Research on improved RFM customer segmentation model based on K-Means algorithm

Yong Huang
Business School
Sichuan University
Chengdu, China
e-mail: huangy_gsgl@scu.edu.cn

Mingzhen Zhang*
Business School
Sichuan University
Chengdu, China
e-mail: zhaozhen093@163.com

Yue He
Business School
Sichuan University
Chengdu, China
e-mail: yuehe321@126.com

***Abstract*—The RFM model used for customer segmentation in the traditional retail industry is not suitable for the industry with distinct attributes of social groups, so the RFMC model is created by introducing the parameter C of social relations. Educational e-commerce enterprise M is selected for empirical study, and k-means algorithm is used for cluster analysis of valid customers of enterprise M, which resulted in 5 distinct customer groups and verified the effectiveness of the model.**

Keywords-K-Means algorithm, Improved RFM model, customer segmentation, Educational e-commerce

I. INTRODUCTION

A. Research Background

RFM model is an important quantitative analysis model in customer relationship management. The RFM model uses three parameters to describe the customer's importance and customer type, that is, recency(R), frequency(F) and monetary(M). Enterprises usually use RFM model and historical data to analyze the sales history and purchase behavior of customers and identify potential customers. However, in industries with distinct community characteristics, such as educational e-commerce industry, RFM model is difficult to evaluate the value of social promotion in customer transaction data, and its parameters show obvious limitations. For example, when the enterprise has a large number of demo courses or products that require customers to buy in a group, the order amount cannot accurately reflect the value benefits brought to the enterprise by customers in the community promotion dimension. This paper intends to combine the community relationship value C with the RFM model, and improve the RFM model to make it more suitable for the industry with the nature of community promotion.

B. Literature Review

P. Anitha et al. applied the k-means algorithm to the RFM model to evaluate the buying behavior of users in different regions [1]. Siti Monalisa et al. applied the RFM model to property insurance, customer investment, telecommunications services, healthcare, and FMCG industries [2]-[7]. Jea Young Lee et al. combined the RFM model with a degree of confidence that takes into account product characteristics to improve the prediction accuracy of the RFM model [8]. Zohre Zalaghi et al. used the extended RFM method to obtain customer behavior characteristics and used the k-means algorithm to classify customers to evaluate customer loyalty [9]. Shohre Haghighatnia et al. considered the impact of discounts on the basis of RFM model to form the RdFdMd model and achieve better customer clustering and value evaluation [10].

II. DESIGN OF RESEARCH METHODS

Educational e-commerce products have strong social attributes. Therefore, constructing a new parameter "C" of social group dimension can help enterprises measure the social radiation breadth of customers which realizes the rapid promotion of products through customers' spontaneous grouping behavior. Based on the improved customer segmentation RFMC model, the appropriate customer clustering method is determined according to the actual needs of the enterprise. In this paper, the K-means clustering method is selected and the final value of "K" (the number of customer groups) is determined according to the elbow method, and clustering analysis is carried out on this basis.

A. Parameters of RFMC Model

The RFMC model is a new segmentation model based on the traditional one. It takes into account the added value of unique demo courses in the education e-commerce industry and forms an improved M-value algorithm. Then, the model

978-1-7281-6042-9/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCIA49625.2020.00012

adds another important parameter "C" (community relation value) to the three basic parameters of "R", "F" and improved "M". The four parameters form RFMC model, which is the basis of customer segmentation and clustering. Four parameters are defined as follows:

R： The number of days from the date when the customer's last valid order was generated within one year to the date when the data set was acquired;

F：Number of customer purchases within one year;

M： The customer's total consumption amount within one year calculated according to the improved parameter "M";

C：The average number of members of the community groups associated with the customer's consumption order within one year.

B. Construction Rules of RFMC Model

1) Construction rules of parameter "R"

R_i is the number of days from the date of the last valid order for customer i to the date of the time node in a year. (Unit: Days)

2) Construction rules of parameter "F"

One purchase behavior is calculated as an order, and purchase behavior is not calculated twice when multiple items are included in the same order.

F_i is the number of valid purchases made by customer i within one year. (Unit: Times)

3) Construction rules of parameter "M"

The key to improving the parameter "M" is to recalculate the benefits brought by each commodity to the enterprise. Set the column matrix R as the purchase price matrix of unit product, R_j as the member of the column matrix R, and R_j as the unit purchase price of the product j (unit: yuan).

