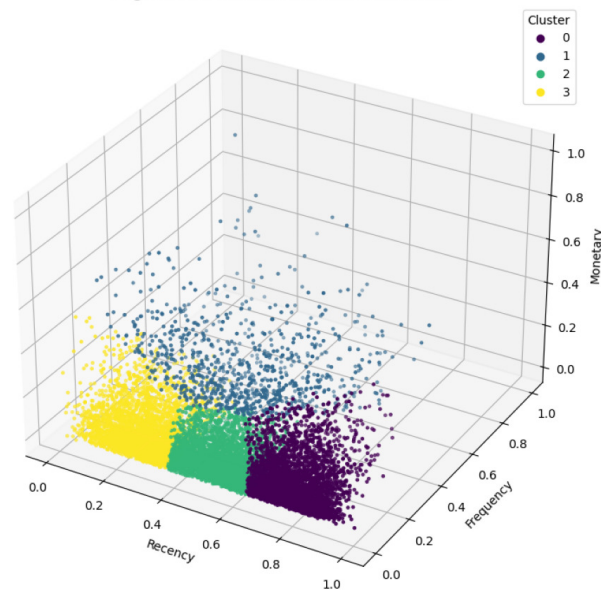
Table 8. The median value in terms of RFM on Dataset 2.

	Rate	Recency ↓	Frequency ↑	Monetary ↑
Cluster 1	28.99%	104	56	142,673.87
Cluster 2	34.20%	312	43.5	111,913.43
Cluster 3	28.11%	213	48	122,743.21
Cluster 4	8.7%	203	278	747,008.31

Segmentation Distribution on Dataset 2

**Figure 5.** Distribution on Dataset 2.

Segmentation scatter3D on Dataset 2

**Figure 6.** Scatter3D on Dataset 2.

5. Conclusions

In this paper, we proposed a novel customer segmentation framework, Gms-Afkmc2. Compared to clustering algorithms based on Spectral and DBSCAN methods, the proposed Gms-Afkmc2 has an advantage in processing large-scale customer data. Meanwhile, the results indicate that the proposed method is more competitive than other classical clustering methods based on K-means. However, there are still some limitations which need to be noted. The proposed Gms-Afkmc2 needs more time to generate the sample dataset and initial clustering centers. In addition, it suffers from a higher sample rate. From the results of the sample rate in Table 1, our Gms still needs to be optimized to obtain a better subdataset with a lower rate and reduced time cost. From Table 2, it can be seen that Gms-Afkmc2 costs twice as K-means++. Meanwhile, in the massive consumption data scenario, these customer segmentation methods, training the initial customer segmentation model via a high-quality sampling dataset and further fitting the final segmentation model with the original dataset and obtaining accurate customer segmentation results, are still valuable since they effectively reduce the iteration time of the customer segmentation model on the source data and make segmentation models more stable. So, there are several aspects which can be considered in the future. First, we could optimize the Gaussian mixture model or design a novel model to better learn the data distribution of the original dataset for obtaining a high-quality sampling dataset. Second, we could further improve the efficiency of the sample dataset algorithm to shorten the running time. Third, we could further study and improve the K-means algorithm so as to enhance the quality of customer segmentation results. In addition, e-WOM can be quantified through semantic analysis and then combined with the current work to obtain more comprehensive and accurate customer segmentation results, because many customers usually give online feedback on their own products, so enterprises have a huge amount of valuable e-WOM of their products [26–28].

Author Contributions: Conceptualization, L.X.; Data curation, L.X.; Formal analysis, L.X.; Funding acquisition, L.X.; Investigation, L.X.; Methodology, L.X.; Project administration, J.Z.; Resources, L.X.; Software, L.X.; Supervision, J.Z.; Validation, L.X. and J.Z.; Writing—original draft, L.X.; Writing—review and editing, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62071396 and the National Science Foundation of Sichuan Province under Grant No. 2022NSFSC0531.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guerola-Navarro, V.; Gil-Gomez, H.; Oltra-Badenes, R.; Soto-Acosta, P. Customer relationship management and its impact on entrepreneurial marketing: A literature review. *Int. Entrep. Manag. J.* **2022**, *20*, 507–547. [\[CrossRef\]](#)
2. Wan, S.; Chen, J.; Qi, Z.; Gan, W.; Tang, L. Fast RFM Model for Customer Segmentation. In Proceedings of the Companion of The Web Conference 2022, Virtual Event/Lyon, France, 25–29 April 2022; Laforest, F., Troncy, R., Simperl, E., Agarwal, D., Gionis, A., Herman, I., Médini, L., Eds.; ACM: New York, NY, USA, 2022; pp. 965–972.
3. Chen, Y.; Liu, L.; Zheng, D.; Li, B. Estimating travellers' value when purchasing auxiliary services in the airline industry based on the RFM model. *J. Retail. Consum. Serv.* **2023**, *74*, 103433. [\[CrossRef\]](#)
4. 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. [\[CrossRef\]](#)
5. Tang, C.; Li, Z.; Wang, J.; Liu, X.; Zhang, W.; Zhu, E. Unified one-step multi-view spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 6449–6460. [\[CrossRef\]](#)
6. Chen, Y.; Zhou, L.; Bouguila, N.; Wang, C.; Chen, Y.; Du, J. BLOCK-DBSCAN: Fast clustering for large scale data. *Pattern Recognit.* **2021**, *109*, 107624. [\[CrossRef\]](#)
7. Yan, J.; Cheng, Y.; Wang, Q.; Liu, L.; Zhang, W.; Jin, B. Transformer and Graph Convolution-Based Unsupervised Detection of Machine Anomalous Sound Under Domain Shifts. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2827–2842. [\[CrossRef\]](#)
8. Lloyd, S.P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–136. [\[CrossRef\]](#)

