International Journal of Smart Business and Technology
Vol.9, No.2 (2021), pp.49-64
http://dx.doi.org/10.21742/ijsbt.2021.9.2.04

Improved Multi-index Customer Segmentation Model Research

Wolfgang Bellotti[1], Daniela N. Davies[2] and Y. H. Wang[3]

[1]*University of Plymouth, Plymouth, UK*
[2,3]*University of Liverpool, Liverpool, UK*
[1]*wolfgang.bellotti@plymouth.ac.uk*, [3]*yh.wang@liverpool.ac.uk*

Abstract

Customer segmentation helps the company's strategy formulation and competitiveness enhancement. To better meet customer needs and preferences, companies must recognize the differences of customers and formulate precise marketing strategies. This article focuses on the current customer segmentation background and combines Data mining tools, proposed a multi-index customer segmentation model. Considering the micro and macro perspectives, the traditional indicators are refined, and new segmentation indicators are added. The indicators are weighted by the entropy method. To reduce the time complexity of clustering, factor analysis is used to reduce the data dimension. Finally, the improved K-means clustering algorithm is used to optimize the determination of the K value and the selection of the initial center point to determine the customer segmentation results. The empirical research results on the segmentation of a retailer's membership data show that the improved algorithm is superior to the classic customer segmentation method in terms of clustering compactness and feature division capabilities. With this, it can help companies to improve the level of customer relationship management and the quality of decision-making.

Keywords: *Data mining, K-means, Customer segmentation model, RFMPQ model*

1. Introduction

Nowadays, market competition in the retail industry has become increasingly fierce, which has brought tremendous pressure to companies, forcing them to understand customer needs more effectively to gain or maintain a competitive advantage in the industry. To improve customer loyalty and satisfaction, provide individuality Customized services and the formulation of precise marketing strategies are crucial for companies. In the preferences and tastes of modern consumers, companies cannot fully satisfy every consumer. However, the advent of the era of big data is a great opportunity for companies. Provides the opportunity to use data analysis and mining technology to segment customers through these massive data, thereby improving the quality of corporate decision-making [1].

Although the traditional segmentation model performs well in customer classification, it ignores the periodicity of customer buying behavior and the purchasing power of products, and these two aspects reflect customer value information. In addition, in the classic RFM model, the defined time variable only considers the latest transaction behavior of the customer. But in many cases, because customers' consumption behaviors show changes in time, such variables cannot accurately reflect customers' repeated purchases or visit tendencies [2]. To make up for the above shortcomings, this paper studies a multi-index customer segmentation model based

Article history:
Received (February 1, 2021), Review Result (April 8, 2021), Accepted (July 3, 2021)

Print ISSN: 2288-8969, eISSN: 2207-516X IJSBT

on data mining to improve the accuracy of the segmentation model. Use entropy method to assign weights, build a new index matrix, and use factor analysis to make a new index matrix The dimensionality reduction of the algorithm reduces the time complexity of the algorithm. Finally, the improved K-means algorithm is used to achieve customer classification. Through empirical analysis, it is verified that our multi-index customer segmentation method can effectively identify customer groups and help companies improve the quality of decision-making and customer relationship management.

2. Theoretical basis

2.1. Customer segmentation

Customer segmentation refers to the process in which a company divides it according to variables such as customer behavior, attributes, needs, and preferences under a specific market environment and operating model, and provides services and products that meet the needs. Research on customer segmentation Mainly from the following four aspects, including customer behavior, demographic methods, lifestyle segmentation, and interest segmentation methods. Currently, the segmentation method based on customer behavior is the most extensive. This method is based on information technology and uses database Existing customer behavior data completes customer segmentation. The most commonly used method is the customer segmentation method based on the RFM (Recency, Frequency, and Monetary) model proposed by Hushes. For example, Dursun and Caber use the RFM model to analyze the hotel customer relationship management system. Value segmentation of customer consumption behavior information [3]. Krishna and Ravi use the RFM model to segment customers, helping companies customize products and services according to customer needs, and improve customer experience and satisfaction [4]. Cho and others believe that customers' importance is not the same, so a weighted RFM model is proposed to mine behavior patterns from customer consumption data to improve the accuracy of recommendations and complete customer segmentation [5].

Secondly, another important issue of customer segmentation is the division of indicator systems. According to customer segmentation variables, the entire customer group is divided into different small groups, which are composed of customers with similar needs and characteristics. For example, Park et al. A model framework for customer segmentation in a multi-category context to predict customer purchase patterns [6]. Kwac et al. subdivide lifestyles based on customer electricity consumption data and determine which lifestyle groups can become a certain lifestyle group based on the results of the segmentation. Some good candidates for energy projects put forward suggestions [7]. Raab et al. identify different market segments according to the role and behavior of customers in service provision and establish close contact with customers to improve service quality [8].