Due to the problem of value deviation of demo courses, it is necessary to recalculate the enterprise benefits generated by the purchase of unit products. Set the column matrix Q as the additional benefit matrix of unit product, and Q_j as the member of the column matrix Q, and Q_j represents the additional benefit of product j (unit: yuan).

The sum of the column matrix R and the column matrix Q gives the column matrix P (P_j=R_j+Q_j). P_j represents the benefits of buying product j. The calculation of column matrix P is shown in equation (1):

$$\begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{bmatrix} + \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{bmatrix} \tag{1}$$

Therefore, let matrix N be the quantity matrix of products purchased by users, N_{ij} be the member of N matrix, N_{i*} represent customer i, and N_{ij} represent the quantity of product j purchased by customer i.

After transposing column matrix into row matrix P^T, multiply matrix N and row matrix P^T to get column matrix M (M_{ij}= N_{ij}*P^T_j). M_{ij} represents the total benefit generated by customer i's purchase of N_{ij} product j. The calculation of M column matrix is shown in equation (2) :

$$\begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m1} & M_{m2} & \cdots & M_{mn} \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1n} \\ N_{21} & N_{22} & \cdots & N_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ N_{m1} & N_{m2} & \cdots & N_{mn} \end{bmatrix} * \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix}^T \tag{2}$$

Therefore, all the benefits brought by the customer i to the enterprise are as follows:

$$\mathrm{M}_i = \sum_{j=0}^{n} M_{ij} \quad ; \quad i = 1,2,3\cdots,m \tag{3}$$

4) Construction rules of parameter "C"

The community is defined as a group in which different customers buy goods in groups at the same time. The community contains information of two major dimensions: community time and community members. Community members are unique in their own communities and independent from each other.

The key to calculating parameter C is to conduct statistics on the community information associated with the customer's consumption order. Set matrix A as the row matrix of the number of community association members, A_j be the member of the row matrix A, and A_j represents the number of association members of the community j(unit: person).

From the data set, calculate whether the customer participates in each community. Set the B matrix as the Boolean matrix (0-1 matrix), which shows whether the customer participates in this community, B_{ij} is a member of the B matrix, B_{i*} represents customer i, and B_{ij} represents whether customer i participates in community j.

Multiply matrix B and row matrix A to get column matrix C (C_{ij}= B_{ij}*A_j). C_{ij} represents the number of people that customer i associated with in community j. The calculation of column matrix C is shown in equation (4):

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mn} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{bmatrix} * [A_1 \quad A_2 \quad \cdots \quad A_n] \tag{4}$$

Therefore, the community relation value of parameter C of customer i is shown in equation (5):

$$C_i = \frac{\sum_{j=0}^{n} c_{ij}}{F_i} \quad ; \quad i = 1,2,3\cdots,m \tag{5}$$

Here we get the initial R, F, M, and C parameter.

C. Normalized Data Transformation

In order to overcome the unreasonable influence of different measurement units and value ranges of parameters in RFMC model on clustering analysis results, it is necessary to normatively transform each parameter. The normalized transformation chooses the transformation of a range normal ratio. It takes the maximum and minimum values of each parameter from the data set and USES the maximum minus the minimum value to make a range. Then subtract the

minimum value of the parameter from the original data and divide by the range to get the normalized data. The calculation is shown in equation (6):

$$X_i' = \frac{X_i - \min\limits_{1\le i\le m}(X_i)}{\max\limits_{1\le i\le m}(X_i) - \min\limits_{1\le i\le m}(X_i)} \quad (6)$$

Four normalization parameters of R 'F' M 'C' with a variable interval of [0,1] are obtained through data normalization calculation, and the clustering and segmentation of enterprise customers are carried out accordingly.

III. THE EMPIRICAL RESEARCH

In this paper, M company is selected as the empirical case of research. M company is mainly engaged in the selection of Internet education products, mall construction, channel cooperation, sales and customer service, etc. There are five categories of enterprise products: demo courses, education software, knowledge payment, online courses and physical products. Since the establishment of an independent micro mall in 2018 with the fixed use of Y channel, the mall has generated a large number of customer order data, which has a distinct feature of social association.

