IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 27 October 2022, accepted 15 November 2022, date of publication 18 November 2022, date of current version 28 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223361

RESEARCH ARTICLE

Multi-Behavior RFM Model Based on Improved SOM Neural Network Algorithm for Customer Segmentation

JUAN LIAO[1,2], **AMAN JANTAN**[1], **YUNFEI RUAN**[3], **AND CHANGMIN ZHOU**[2]
[1]School of Computer Science, Universiti Sains Malaysia, Penang 11800, Malaysia
[2]School of Big Data Engineering, Kaili University, Kaili 556000, China
[3]School of Science, Kaili University, Kaili 556000, China

Corresponding author: Juan Liao (liaojuan@student.usm.my)

This work was supported in part by the Major Pilot Project of Data Science and Big Data Technology–Golden Major Construction of Kaili University under Grant 0204051002.

ABSTRACT Previous research on RFM (recency, frequency, and monetary value) models focused on only one type of user behavior data, i.e., the purchase behavior, without considering the interactions between users and items, such as clicking, favorite, and adding to cart. In this study, we propose a novel solution for deconstructing the multiple behaviors of consumers in a specific period and performing customer segmentation in an application promotion system called multi-behavior RFM (*MB-RFM*) based on the self-organizing map (SOM) algorithm. First, using the R, F, and M values, we analyzed the weight relationship between multiple behaviors of users and items using the superiority chart and entropy value methods. Each behavior ascribed to a customer was considered to be a part of *MB-RFM* model values, which were then used to classify customers using an improved SOM neural network. Subsequently, various promotion strategies were developed according to customer categories that can help application vendors in improving their application utilization and implementing targeted promotion strategies. To prove the effectiveness of the proposed method in sparse datasets, two real-world datasets were used to perform experiments, whose results demonstrated that the classification performance of our method was significantly more accurate.

INDEX TERMS Application promotion system, customer segmentation, multi-behavior RFM model, self-organizing map neural network.

I. INTRODUCTION

The emergence of online platforms has led to a boom in online shopping, particularly due to the COVID-19 pandemic. To attract more users, e-commerce platforms have launched mobile applications. Statistically, in China, according to the Ministry of Industry and Information Technology, the number of applications monitored in China's domestic market was 3.28 million by the end of February [1]. From 2016 to 2020, China's e-commerce transaction volume increased from 26.10 to 37.21 trillion yuan, with an average annual growth rate of 9.3% [1]. Hence, the effective use of user-item interaction data by mobile application manufacturers to ensure that these applications persevere has become an urgent problem in e-commerce industries.

User-item behavior is an extensively studied research topic in e-commerce, which has helped enterprises in gaining large profits and promoting applications. Different models have been employed to analyze customer behavior. Martínez et al. [30] proposed a customer segmentation method for an outfitter company in Taipei, Taiwan, focusing on promotion strategies; they considered gender, birth date, shopping frequency, and total spending as influence factors to analyze different types of customers. Peter et al. [2] proposed an RFM (recency, frequency, monetary value) model as a customer segmentation model wherein users were classified by analyzing the user-item behavior data. This model mapped dynamic behaviors according to three variables: R, which

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

refers to the interval between the last shopping time of the user; F, which refers to the sum of shopping times of the user in a certain period; and M, which refers to the total shopping amount of the user in a certain period. Wei and Lin [3] reviewed the application of this RFM model, which was widely applied in marketing, and could effectively identify valuable customers to develop a marketing strategy. Moreover, Wei et al. described the RFM model along with its application, advantages, and disadvantages in combination with other variables.

Zong and Pan [9] used customer consumption data from a retail company for customer segmentation and developed targeted marketing strategies using the SOM neural network algorithm and optimized the RFM model; the proposed method implemented four typical features-based categories of the optimized RFM model. In the SOM-based RFM model proposed by Juha Vesanto, and Esa Alhoniemi [8], data mining was employed, which projected the input space on a low-dimension regular grid. It could be effectively utilized to visualize and explore data features. Hu et al. [10] proposed customer clustering based on the SOM algorithm for the RFM model and provided corresponding marketing strategies for classified customers using marketing knowledge; this study also classified customers in specific mobile communication industries. Rachid et al. [11] used the SOM method to determine the best number of clusters and proposed an RFM model for customer segmentation; the research classified 730 customers into eight groups via the K-means method. Moreover, Wu et al. [12] empirically studied the application of K-means-based customer segmentation to purchase behavior models; they used the principal component analysis method to determine the weight of the RFM model, combined it with the K-means algorithm to perform customer segmentation, and proposed managing strategies. Wu et al. [19] obtained the user characteristics of an improved RFM model and established an e-commerce platform using the K-means++ algorithm. Cheng and Chen [13] proposed a customer relationship management (CRM) system based on the K-means algorithm and RFM model. This system was different from customer clustering because in this study, the RFM model was used to yield quantitative values as input attributes, followed by clustering; finally, rough sets were employed to mine classification rules to ensure an excellent CRM system for enterprises. Chen et al. [15] studied customer-centric business intelligence for an online retailer by analyzing the main consumer characteristics using K-means clustering and decision trees. Vohra et al. [16] presented a neural network framework that used the K-means and SOM clustering algorithms; in this study, customer segments were obtained by computing the RFM values.