Nowadays, customer segmentation can not only effectively identify key customer groups, but also help companies understand customer behavior and preferences at a deeper level. Using customer segmentation results to help companies develop differentiated customer management and marketing strategies to achieve a win-win situation for both companies and customers.

2.2. Data mining

Data mining refers to the process of discovering hidden information from a large amount of data, such as Patterns, Trends, and Relationships. It can also be said to extract information or knowledge from data. Through the use of complex data analysis tools to highlight the

information structure under large data sets and discover the hidden potential relationships between these data. For customer consumption data, data mining technology can help companies better maintain customer relationships, multi-attribute and multi-dimensional discovery The differences in consumer demand and behavior patterns of customer groups can achieve precise customer relationship management. Data mining technology mainly has the following aspects: Clustering, Classification, Regression analysis, Prediction、Association rules [9].

In data mining technology, several commonly used algorithms for customer relationship mining are as follows: clustering algorithm, classification algorithm, and association rule mining. The clustering algorithm can discover the differences in consumer behavior of different customer groups and help companies formulate precise marketing strategies. Classification Algorithms can predict the trend of future customer consumption behavior. Association rule mining can find out the relationship between customers and products and guide companies to cross-sell. Murray [10] and others use data mining methods to identify behavioral patterns in historical noisy transmission data, thereby Better achieve customer segmentation. Tleis et al. [11] use the K-means clustering algorithm to achieve customer value segmentation in the organic food market. Peker et al. [12] achieve customer segmentation in the grocery retail industry through LRFM model clustering. Lotko et al. [13] used neural networks to model and analyze customer loyalty in the maintenance service industry.

Data mining technology has become an important tool for enterprises to assist decision-making. Effective customer relationship management requires the use of data mining technology to realize the feature extraction and value classification of customer information. Making full use of customer consumption information can improve the quality of customer loyalty and relationship management. At the same time, it effectively allocates resources to maximize the company's profits and maintain the competitiveness of the same industry. Therefore, it is of great significance for companies to use data mining techniques in customer relationship management.

3. Model establishment

This section mainly introduces the customer segmentation model and customer segmentation process of data mining. It mainly includes the following steps: data acquisition and preprocessing; analysis and modeling; model evaluation and optimization. Among them, the innovation of the model is mainly reflected in the “Analysis and Modeling” stage. Including the construction of RFMPQ multi-index segmentation system, objective weighting by entropy method, factor analysis dimensionality reduction, clustering to achieve customer segmentation.

3.1. Data acquisition and preprocessing

Data acquisition is the basis of data mining work. It is extracted based on the results of demand analysis and collected data, mainly from network data and local databases. However, there are a lot of abnormal data in the original data, such as missing data, outliers, inconsistencies, etc. Seriously affect the efficiency of the data analysis model, and even lead to deviations in the analysis results. Therefore, data cleaning becomes especially important. After the data cleaning is completed, the next thing that needs to be done is a series of operations such as data conversion, integration, and protocol. It is the data acquisition and preprocessing. On the one hand, data preprocessing can improve the quality of basic data, and on the other hand, it can better adapt the data to specific data mining patterns and reduce the time spent on the model.

3.2. Build the model

In the classic RFM model, only the recent transactions of the customer are considered, and the overall behavioral characteristics of the customer cannot be fully characterized. Combined with the multi-dimensional characteristics of the data, we update and optimize the traditional customer segmentation indicators, which are mainly reflected in the following aspects: 1. Divide each dimension into macro and micro aspects. The macro aspect can reflect the customer's overall consumption situation, and the micro aspect reflects the characteristics of their recent purchase behavior. The periodicity and quantity of customer purchase behavior are increased. On the one hand, it can It accurately reflects the customer's transaction behavior, and on the other hand, reflects the customer's purchasing power.

First of all, in the selection of R (Recency), the classic customer segmentation model usually selects the time interval from the customer's last visit date to the observation period. On this basis, we modify the proximity variable to the customer N visit date to the observation period The average number of days between, you can observe the degree of customer visits to the company, and provide information about the tendency to repeat purchases, the model calculation formula is as follows

$$y = \frac{1}{n}\sum_{i=1}^{n} date\_dis\ (t_{enddate}, t_{m-i}) \tag{1}$$

Where $\mathrm{date\_dis}\ (t_{\mathrm{enddate}}, t_{m-i})$ represents the difference between the date of the observation period and the date of the customer's visit. t_m is the last visit of the customer. n is the total number of visits by the customer. When n=1, The newly defined proximity value variable is equal to the traditional proximity value, so the new feature variable contains the classic variable characteristics. R1 is the ratio of the average proximity value of customer consumption to the average proximity value of all customers, and R2 is the customer proximity within a year The ratio of the value to its historical proximity value. The higher the ratio, the closer the customer's consumption time is to the observation period, the smaller the customer's churn. On the contrary, the greater the customer's churn.