From Y, the main sales channel of M company, the accumulated data of 362 days are selected from July 2,2018 to June 28,2019. The order data set contains 25 variable fields, and each order is recorded as a single line of data, with 263,595 pieces of original data. According to product data dimension, product additional benefit dimension, community group relationship dimension, customer buying behavior dimension and RFMC model, the parameter data of RFMC model based on customer ID is calculated. There are a total of 98,788 valid cases. The result of RFMC for first five customers is shown in Table I:

TABLE I. RFMC VALUES

Customer ID	R	F	M	C
C00001	63	3	393	1.333333333
C00002	235	2	776	0
C00003	127	1	15.9	0
C00004	79	1	10	1
C00005	126	2	167.9	4.5

Descriptive statistics are conducted on the data to obtain the value range of parameters. The results are shown in Table II:

TABLE II. DESCRIPTIVE STATISTICS

	N	Minimum	Maximum	Average
R	98788	0	358	105.29
F	98788	1	99	1.57
M	98788	2	48301.9	330.964
C	98788	0	98	3.778
Valid Cases	98788			

The maximum and minimum values of the four parameters are read from table (2). According to the normalized data transformation, four normalization parameters R 'F' M 'C' are calculated. The result of R'F'M'C' for first five customers is shown in table III:

TABLE III. R′ F′ M′ C′ VALUES

Customer ID	R'	F'	M'	C'
C0001	0.1759777	0.02040816	0.00809525	0.013605442
C0002	0.6564246	0.01020408	0.01602488	0
C0003	0.3547486	0	0.00028779	0
C0004	0.2206704	0	0.00016563	0.010204082
C0005	0.3519553	0.01020408	0.00343479	0.045918367

The cross - factor correlation analysis of the overall data is carried out, and the results are shown in Table IV:

TABLE IV. RESULTS OF CROSS - FACTOR CORRELATION ANALYSIS

		R'	F'	M'	C'
R'	Pearson correlation	1	-.154**	-.157**	.096**
	Sig.（double-tailed）		0.000	0.000	0.000
F'	Pearson correlation	-.154**	1	.579**	-.012**
	Sig.（double-tailed）	0.000		0.000	0.000
M'	Pearson correlation	-.157**	.579**	1	-.091**
	Sig.（double-tailed）	0.000	0.000		0.000
C'	Pearson correlation	.096**	-.012**	-.091**	1
	Sig.（double-tailed）	0.000	0.000	0.000	
	Valid Cases	98788	98788	98788	98788
**. At level 0.01 (double-tailed), the correlation is significant.					

It can be concluded from the table that among the two-factor variables, only the correlation coefficient of parameter F' and parameter M' is greater than 0.5. This result conforms to the characteristics of the traditional RFM model that the correlation coefficient of F and M is high, and the correlation coefficient of other parameters is low.

In addition, the absolute value of correlation coefficient between parameter C and other parameters is less than 0.1, which proves that parameter C has little influence on RFM and is a parameter with high independence

The K value is established based on the elbow method. Using Python3.6, the normalized parameter data of R 'F' M 'C' is imported and the sse-k value line graph is made. The results are shown in Fig. 1:

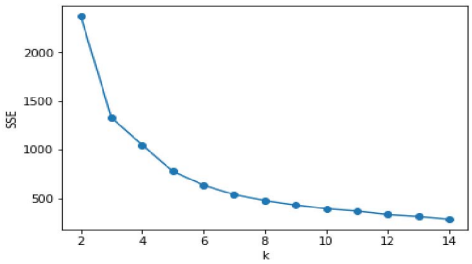

Figure 1. Line graph of SSE- K value.