K-means or SOM-based RFM models affect customer classification. Attempts have been made to improve the RFM model according to the particularity of each industrial sector. Yan Chun et al. [4] proposed a customer classification model based on an improved SOM neural network and RFMC model; it is worth noting that because this study was related to the insurance business, an additional C factor was included, which denotes the insurance compensation. Thus, different marketing strategies can be developed according to customer classification for enterprises to make operational changes with efficient resources based on customer segmentation. Mahboubeh and Mohammad [17] added weight to the RFM model based on the customer's lifetime value. With the advent of machine learning, data mining is no longer limited to customer classification and behavior data can now be analyzed. Cheng et al. [18] presented an expanded RFM model by including the time since the first purchase and churn probability. Before the introduction of the RFM model, customer purchase behaviors were analyzed using real-time transactional and retail datasets to deploy data segmentation (for example, P. Anitha and Malini M. Patil [33]). Recently, Rahim et al. [20] proposed the analysis of repurchase behaviors for customer classification using the RFM model; they focused on customer behavior modeling based on point-of-sale data. Zhou et al. [21] added a dimension of inter purchase time in the RFM model (called the RFMT model); they obtained a high customer classification rate compared to multi-layer perceptron and support vector machine systems. Monalisa et al. [14] designed a customer clustering fuzzy C-means algorithm using the RFM model, wherein analytical hierarchy process weights were employed for ranking.

With the emergence of various industry data sources, customer classification plays a vital role in customer segmentation. Qian et al. [5] used the CLARA algorithm (large clustering application) to extract ETC data with respect to the recency, frequency, and monetary factors, and constructed a segmentation index system of ETC customers. This customer segmentation index system was based on ETC consumption characteristics; it analyzed the travel characteristics of each customer to develop service strategies for the enterprise. Wei et al. [6] proposed a CRM RFM model based on customer classification, which attracted customers in a veterinary hospital. Wu et al. [7] used a SOM neural network, including two layers, with an RFM model for telegraphic customer classification. Coussement et al. [32] constructed variants of RFM-based predictive models, such as logistic regression, decision trees, and neural networks. Olson and Chae [31] discussed three prominent segmentation techniques for direct marketing, namely, RFM analysis, logistic regression, and decision trees. They recommended the use of decision trees in the context of customer segmentation for direct marketing, even in the case of problems with data accuracy.

However, these studies only concentrated on one user-item interaction behavior, as shown in Figure 1. In this study, we analyzed mobile phone applications' content and proposed an *MB-RFM* model (as shown in Figure 2) with an improved SOM algorithm. The contributions of our work are given below:

(1) We propose a novel RFM model based on SOM for customer segmentation using user-item interaction behavior data. The model shares multi-behavior

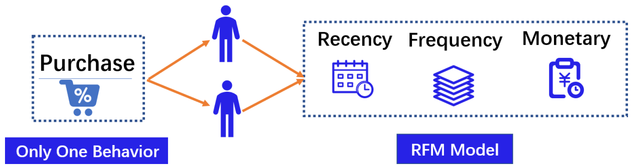


FIGURE 1. One behavior RFM model.

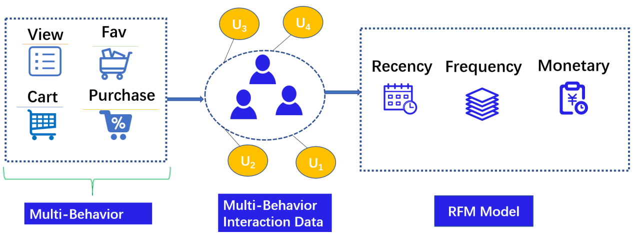


FIGURE 2. Multi-behavior RFM model.

(such as clicking, add to cart, favorite, purchase) data and calculates the value of R and F for each behavior type.

(2) To balance the weight relations between behavior types, we propose the prediction of each weight value of multi-behavior using a new method, i.e., the combination of superiority chart and entropy value methods.

(3) To ensure better customer segmentation, our model *MB-RFM* adopts an improved SOM algorithm to cluster the calculated R, F, and M values (Section III(A)). Moreover, in this study, we systematically analyzed the theoretical relationship between RFM and SOM in terms of customer segmentation.

(4) To demonstrate the effectiveness of our model, we conducted extensive experiments on two real-world datasets. This work has been adopted by a company in China and the results obtained demonstrate the efficacy of this method in increasing business turnover.