In the selection of F (Frequency), the total number of consumptions during the customer's observation period is the numerator, and the average number of consumptions of all customers is the denominator, and the ratio of the two is recorded as F1. The macro aspect reflects the level of the customer among all customers. The micro aspect, Select the total number of consumptions in the past year and the total number of consumptions in the past year, the ratio of the two is F2. The purpose is to observe the recent changes in customer loyalty, if the ratio of the total number of consumptions in the past year to the total number of consumptions is larger, which shows that the loyalty of customers is on the rise.

Select M (Monetary), count the total consumption number of customers during the observation period and calculate the average consumption amount of all customers. M1 is the ratio of the total consumption number of customers to the average consumption amount of all customers, and M2 is the customer's recent consumption and its history The ratio of the total consumption amount. Through the ratio of the consumption amount, you can observe the level of the customer's contribution to the company. If the ratio is larger, it means that the customer's purchasing power is greater. The company should invest resources in this part of the customer to increase customer satisfaction and customer value. On the contrary, the smaller the customer's purchasing power, the company should allocate resources appropriately and formulate effective marketing strategies to stimulate customer consumption.

In the determination of P (Periodicity), we define it as the standard deviation of the customer visit interval, which can reflect whether the customer regularly visits the store. The calculation formula is as follows

$$Periodicity = stdev(VT_1, VT_2, \cdots, VT_n) \quad (2)$$

Where n represents the number of customer visit interval values. VT represents the visit time interval, which refers to the time elapsed between two consecutive visits by the customer. P1 is the ratio of the period value of the customer's purchase of the product to the average period value of all the customers' purchases of the product. P2 is the ratio of the cyclical value of the customer's recent purchases to the total historical total purchasing cyclical value. The periodicity indicates whether the customer visits tend to be carried out regularly. If a customer's periodicity value is low, it means that the customer visits or purchases the time interval is relatively fixed and can be considered regular.

Q (Quantity) is the number of products purchased in the customer's consumption record, Q1 is the ratio of the number of products purchased by the customer to the average number of purchases by all customers. Q2 is the ratio of the number of products purchased by the customer shortly to the total number of purchases in history. By observing this indicator, the purpose is to find from the customer's purchase record that the more types of customer consumption, the higher the possibility of cross-selling to such customers. After analyzing the commodity shopping basket, they are more inclined to buy more types of products. Enterprises can complete product cross-selling according to this psychological trend of customers and increase product sales. The index system for constructing the RFMPQ model is shown in [Table 1].

Table 1. Customer segmentation index system

Index system	Subdivided categories
R(Recency)	R1, the ratio of the average proximity value of customer consumption to the average proximity value of all customers R2, the ratio of the customer's recent purchase proximity value to its historical proximity value
F(Frequency)	F1, the ratio of the total number of customer consumption to the average number of consumptions of all customers F2, the ratio of the number of customers' recent consumption to the total number of consumptions in history
M(Monetary)	M1, the ratio of the total customer consumption to the average consumption of all customers M2, the ratio of the customer's recent consumption to its total historical consumption
P(Periodicity)	P1, the ratio of the periodicity value of products purchased by customers to the average periodicity value of products purchased by all customers P2, and the ratio of the periodicity value of products purchased by customers shortly to their total historical purchase periodicity value
Q(Quantity)	Q1, the ratio of the number of products purchased by customers to the average number of purchases by all customers Q2, the ratio of the number of products purchased by customers in the recent past to the total number of purchases in their history

Further research, the classic RFM model has different opinions on the division of indicator weights. Hughes and Arthur believe that the RFM model is the same in terms of weight division and should be given the same weight value. The empirical analysis of Stone and Jacobs using credit card user data shows that each of the weights of the indicators is not the same. The frequency value should be assigned the highest, the recency is the second, and the cost is the lowest.

At present, the research on the weight of customer segmentation indicators mainly includes the following two aspects: First, subjective weighting method, including analytic hierarchy process and eigenvalue method, etc., subjective evaluation method is related to the decision-makers own understanding ability, and the influence of human factors is relatively large. Large. The second is the objective weighting method, including the range method, entropy method, etc. The objective evaluation method emphasizes the application of mathematical theories, starting from the degree of dispersion of data and the degree of information contribution, and is not affected by the decision-makers themselves.

To obtain more objective customer segmentation results and highlight the importance of indicators, the entropy method is used to calculate the weight of the segmentation indicators. The weight value is determined according to the ability of the information provided by the observed value of each indicator. The specific steps of the entropy method are as follows:

(1) Build a data matrix

$$A = \begin{pmatrix} X_{11} & \cdots & X_{1j} \\ \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} \end{pmatrix}_{i \times j} \tag{3}$$

Where X_{ij} is the i-th customer and the value of the j-th segment index.