According to the elbow method and the inflection point of the line graph, it can be established that the clustering effect is more significant

After K=5 is established, k-means clustering is run through SPSS, in which the variable is the parameter of R 'F' M 'C', the case labeling is based on ID Code, and the initial clustering center is randomly selected

The k-means algorithm is stable when iterating to 44 times, and the clustering center does not change. At this time, the final clustering center is obtained as shown in table V:

TABLE V. FINAL CLUSTERING CENTER

	Cluster Groups				
	A	B	C	D	E
R'	0.03721	0.35357	0.04302	0.76673	0.17794
F'	0.00908	0.00358	0.00347	0.00201	0.00777
M'	0.01032	0.00519	0.00531	0.00321	0.00741
C'	0.02686	0.02705	0.49984	0.06044	0.02295

The final cluster center R'F'M'C parameter is processed inversely, and the values of the four parameters of the intuitive RFMC are obtained. Clustering center reflects the average index level of the current cluster group, which has more distinct characteristics. Clustering results and the number of cases in each group are shown in table VI:

TABLE VI. CLUSTERING RESULTS OF RFMC MODELS

	Cluster Groups				
	A	B	C	D	E
R	13.32	126.57	15.40	274.48	63.70
F	1.89	1.35	1.33	1.19	1.76
M	500.52	252.70	258.25	156.93	359.72
C	2.63	2.65	48.98	5.92	2.25
M	28460	28371	1328	18135	22494

Marketers in enterprise M can obtain effective customer classification information and target marketing activities from table VI. For example, the number of members of the A and B customer groups is large, the Class A customer base accounts for 28.81% of the total active customers, and the contribution of profits accounts for 43.57% of the total profits provided by the effective customers of enterprises during the same period. Class A contributes most of the value benefits of enterprises, is the enterprise needs to pay high attention to the high-value customer base.

IV. CONCLUSION

The construction of the segmentation model directly affects the accuracy of the segmentation of enterprise customers. By observing the product and sales characteristics of educational e-commerce enterprises, this paper combines the radiation values of community relationship with the RFM model, and improves the algorithm of M index to form the RFMC model, making it more suitable for e-commerce enterprises with the nature of community promotion.

Through k-means clustering, the effective customer base is classified into 5 categories, which lay a foundation for enterprises to carry out customer management and marketing work.

The three parameters in the traditional RFM model are correlated, and the influence degree of parameter R in the clustering results is often much higher than that of parameter F and M, which is the same in the RFMC model. Although this is in line with the actual situation that customers' perception of the enterprise brand and stickiness decrease with the increase of the last purchase time index, the weight of the four indexes of RFMC still needs the enterprise to explore in the actual operation and customer segmentation management.

REFERENCES

[1] P. Anitha,Malini M. Patil. RFM model for customer purchase behavior using K-Means algorithm[J]. Journal of King Saud University - Computer and Information Sciences,2019.

[2] Siti Monalisa,Putri Nadya,Rice Novita. Analysis for Customer Lifetime Value Categorization with RFM Model[J]. Procedia Computer Science,2019,161.

[3] Chun Yan, Haitang Sun, Wei Liu, et al. An integrated method based on hesitant fuzzy theory and RFM model to insurance customers' classification and lifetime value determination. 2018, 35(1):159-169.

[4] Meina Song,Xuejun Zhao,Haihong E,Zhonghong Ou. Statistics-based CRM approach via time series segmenting RFM on large scale data[J]. Knowledge-Based Systems,2017,132.

[5] Mehdi Mohammadzadeh,Zeinab Zare Hoseini,Hamid Derafshi. A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran[J]. Procedia Computer Science,2017,120.

[6] Vahid Baradaran, Mohammad Biglari. Customer segmentation in Fast Moving Consumer Goods (FMCG) Industries by using developedRFM model. 2015, 7(1):23-42.

[7] Morteza Maleki Minbashrazgah,Azim Zarei,Zahedeh Hajiloo. Identifying & Segmenting Key Customers for Prioritizing them Based on Lifetime Value using RFM Model (Case study: Internet customer of Qom Telecommunications Company)[J].2016,8(2):461-478.

[8] Jea Young Lee, Ho Kuen Lee. Combined Response Modeling for Individual Marketing by RFM and Confidence. 2008, 19(2):597-608.

[9] Zohre Zalaghi,Yousef Abbasnejad Varzi. Measuring customer loyalty using an extended RFM and clustering technique[J]. Management Science Letters,2014,4(5).

[10] Shohre Haghighatnia,Neda Abdolvand,Saeedeh Rajaee Harandi. Evaluating discounts as a dimension of customer behavior analysis[J]. Journal of Marketing Communications,2018,24(4).