The remainder of this paper is organized as follows. Section II presents the preliminaries. In Section III, we discuss the proposed method. Section IV describes the experiments conducted in this study. Finally, Section V summarizes our work.

II. PRELIMINARIES

A. RFM MODEL

An RFM model is a generic data analysis model that classifies users based on business marketing data. It implements customer management relationships for customer segmentation and represents each user via three indexes: (1) R value, which denotes the last consumption interval, i.e., the last purchase behavior of a user from a specific period; (2) F value, which denotes the consumption frequency, i.e., the total number of purchase behaviors for a user in a specific period; and (3) M value, which denotes the consumption amount, i.e., the total amount of consumption by a user in a specific period.

B. SOM MODEL

SOM is a dimension reduction algorithm, which is widely used in clustering algorithms; it generates low-dimensional discrete maps by learning the data in the input space. SOM neural networks are composed of input and competition layers. People denote different numbers of vectors in the input layer depending on the situation; when an external signal is input into the network, the competition layer analyzes and compares the input variables, determines the rules, and classifies the vectors. A SOM neural network model comprises four steps:

(1) First, the parameters of each node are randomly initialized, while ensuring the same number of parameters and input data dimensions.

(2) Second, for each input vector, the best matching vector is determined. Given a dataset D with dimension N, let $d_m(x)$ denote the distance between two nodes. To find $d_m(x)$, the following discriminant function-Euclidean distance function can be used:

$$d_m(x) = \sum_{n=1}^{N} (x_n - w_{nm})^2 \tag{1}$$

where $X = \{x_n\}, n = 1, 2, \ldots N$ denotes each node.

(3) Third, to find closer nodes, the SOM neural network must constantly update nodes. The objective function for optimization is given below:

$$W_{m,I(x)} = \exp(D^2_{m,I(x)}/2\sigma^2)D_{m,n} \tag{2}$$

where $I(x)$ denotes node activation, $D_{m,n}$ denotes the Euclidean distance between m and n, and σ represents an attenuation function (generally, $\sigma(t) = \sigma_0 \exp(-t/T_\sigma)$). In this step, to find the node adjacent to the active node, the SOM model defines the objective function adjacent to the active node. According to Eq. (2), the degree of updating a neighboring node is affected by its distance from the active node ($I(x)$).

(4) Finally, parameter adjustment, i.e., to update node parameters, we adopted the gradient descent method, as follows:

$$\nabla w_{mn} = \eta(t).W_{m,I(x)}(t).(x_n - w_{mn})\eta^1 \ldots \eta^i \tag{3}$$

Eq. (3) is an iterative functional that keeps iterating until it converges.

III. MB-RFM BASED ON IMPROVED SOM ALGORITHM

Figure 3 illustrates the proposed customer classification method.

A. CALCULATION OF R, F, M FOR MB-RFM MODEL

Typically, previous RFM models analyzed a single user behavior, i.e., purchase behavior. However, to promote the application of this model, it is essential to analyze the clicking and adding to cart behaviors of users along with their purchasing behavior. When only one user action is considered, sparse data are obtained, which may not include some important customers; for example, a user who often browses an application without purchasing. By applying promotion methods, these

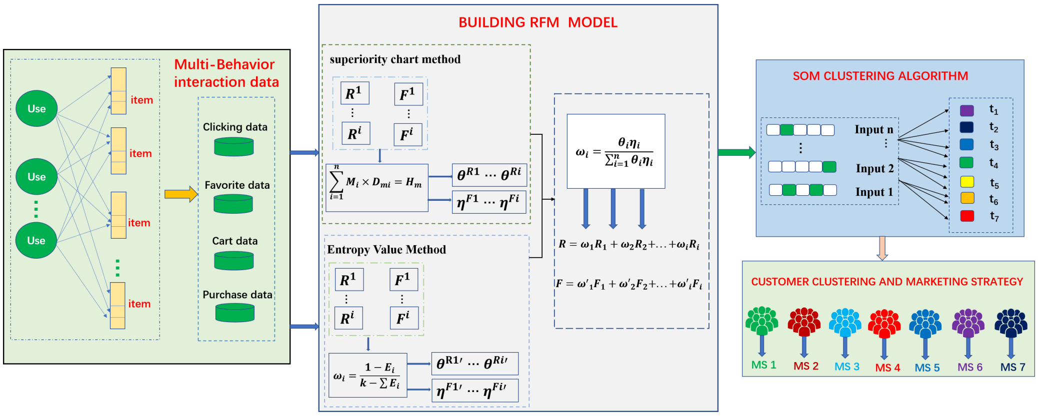


FIGURE 3. Framework of our MB-RFM based on SOM algorithm.

users may become important customers. Thus, in this study, we redefined R and F.