(2) Data standardization processing

Where to avoid the meaninglessness of the logarithm when calculating the entropy value, the data is translated, and the positive indicator.

$$X'_{ij} = \frac{X_{ij} - \mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right)}{\mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right) - \mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right)} + 1 i = 1,2, \cdots, n; j = 1,2, \cdots, m \tag{4}$$

negative index

$$X'_{ij} = \frac{\mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right) - X_{ij}}{\mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right) - \mathrm{m}\ \left(X_{1j}, \cdots, X_{nj}\right)} + 1\ i = 1,2, \cdots, n; j = 1,2, \cdots, m \tag{5}$$

(3) Calculate the proportion of the i-th customer and the j-th indicator

$$P_{ij} = \frac{X'_{ij}}{\sum_{i=1}^{n} X'_{ij}}, (j = 1,2, \cdots, m) \tag{6}$$

(4) Calculate the entropy value of the jth index

$$e_j = -k \sum_{i=1}^{n} P_{ij}\, ln\left(P_{ij}\right) \tag{7}$$

(5) Calculate the coefficient of variance for the j-th index

$$g_j = 1 - e_j \tag{8}$$

For the j-th index, the greater the difference of the index value X'_{ij}, the greater the effect on program evaluation, the smaller the entropy value, and the larger the g_j value. The more important the index is.

(6) Calculate the weight of each indicator

$$W_j = \frac{g_j}{\sum_{j=1}^{m} g_j}, j = 1,2,\cdots,m \tag{9}$$

Make $W = \begin{pmatrix} W_1 & & \\ & \ddots & \\ & & W_m \end{pmatrix}$, calculate it with the original indicator matrix, namely

$$A' = AW \tag{10}$$

The entropy method determines the size of the weight value according to the degree of difference of various indicators, avoiding the deviation caused by subjective factors, but the entropy method cannot reduce the dimensionality of the evaluation indicators, and there is a phenomenon that the clustering time complexity is relatively high. So, we used factor analysis to reduce the data dimension of the new index matrix.

Factor analysis model: Generally, $X = (x_1, x_2, \cdots, x_p)'$ is an observable random variable, and there is

$$X_i = \mu_i + a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{in}f_m + e_i \tag{11}$$

where $f = (f_1, f_2, \cdots, f_m)'$ are common factors, $e = (e_1, e_2, \cdots, e_p)'$ are special factors, f and e are random variables that cannot be directly observed. $\mu = (\mu_1, \mu_2, \cdots, \mu_p)'$ is the mean value of the population X. $A = (a_{ij})_{p*m}$ is the factor loading matrix.

Usually, X is first standardized so that its mean value is zero and variance is 1, so there is: Suppose:

(1) The mean of f_i is 0, and the variance is 1;

(2) The mean of e_i is 0, and the variance is δ_i;

(3) f_i and e_i are independent of each other.

Then X is called a factor model with m common factors.

If it is satisfied that f_i and f_j are mutually independent ($i \neq j$), then the factor model is called an orthogonal factor model. The orthogonal factor model has the following characteristics: The variance of X can be expressed as

$$Var(x_i) = 1 = {a_{i1}}^2 + {a_{i2}}^2 + \cdots + {a_{im}}^2 + \delta_i \tag{12}$$

Assume

$${h_i}^2 = {a_{i1}}^2 + {a_{i2}}^2 + \cdots + {a_{im}}^2 \tag{13}$$

Then

(1) ${h_i}^2$ is the contribution of m common factors to the i-th variable, which represents the i-th common degree or common variance;

(2) δ_i is the special variance, which represents the part that cannot be explained by the common factor. The factor loading is the correlation coefficient between the random variable and the common factor.

$${g_j}^2 = \sum_{i=1}^{p} {a_{ij}}^2, j = 1,2,\cdots,m \tag{14}$$

Call $g_j{}^2$ the "contribution" of the public factor f_j to X, which is an indicator to measure the importance of the public factor.

Factor analysis steps:

(1) Input the original data X_{n*p}, calculate the sample mean and variance;

(2) Find the sample correlation coefficient matrix $R = (r_{ij})_{p*p}$;

(3) Find the characteristic root $\lambda_i(\lambda_1, \lambda_2, \cdots, \lambda_p > 0)$ of the correlation coefficient matrix and the corresponding normal orthogonal eigenvector;

(4) Determine the number of common factors;

(5) Calculate the common variance $h_i{}^2$ of the common factor;

(6) Rotate the load matrix to better explain the common factors. The factor analysis method uses the relationship between variables and a few factors to express the correlation between multiple indicators. See Kaiser metrics [Table 2].