9. Inaba, M.; Katoh, N.; Imai, H. Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k -Clustering (Extended Abstract). In Proceedings of the Tenth Annual Symposium on Computational Geometry, New York, NY, USA, 6–8 June 1994; Mehlhorn, K., Ed.; ACM: New York, NY, USA, 1994; pp. 332–339.
10. Arthur, D.; Vassilvitskii, S. k -means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, LA, USA, 7–9 January 2007; Bansal, N., Pruhs, K., Stein, C., Eds.; SIAM: Philadelphia, PA, USA, 2007; pp. 1027–1035.
11. Makarychev, K.; Reddy, A.; Shan, L. Improved Guarantees for k -means++ and k -means++ Parallel. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020.
12. Bachem, O.; Lucic, M.; Hassani, H.; Krause, A. Fast and Provably Good Seedings for k -Means. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016.
13. Maryani, I.; Riana, D.; Astuti, R.D.; Ishaq, A.; Pratama, E.A. Customer segmentation based on RFM model and clustering techniques with K-means algorithm. In Proceedings of the 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 17–18 October 2018; pp. 1–6.
14. Dzulhaq, M.I.; Sari, K.W.; Ramdhan, S.; Tullah, R. Customer segmentation based on RFM value using K-means algorithm. In Proceedings of the 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 16–17 October 2019; pp. 1–7.
15. Anitha, P.; Patil, M.M. RFM model for customer purchase behavior using K-Means algorithm. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1785–1792. [[CrossRef](#)]
16. Khalili-Damghani, K.; Abdi, F.; Abolmakarem, S. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Appl. Soft Comput.* **2018**, *73*, 816–828. [[CrossRef](#)]
17. Zhang, Y.; Bradlow, E.T.; Small, D.S. Predicting customer value using clumpiness: From RFM to RFMC. *Mark. Sci.* **2015**, *34*, 195–208. [[CrossRef](#)]
18. Heldt, R.; Silveira, C.S.; Luce, F.B. Predicting customer value per product: From RFM to RFM/P. *J. Bus. Res.* **2021**, *127*, 444–453. [[CrossRef](#)]
19. Elkan, C. Using the Triangle Inequality to Accelerate k -Means. In Proceedings of the Machine Learning, Proceedings of the Twentieth International Conference ICML 2003, Washington, DC, USA, 21–24 August 2003; Fawcett, T., Mishra, N., Eds.; AAAI Press: Washington, DC, USA, 2003; pp. 147–153.
20. Vemuri, R.T.; Azam, M.; Bouguila, N.; Patterson, Z. A Bayesian sampling framework for asymmetric generalized Gaussian mixture models learning. *Neural Comput. Appl.* **2022**, *34*, 14123–14134. [[CrossRef](#)]
21. Wang, F.; Liao, F.; Li, Y.; Wang, H. A new prediction strategy for dynamic multi-objective optimization using Gaussian Mixture Model. *Inf. Sci.* **2021**, *580*, 331–351. [[CrossRef](#)]
22. Jin, P.; Huang, J.; Liu, F.; Wu, X.; Ge, S.; Song, G.; Clifton, D.; Chen, J. Expectation-maximization contrastive learning for compact video-and-language representations. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 30291–30306.
23. Cong, Y.; Li, S. Big Learning Expectation Maximization. *arXiv* **2023**, arXiv:2312.11926.
24. Bagirov, A.M.; Aliguliyev, R.M.; Sultanova, N. Finding compact and well-separated clusters: Clustering using silhouette coefficients. *Pattern Recognit.* **2023**, *135*, 109144. [[CrossRef](#)]
25. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. k -means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [[CrossRef](#)]
26. Narayanan, S.; Samuel, P.; Chacko, M. Product Pre-Launch Prediction From Resilient Distributed e-WOM Data. *IEEE Access* **2020**, *8*, 167887–167899. [[CrossRef](#)]
27. Anastasiei, B.; Dospinescu, N.; Dospinescu, O. Individual and Product-Related Antecedents of Electronic Word-of-Mouth. *arXiv* **2024**, arXiv:2403.14717.
28. Kim, W.; Nam, K.; Son, Y. Categorizing affective response of customer with novel explainable clustering algorithm: The case study of Amazon reviews. *Electron. Commer. Res. Appl.* **2023**, *58*, 101250. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.