R Denotes Recency: R consists of three parts, i.e., the time elapsed since a user's most recent view, add to cart, and purchase actions. We carefully selected important periods, such as China's Double Eleven and May Day, to ensure consistency with actual model analysis. Liu et al. [29] proposed the standard normal distribution of periods, which had certain limitations for some applications; we confirmed this through experiments, as discussed in Section 4. Considering the different proportions of the view, add to cart, and purchase actions under R, we used the following equation to define it:

$$R = \omega_1 R_1 + \omega_2 R_2 + \ldots + \omega_i R_i \tag{4}$$

where R denotes the recent time interval values in the *MB-RFM* model, R_i denotes the i behavior type of user's characteristic recently time interval values (such as clicking, adding to cart (or collecting), purchasing behavior, preferences) and $\omega_1, \omega_2, \ldots, \omega_i$ denotes the weight value of $R_1, R_2, \ldots R_i$.

Furthermore, ω was comprehensively evaluated using the superiority chart and entropy value methods (i.e., techniques of subjective analysis and objective calculation, respectively).

1) CALCULATION OF ω (SUPERIORITY CHART METHOD)

We employed the superiority chart to calculate weights, as described below:

(1) We considered the final score of each element via the horizontal summation of the assigned score of each element to calculate the weight. Then, they were sorted according to the score level to determine the relative importance of each element, as follows:

$$\sum_{i=1}^{n} A_{ij} = A_i (i = 1, 2, \ldots, n) \tag{5}$$

where A_i denotes the summation of row i in Table 1. According to the value of A_i, we calculated the weight of each influencing factor in our model, as shown below:

$$A_i / \sum_{i=1}^{n} A_i = M_i \tag{6}$$

where M_i denotes the weight of factor i.

(2) The weight of the overall scheme was calculated according to that of each factor. The final weight of each element was determined by combining the weight (M_i) of influencing factors in Eq. (1) and D_{mi} in Table 1, as follows:

$$\sum_{i=1}^{n} M_i \times D_{mi} = H_m \quad (m = 1, 2, \ldots) \tag{7}$$

where H_m denotes the final weight of factor i.

2) CALCULATION OF ω (ENTROPY VALUE METHOD)

First, we adopted the data processing method of normalization proposed by Zhu et al. [23], as follows:

$$Y_{ij} = \frac{X_{ij} - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{8}$$

where $X_i = \{x_1, x_2, \ldots x_n\}$ denotes the number of factors (n denotes the total number of elements); further, $Y_1, Y_2, \ldots, Y_k$ denotes the normalized value.

Second, we calculated the information entropy of each factor, which can be given by [34]:

$$E_j = -\frac{1}{\ln n} \sum\nolimits_{i=1}^{n} p_{ij} \ln p_{ij} \tag{9}$$

where $p_{ij} = Y_{ij} / \sum\limits_{i=1}^{n} Y_{ij}$, denoting the specific gravity of factor j in sample factor i, and E_j denotes the information entropy of element j.

Finally, we calculated the weight of each factor according to the $(E_1, E_2, \ldots, E_k)$ information entropy.

$$\omega_i = \frac{1 - E_i}{k - \sum E_i} \quad (i = 1, 2, \ldots, k) \tag{10}$$

3) CALCULATION OF ω

Next, we calculated ω following the process described in Sections 3.1.1 and 3.1.2 as follows:

$$\omega_i = \frac{\theta_i \eta_i}{\sum\limits_{i=1}^{n} \theta_i \eta_i} \tag{11}$$

where θ_i and η_i denote the weight calculated by the superiority chart and entropy value methods, respectively.

F Denotes Frequency: Similar to R, F also consists of three parts, i.e., the frequency of viewing an item, adding to cart, and purchasing by a user in a specific period. It can be given by:

$$F = \theta_1 F_1 + \theta_2 F_2 + \ldots + \theta_i F_i \tag{12}$$

where F denotes the consumption frequency in the *MB-RFM* model, F_i denotes the frequency of user's i behavior (such as clicking, adding to cart (or collecting), or purchasing) in the latest period of time, and $\theta_1, \theta_2, \ldots, \theta_i$ denotes the weight of $F_1, F_2, \ldots, F_i$. θ was computed in the same manner as ω in Sections 1), 2), and 3).

M Denotes Monetary: M denotes the total amount of money spent by users over a specific period. This study aimed to segment the customer base more accurately without considering the profit of selling products; hence, we simply calculated the total expenditure by adding the cost of all items purchased.

B. CUSTOMER SEGMENTATION BASED ON IMPROVED SOM ALGORITHM

In this section, we establish the proposed *MB-RFM* model. We obtained the new values of R, F, and M according to Eq. (4) and (12) as the input vector for the SOM model. Figure 4 Illustrates the structure diagram of the improved SOM neural network.