Table 2. Factor analysis metrics

Identify category	Range of values	Factor analysis fits the situation
KMO	0.90~1.00	Marvelous
	0.80~0.89	Meritorious
	0.70~0.79	Meddling/Middle
	0.60~0.69	Mediocre
	0.50~0.59	Miserable
	小于 0.5	Unacceptable
Bartlett‘s test	≤0.01	Acceptable

For the new index matrix, we determine whether the variables are suitable for factor analysis according to KMO and Bartlett's test, and determine the number of factors by referring to and through the cumulative variance contribution rate and characteristic root. The cumulative variance contribution rate is generally not less than 85%, and the characteristic root is required to be greater than 1.

Next, you need to cluster the factor variables to complete customer segmentation. The commonly used clustering algorithm is the K-means algorithm, which randomly selects a set of initial clustering centers and continuously updates iteratively until the clustering results no longer change [14]. However, the determination of the K value in the K-means algorithm is difficult to estimate. At first, we were not sure how many categories are the most appropriate to divide the data set into. Some of the K values were determined based on research experience. In addition, in the clustering algorithm, the choice of the initial center point has a greater impact on the classification results. If the initial value is not well selected, the expected effect may not be obtained. Therefore, we use an improved K-means algorithm to make up for the above shortcomings.

First, determine the optimal number of clusters K according to the SSE (elbow method). SSE is defined as the sum of the squares of the distance between the object of each cluster and its cluster center. Generally, the more categories, the smaller the SSE. A suitable K The value can be defined as the value at which the SSE decline rate significantly slows down. Because when the K value is less than the true number of clusters, the increase in K will greatly increase the degree of aggregation of each cluster, so the downward trend of SSE is obvious. And when the K value When the actual number of clusters is reached, the degree of aggregation obtained by

increasing K will be extremely small, and the decline of SSE will also be sharply reduced. Therefore, the trend graph of SSE and K is an elbow shape, and the position of the elbow is the actual number of clusters corresponding to the K value.

In addition, after determining the number of clusters, in the selection of initial points, we select K points as far as possible. This improvement is simple and intuitive, but it is very effective. The specific algorithm is described as follows:

(1) Randomly select a point from the input data set as the initial cluster center point;

(2) For each point X in the data set, calculate the distance $D(x)$ from the initial cluster center point, and put it in an array, and then add the distance to get $Sum(D(x))$;

(3) Select the next new cluster center point. The selection principle is the point with the larger $D(x)$, that is, the point farthest from the initial center point, has a greater probability of being selected. Obtained by the weight method The next initial seed point. The steps are as follows: 1. Take a random value Random that can fall in $Sum(D(x))$, and the calculation method is $Sum(D(x))$ multiplying a random number between 0 and 1; 2. Find out the interval where Random is currently located. Random is equal to Random minus $D(x)$ until it is less than or equal to 0. At this time, the corresponding point is the next initial seed point. As shown in Figure 1, Random is larger probability falls in $D(x_3)$. 3. Repeat steps (2) and (3) until K initial clustering center points are selected. 4. According to the selected K initial clustering center points, Run the standard K-means algorithm.

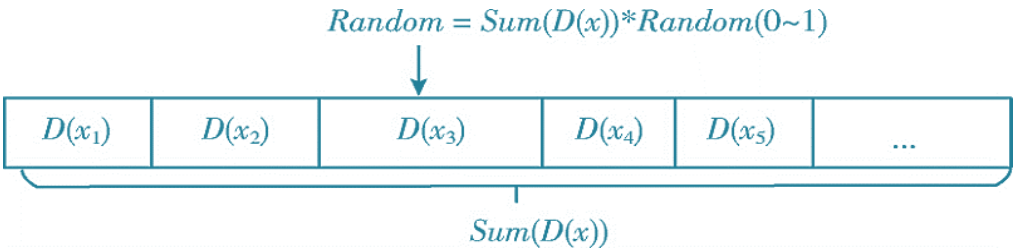

Figure 1. Selection of initial clustering center point

In addition, in terms of distance calculation, we use Euclidean distance

$$D_{ij} = \|X_i - C_j\| = \sqrt{\sum_{u=1}^{n} |x_{iu} - c_{ju}|^2} \quad (15)$$

Where X_i is the vector formed by all indicators of sample i, C_j is the vector of the center point of cluster j corresponding to these indicators, and n is the number of indicators.

3.3. Model evaluation

To verify the effectiveness of the results, we compare the results of customer segmentation with classic RFM indicators. And verify the optimization of the clustering time and the number of iterations after the initial center point is selected. The main consideration in the evaluation of the clustering effect is the category Therefore, we use the average Euclidean distance between each customer point and its cluster center point as the standard.

$$\tilde{d} = \frac{\sum_{i=1}^{m} \sqrt{\sum_{u=1}^{n} \left| x_{iu} - c_{ju} \right|^2}}{m} \tag{16}$$

X_i is the vector formed by all indicators of sample i, C_j is the vector of the center point of cluster j corresponding to these indicators, n is the number of indicators, and m is the number of samples in the class.

4. Application

4.1. Data preprocessing

We use Point of Sales data (POS data) provided by a retailer in the past 3 years as an example. The data set contains more than 30,000-member information and approximately 380,000 consumption records. The missing attribute information was deleted, and a small part of the missing information was interpolated and completed. Through the cleaning and integration of the data, 31099 pieces of basic member information and 362,368 pieces of consumption information were finally retained, about 94% of the original data set.

4.2. Analysis and modeling

First, the weight of each indicator obtained according to the entropy method is W = (0.12044223, 0.13227003, 0.00438084, 0.26321809, 0.00650823, 0.16555703, 0.00389175, 0.14118024, 0.00505721, 0.15749435). The weight value obtained is calculated according to formula (10) New data matrix.

Next, use KMO and Bartlett's test to determine whether the new data matrix is suitable for factor analysis. It can be seen that KMO=0.857, indicating that the data matrix is more suitable for factor analysis. Bartlett's test Sig value is less than 0.05, indicating that zero is rejected the assumption is that the correlation matrix is not a unit matrix, and there are common factors between the original matrices, which is suitable for factor analysis. Further, the number of factors is determined by calculating the cumulative variance contribution rate and the characteristic root, as shown in [Table 3].

Table 3. Explanation of total variance

Element	Initial eigenvalue			Rotate the sum of squares loading		
	Total	Variance %	Total %	Total	Variance %	Total %
1	5.74	57.42	57.42	4.25	42.56	42.56
2	1.59	15.96	73.39	2.52	25.26	67.82
3	1.36	13.60	87.00	1.91	19.18	87.00
4	0.59	5.92	92.92	-	-	-
5	0.26	2.60	95.52	-	-	-
6	0.21	2.11	97.63	-	-	-
7	0.11	1.13	98.76	-	-	-
8	0.05	0.54	99.31	-	-	-
9	0.05	0.51	99.81	-	-	-
10	0.02	0.18	100.00	-	-	-

As can be seen from [Table 3], the variance percentage of factor 1 is 57.42%, the variance percentage of factor 2 is 15.96%, the variance percentage of factor 3 is 13.605%, and the cumulative contribution rate of the first three factors is 87%. In addition, observe the characteristic value Load the data with the rotated sum of squares, and finally, we selected 3 factors.

In terms of clustering, to make up for the shortcomings of traditional clustering algorithms, we first determine the number of optimal clusters according to the SSE method and determine the K value by observing the position of the elbow. After dimensionality reduction, data of 3 common factors are selected. Set as input to find out the position of the elbow, as shown in [Figure 2]. The K value corresponding to the elbow is 5, so for this data set, the optimal number of clusters should be 5 categories.

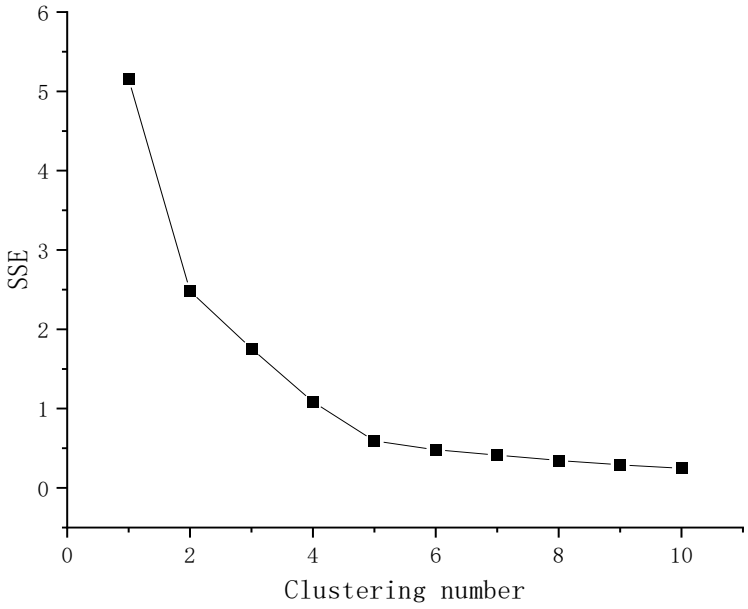

Figure 2. SSE diagram

Next, in the selection of the initial clustering center point, 5 points as far away as possible are selected as the initial clustering center, and the results are shown in [Table 4].

Table 4. Initial cluster centers

	1	2	3	4	5
FAC1	-1.02428	2.94218	-0.54740	-0.17305	-0.50178
FAC2	0.95368	0.33829	-1.67832	0.91526	0.95785
FAC3	-1.55867	0.87127	-0.14526	1.55275	-0.19257

Finally, we divide the customers into 5 categories according to the standard K-means algorithm, and the clustering information is shown in [Table 5].