Input Layer: Considering that user-item interactions have layered features, which are initialized by obtaining the view, add to cart, and purchase actions, among others, we calculated the input vector using Eq. (4) and (12).

Training: In comparison to baseline SOM, in this study, the learning rate and neighborhood function between adjacent nodes were mainly improved. It is difficult to select a reasonable learning rate in baseline SOM networks. If the learning rate is close to 1, the weight vector oscillates and gets repeatedly updated, leading to unstable learning. When the network learning rate gradually approaches 0, although it can improve the stability of learning, it reduces the convergence speed of the network. Hence, in this study, a dynamic learning speed function, λ(t), was considered:

$$\lambda(\mathrm{t}) = \lambda(0)\exp(-\mathrm{t}/\mathrm{nT}) \tag{13}$$

where λ(0) denotes the learning rate, T denotes the learning step, and n is a multiple of T (usually between 3 and 5) ([24]). This method helps in achieving a balance between fast learning and stable learning rate.

In baseline SOM networks, the improper selection of neighborhood values may lead to slow convergence or even non-convergence; thus, the neighborhood function must be improved. In this study, we defined the neighborhood function as follows:

$$\mathrm{U}_c(t) = f(t)\exp(-d_c^2/2\sigma(t)^2) \tag{14}$$

where $f(t)$ denotes a monotonic decreasing function; it should be noted that in the actual weight self-organization process ([25]), $f(t) = \frac{1}{t}$ for a continuous system and $f(t+k) = \frac{1}{t+k}$ for a discrete system. In our *MB-RFM* model, we considered continuous and layered user-item interaction behavior data with $f(t) = 1 - \frac{t}{T}$. Moreover, $\sigma(t)$ was considered to be a monotonic decreasing function; it is denoted by the radius of function $U_c(t)$. In our model, $\sigma(t)$ was not zero because $\sigma(t) = 0$ results in disordered classification. Thus, we determined $\sigma(t)$ as follows ([26]):

$$\sigma(t) = \sigma_0 \exp(-t/T) \tag{15}$$

After the aforementioned improvement of the SOM neural network, the improved SOM model's final formula can be given by:

$$W_j(t+1) = W_j(t) + \sigma(t)U_c(t)[p_k(t) - W_j(t)] \tag{16}$$

C. APPLICATION PROMOTION STRATEGIES

In Section 3.2, user-item interaction data were classified using the improved SOM clustering algorithm. In this study, customers were divided into seven categories, namely, key customers, key development customers, general development customers, general retention customers, key retention customers, general maintenance users, and key maintenance users. Our customer stratification is different from the traditional core framework of customer stratification ([27]). We innovatively propose to invest in the most critical customer stratification, which is the most effective investment method for maximizing enterprise profits. To obtain significantly larger profits than the invested amount, applications should invest in key customers that bring high profits with effective services; simultaneously, they can provide personalized services for such customers ([28]). Customers with high F and M scores are the most valuable users; they should be offered personalized services carefully to avoid aversion during launching activities. In addition, applications should focus on potential customers because they possess significant potential value and proper investment may turn them into key development customers in the future, even if they do not currently generate high profits. Furthermore, general retention customers exhibit low profits and transformation rate, with weak customer engagement; to reduce the operating cost, applications should not considerably focus on such customers. Figure 5 demonstrates the specific analysis of various customers.

(1) Key customers: This refers to customers who have frequented the mobile application recently and purchased a high amount of money. From the perspective of customer value, such customers can bring high profits to the application and demonstrate high loyalty; thus, we should focus on and maintain them.

(2) Key development customers: The average degree of closing the application, frequency, and value of this type of customer are higher than the overall average. Moreover, their application closing degree, frequency, and value are higher than the overall average; although the customer has had no recent interaction with the application, however, as the

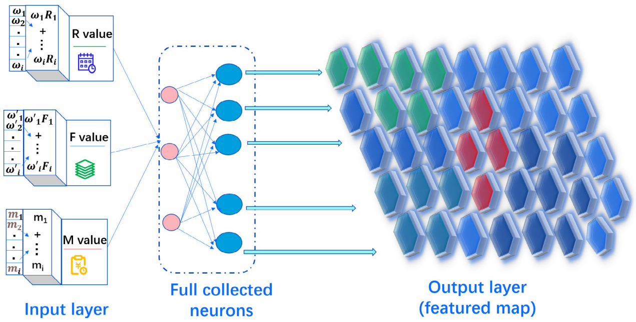


FIGURE 4. Improved SOM neural network model.