Table 5. Multi-index customer segmentation results

Category	Closeness	Frequency	Money	Cycle	Quantity	Number of samples	$\bar{d}$
C1	2.1	2.7	9116.9	3.5	16.4	9458	0.84
C2	4.1	3.4	12790.3	4	23.2	4706	0.69
C3	3	1.6	4832.4	2.2	8.6	6641	0.27
C4	4.8	7.7	36356.1	4.5	63	2434	0.97
C5	2.6	1.1	5785.9	2.8	10.3	7860	0.57
Total						31099	

Based on the clustering results, we subdivide customers into 5 levels, namely: C1 medium-value customers, C2 important development customers, C3 low-value customers, C4 high-value customers, C5 general customers.

4.3. Model evaluation

For the same data set, the customers are segmented according to the classic customer segmentation indicators and calculated according to the evaluation method described in this article. The clustering information is shown in [Table 6]. Simultaneously monitor the algorithm running time and number of iterations of the new model and the classic model, And the change in the size of the cluster center.

Table 6. Results of the traditional segmentation model

category	Closeness	frequency	Money	Number of samples	$\bar{d}$
C1	2.19	1.69	7069.11	8312	0.77
C2	4.46	1.41	5406.69	12217	0.56
C3	2.99	5.9	15516.9	7105	0.85
C4	2.59	7.06	21398.7	2631	1.3
C5	2.52	10.4	34077.5	834	2.99
Total				31099	

Through actual cases, it is found that after the initial cluster center is found, the average value of cluster center changes (the average of the changes of 5 cluster center points) has dropped from the initial 1.87 to 0.57, indicating that the selection of the initial point has a significant impact on the clustering iteration. Large impact. From [Figure 3 (the abscissa is the number of clustering iterations, the ordinate is the mean value of cluster center changes), it can be seen that the standard clustering algorithm has been iterated more than 70 times, and iterated after adding the initial point After more than 30 times, the number of clustering iterations is about 1/2 of the original, indicating that the selection of the initial clustering center points has been optimized. In short, the initial clustering center points are dispersed as much as possible, which can be effective Reduce the number of iterations and speed up the calculation. Moreover, the time spent on clustering has dropped from 00:01.13 to 00:00.50. It can be seen that the number of iterations and clustering time has been optimized after the algorithm is improved.

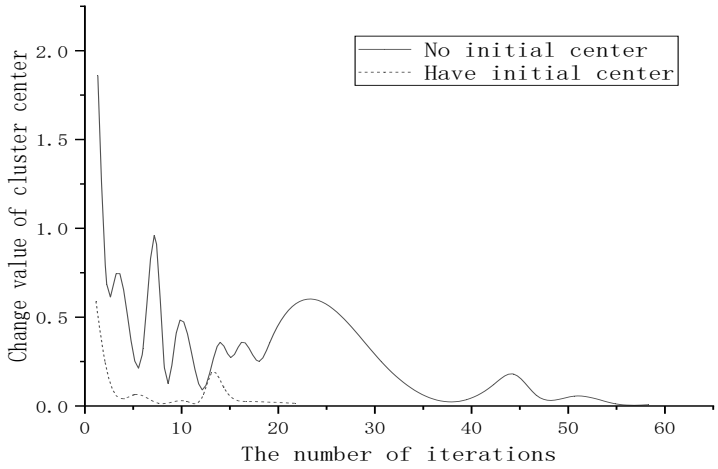

Figure 3. Comparison of results

We compare the segmentation results of the two models, and it is easy to find that in the classic RFM model segmentation results, except for the large difference in the amount of cost, the difference in other characteristics is small. In addition, the average within each category can be seen the distance is larger. It can be seen from Table 6 that the segmentation results obtained by using the multi-index customer segmentation model have larger differences between classes, smaller intra-class differences, and more compact clustering effects. This shows that the model is superior to traditional customer segmentation methods in terms of clustering compactness and feature division capabilities, and can effectively help companies distinguish different types of customer groups, and improve the level of customer relationship management and the quality of decision-making.

5. Customer segmentation strategy

The multi-index customer segmentation model proposed in this article, based on the segmentation results, can help corporate decision-makers formulate precise marketing strategies, strengthen the connection between the company and customers, and thus bring higher profits. In this section, we will provide examples of management strategies based on customer segmentation. The purpose is to retain high-value customers, attract general customers, and strive for important development customers, thereby improving corporate profits and customer satisfaction.

Medium value customers (C1), are the customers with the largest proportion of the company, accounting for about 30% of the total, and the consumption level is the average level of the overall customer. However, in this group, the average recency value of customer consumption is low, indicating that the customer purchases the product The time interval is longer and the possibility of churn is higher. Companies should pay attention to the latest news of such customers and adopt certain marketing methods to reduce the possibility of customer churn.