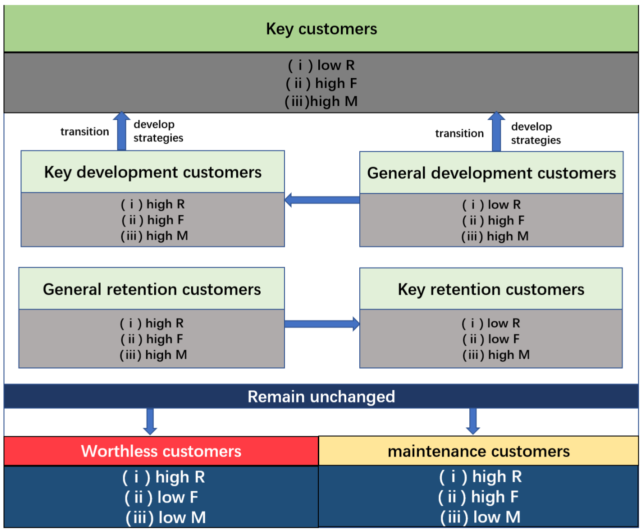


FIGURE 5. Customer stratification.

leading target group of using the application, the customer has good prospects for development. Therefore, enterprises should focus on developing and retaining this type of customers to turn them into key customers.

(3) General development customers: The average recency, money, and amount of application usage for such customers are lower than the overall intermediate level. However, their purchase frequency is higher than the general average level, indicating that such customers use the application frequently without high consumption. Hence, such customers can contribute to company profits.

(4) General retention customers: The average recency and frequency of products purchased by such customers in the application is lower than the overall intermediate level. However, the value and amount are higher than the general intermediate level, indicating that such customers rarely buy products, even though the purchase amount may be increased. Such customers bring minimal profits to the company.

(5) Key retention customers: The average recency and money in terms of application usage for such customers are higher than the overall intermediate level, while the frequency and amount are lower than the general intermediate level, indicating that such customers used the application more frequently but did not purchase any goods recently, suggesting that they may lead to loss. Enterprises should improve communication with such customers to retain them.

(6) Maintenance customers: The average recency and value of products purchased by these customers are lower than the general intermediate level, while the intermediate frequency and amount are higher than the general intermediate level, indicating that these customers purchased products more frequently in the past year; thus, they are the maintenance customers of insurance companies.

(7) Worthless customers: The average recency of products purchased by such customers is higher than the overall average level, while the intermediate frequency and amount are lower than the general intermediate level, indicating that such customers rarely purchase products and can hardly create profits for the company. Therefore, they are worthless to the company.

IV. EXPERIMENTS AND DISCUSSION

A. DATASETS AND DATA PREPROCESSING

1) DATASETS

We conducted experiments on two real-world e-commerce datasets that include multiple forms of user behaviors, such as view, add to cart, and purchase. To evaluate the accuracy of our model in sparse datasets, the two selected datasets were compared (we used the Yunji dataset, which is widely used in China, and Mengba data, which is a newly emerging dataset).

a: YUNJI DATASET

This dataset comprises data from Yunji, a popular e-commerce platform for daily necessities in China, collected between 2014/11/18 and 2014/12/18. We used the Yunji dataset in our experiment as a sample of user interaction data (it includes features such as view, add to cart, and purchase).

b: MENGBA DATASET

Mengba is a publicly available maternal and infant dataset from the WeChat applets containing 4339 users and their interaction behaviors from 2021/5/1 to 2021/5/30.

For both datasets, we combined the actions of add to cart and favorite as a data category because they do not affect the total amount of money spent by users; in other words, it does not affect M. The Mengba dataset contains data from enterprises, including private information; thus, we cleaned the data. When the data met the analysis requirements, we built the proposed *MB-RFM* model according to the objective of splitting.

2) DATA PREPROCESSING

The Yunji dataset contains approximately 100 million records; we selected only 2 million data records as representatives, including the behavior records of 61,198 users. Moreover, the Mengba dataset contains geographic data, which was not required in this study; hence, they were not included. Further, five factors of customer ID, transaction data, transaction behavior (view, favorite, add to cart, purchase), and amount of trade were selected according to the requirements of the RFM model. The datasets used in this study comprise accurate purchasing behaviors with respect to time, and a small number of users may repeatedly purchase or browse unified products within an hour; thus, such data were not processed and deleted. Moreover, some abnormal data, such as negative amount and date were deleted. We attempted to separate the required data from the datasets, i.e., data columns associated with R, F, and M.

B. COMPARISONS

We first calculated R, F, and M using the baseline RFM and proposed *MB-RFM* models. To achieve the best distance effect, we compared the clustering accuracy of the proposed *MB-RFM* model with that of the baseline RFM model.

1) VALUE OF *R*

To demonstrate the superior performance of our *MB-RFM* model, the values of R, F and M were compared with those calculated using the baseline RFM and other models. It should be noted that instead of demonstrating the calculation of each value, we have presented the final clustering effect. Figure 6 illustrates R calculated on the Yunji and Mengba datasets by the baseline RFM model. In the proposed *MB-RFM* model, the user-item interaction behavior data include view, add to cart, and purchase; Table 1 and Table 2 list the weight, ω, of R for the Yunji and Mengba datasets, respectively.

It is evident from Table 2 that for view, $\omega = 0$ because the data were selected during the Double Eleven grand promotion in China. In this period, the add to cart and purchase behaviors of users increased, with negligible impact of view. Figure 7 demonstrates R for both datasets evaluated using the proposed *MB-RFM* model.

2) VALUE OF *F*

The value of F (shown in Figure 8) was calculated in the same manner as that of R. Table 3 and Table 4 list the weight of F for both datasets.

3) VALUE OF *M*

Figure 9 illustrates the value of M in both datasets. Because view and add to cart (and favorite) were not included in user-item behavior, the calculation of M did not change.

4) CLUSTERING RESULTS OF IMPROVED SOM NEURAL NETWORK ALGORITHM

We performed the improved SOM algorithm to compare the baseline RFM model and *MB-RFM* model for clustering. Figure 10 shows the two datasets using a two-dimensional space to visualize the results. The ''leave-one-out'' algorithm was used for comparison, where one dataset was left as comparison data and the other as test data. The distance between the two datasets affected the results; however, in comparison to other models, the proposed model exhibited the best effect.

TABLE 1. Weight Of R (Yunji dataset).

Factor	Superiority Chart Method (θ_i)	Entropy Value Method (η_i)	Final weight value $\omega_i = \frac{\theta_i\eta_i}{\sum_{i=1}^{n}\theta_i\eta_i}$
View	0.025	0.31	0.020
Cart and Favorite	0.22	0.27	0.155
Purchase	0.755	0.42	0.825

TABLE 2. Weight Of R (Mengba dataset).

Factor	Superiority Chart Method (θ_i)	Entropy Value Method (η_i)	Final weight value $\omega_i = \frac{\theta_i\eta_i}{\sum_{i=1}^{n}\theta_i\eta_i}$
View	0.005	0.01	0.000
Cart and Favorite	0.42	0.45	0.378
Purchase	0.575	0.54	0.622

TABLE 3. Weight Of F (Yunji dataset).

Factor	Superiority Chart Method (θ_i)	Entropy Value Method (η_i)	Final weight value $\omega_i = \frac{\theta_i\eta_i}{\sum_{i=1}^{n}\theta_i\eta_i}$
View	0.037	0.246	0.022
Cart and Favorite	0.242	0.301	0.178
Purchase	0.721	0.453	0.799

TABLE 4. Weight of F (Mengba dataset).

Factor	Superiority Chart Method (θ_i)	Entropy Value Method (η_i)	Final weight value $\omega_i = \frac{\theta_i\eta_i}{\sum_{i=1}^{n}\theta_i\eta_i}$
View	0.062	0.021	0.003
Cart and Favorite	0.325	0.42	0.284
Purchase	0.613	0.559	0.713

Figure 11 (a) and (b) illustrate the results in two clustering dialogs for better demonstration.

C. CUSTOMER SEGMENTATION

Figure 11 shows the segmentation results for the Yunji and Mengba datasets calculated using the RFM (d, c) and *MB-RFM* (b, a) models, including seven clusters, for a total of 61198 (a, c) and 4339 (b, d) customers, respectively. For the Yunji dataset, the ''general development customers'' cluster in Figure 10 (d) had the lowest percentage of customers (0.65%), while the ''maintenance customers'' cluster had the highest rate of customers (40.19%). However, in the proposed *MB-RFM* model, the highest and lowest percentage of customers changed. Figure 10 (b) shows that