Important development customers (C2). They are potential value customers of the company. The number of customers accounts for 15.1% of the total. Although the consumption level is lower than that of high-value customers, they are loyal customers of the company as a whole and have great potential for development. In marketing activities, companies should pay attention to the relationship with such customers, formulate appropriate user strategies to

stimulate their consumption. In addition, promote the transformation of important development customers to high-value customers, and achieve long-term stable returns for the company.

Low-value customers (C3) and general customers (C5), these two types of customers account for about 50% of the total number of customers. The overall performance is that the purchase amount is small, the frequency is low, and the time interval is relatively long. The purchase behavior is very random. Generally, product promotion and price reduction are very attractive to such customers. Companies can formulate marketing activities regularly to promote their transformation to developing customers. At the same time, companies should appropriately reduce the resource input of such customers and transfer to valuable Customer groups, to achieve the effective use of corporate resources.

High-value customers (C4), their purchase amount is large, the consumption frequency is more, the purchase type is various, and the contribution to the company is the largest, but they account for the smallest proportion, accounting for 7.8% of the overall customer. The company is in customer relationship management, should focus on such customers. Prioritize corporate resources to them, and carry out personalized management and precise marketing strategies to improve their satisfaction and loyalty, and extend the consumption cycle of such customers.

6. Conclusion

To provide companies with a deeper understanding of consumer behavior and preferences, to help companies make decisions and develop customer relationships, combined with existing customer segmentation methods, propose a multi-index customer segmentation model. From a macro and micro perspective, the traditional indicators Optimize and construct RFMPA multi-index customer system; adopt entropy method to objectively weight; adopt factor analysis to reduce dimensionality; adopt improved K-means algorithm to complete customer segmentation. Use supermarket chain customer consumption data to conduct empirical research and compare data experimental results It shows that the model can better solve the problem of customer segmentation and improve the quality of enterprise customer relationship management and decision-making. Based on the research of customer segmentation problems, future work will focus on more detailed customer classification and analyze various user characteristics of different customers. Combined with data mining technology to assist the decision-making and optimization of customer segmentation.

References

[1] M. Kowalczyk and P. Buxmann, “An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study,” Decision Support Systems, vol.80, no.10, pp.1-13, **(2015)**

[2] P. Sarantopoulos, A. Theotokis, K. Pramatari, “Shopping missions: An analytical method for the identification of shopper need states,” Journal of Business Research, vol.69, no.3, pp.1043-1052, **(2016)**

[3] A. Dursun and M. Caber, “Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis,” Tourism Management Perspectives, vol.18, no.3, pp.153-160, **(2016)**

[4] G. J. Krishna and V. Ravi, “Evolutionary computing applied to customer relationship management: A survey,” Engineering Applications of Artificial Intelligence, vol.56, no.1, pp.30-59, **(2016)**

[5] Y. S. Cho and S. C. Moon, ‘Weighted mining frequent itemsets using FP-tree based on RFM for personalized u-commerce recommendation system,” Lecture Notes in Electrical Engineering, vol.274, no.1, pp.441-450, **(2014)**

[6] C. H. Park, Y. H. Park, and D. A. Schweidel, “A multi-category customer base analysis,” International Journal of Research in Marketing, vol.31, no.3, pp.266-279, **(2014)**

[7] J. Kwac, J. Flora, and R. Rajagopal, "Lifestyle segmentation based on energy consumption data," IEEE Transactions on Smart Grid, vol.9, vol.4, pp.2409-2418, **(2018)**

[8] C. Raab and S. Tanford, "Segmenting customers by participation: An innovative path to service excellence," International Journal of Contemporary Hospitality Management, vol.29, no.5, pp.1468-1485, **(2017)**

[9] M. Bejaei, K. Wiseman, and K. M. Cheng, "Developing logistic regression models using purchase attributes and demographics to predict the probability of purchases of regular and special tyeggs," British Poultry Science, vol.56, no.4, ppp.425-435, **(2015)**

[10] P. W. Murray, B. Agard, and M. A. Barajas, "Market segmentation through data mining: A method to extract behaviors from a noisy data set," Computers and Industrial Engineering, vol.4, no.17, pp.233-252, **(2017)**

[11] M. Tleis and R. Callieris, "Segmenting the organic food market in Lebanon: An application of k-means cluster analysis," British Food Journal, vol.119, no.7, pp.1423-1441, **(2017)**

[12] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: A case study," Marketing Intelligence and Planning, vol.35, no.4, pp.544-559, **(2017)**

[13] A. Lotko, P. A. Korneta, and M. A. Lotko, "Using neural networks in modeling customer loyalty in passenger cars maintenance and repair services," Applied Sciences, vol.8, no.5, pp.713-729, **(2018)**

[14] D. L. Huerta- Mu oz, R. Z. Ríos-Mercado, and R. Ruiz, "An iterated greedy heuristic for a market segmentation problem with multiple attributes," European Journal of Operational Research, vol.261, no.1, pp.75-87, **(2017)**

This page is empty by intention.