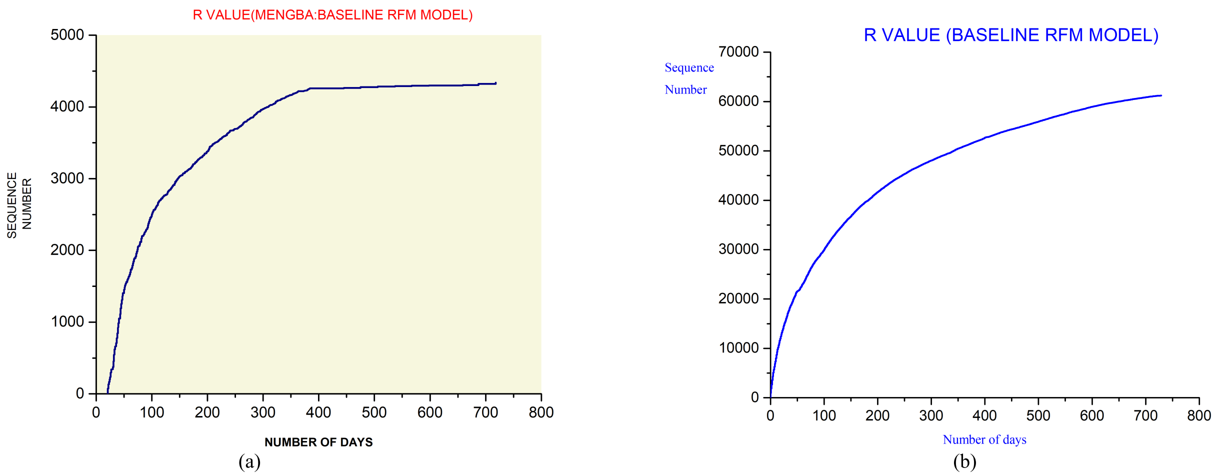


FIGURE 6. (a) Mengba dataset: R value based on Baseline RFM. (b) Yunji dataset: R value based on Baseline RFM.

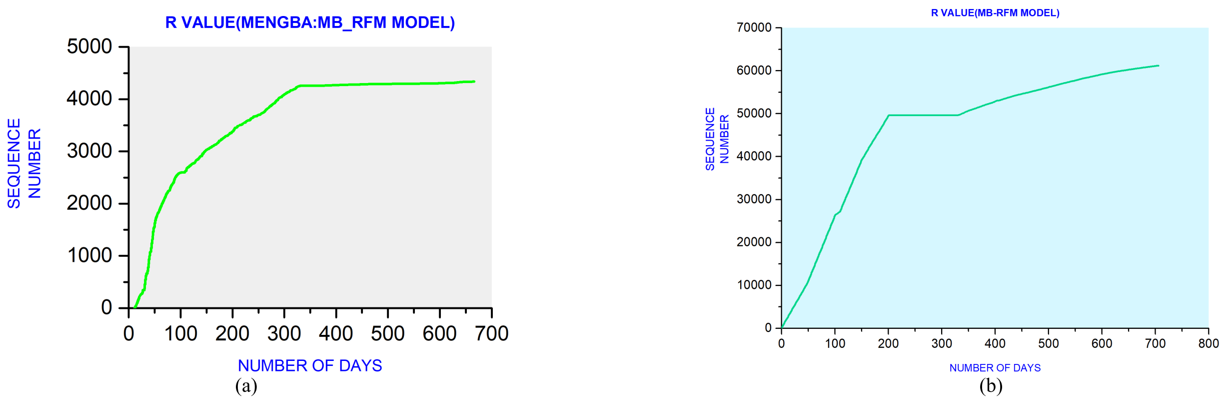


FIGURE 7. (a) Mengba dataset: R value based on *MB-RFM* model. (b) Yunji dataset: R value based on Baseline *MB-RFM* model.

the highest rate of "maintenance customers" is 28.5%, while the lowest rate of customers is 1.94%. In particular, "worthless customers" decreased by more than half from 40.12% to 13.54%, which reduced the customer churn rate according to our experimental comparison using the "leave-one-out" method. Concurrently, "key customers" increased from 4.63% to 14.47%. For the sparse Mengba dataset (Figure 11(c)), the "maintenance customers" cluster contained the lowest percentage of customers (1.87%), while the "general development customers" cluster had the highest rate of customers (38.35%). However, in the proposed *MB-RFM* model, the highest and lowest rates of customers changed. Figure 11 (a) shows that the highest and lowest rates of customers were 38.81% and 1.41%, respectively. Evidently, "worthless customers" decreased from 35.81% to 21.99%, which is not as good as that in the Yunji dataset, but also the best in terms of experimental results. In the comparison algorithm, we used the "leave-one-out" method. Table 5 lists the number of customer classifications in each category. A comparison between the remaining group of data and classification results for each test dataset showed that our model demonstrated the best effect. Moreover, only one group of user types was different, while the others were the same as our prediction, with the highest accuracy.

D. DISCUSSION

Figure 11 shows *MB-RFM* distributions in seven clusters, and Table 5 lists the number of customers and average *MB-RFM* values for each cluster; this information can be used to analyze the characteristics of clusters and formulate corresponding strategies to promote the application. The Yunji dataset was used to demonstrate an analysis example, as shown in Table 6.

(1) Key customers (T1). T1 included 14.47% of total customers (Figure 11(b)). On average, this cluster demonstrated low R, high F, and high M, suggesting that the latest user-item interaction on the application for most customers in this cluster around 11 days. For such customers, enterprises should focus on using the application behavior to timely understand their real-time needs and dynamics, and provide the services required by customers to ensure lasting loyalty.

(2) Key development customers (T2). T2 included 13.59% of all customers (Figure 10 (b)). This cluster exhibited high R, F, and M. Such customers are the leading target group of application sales and demonstrate good prospects for development and strong consumption ability. Thus, enterprises should consider the following: first, ensuring timely guidance and good service; then, applications can intensify its preferential appropriately to incite the customer purchase desire;

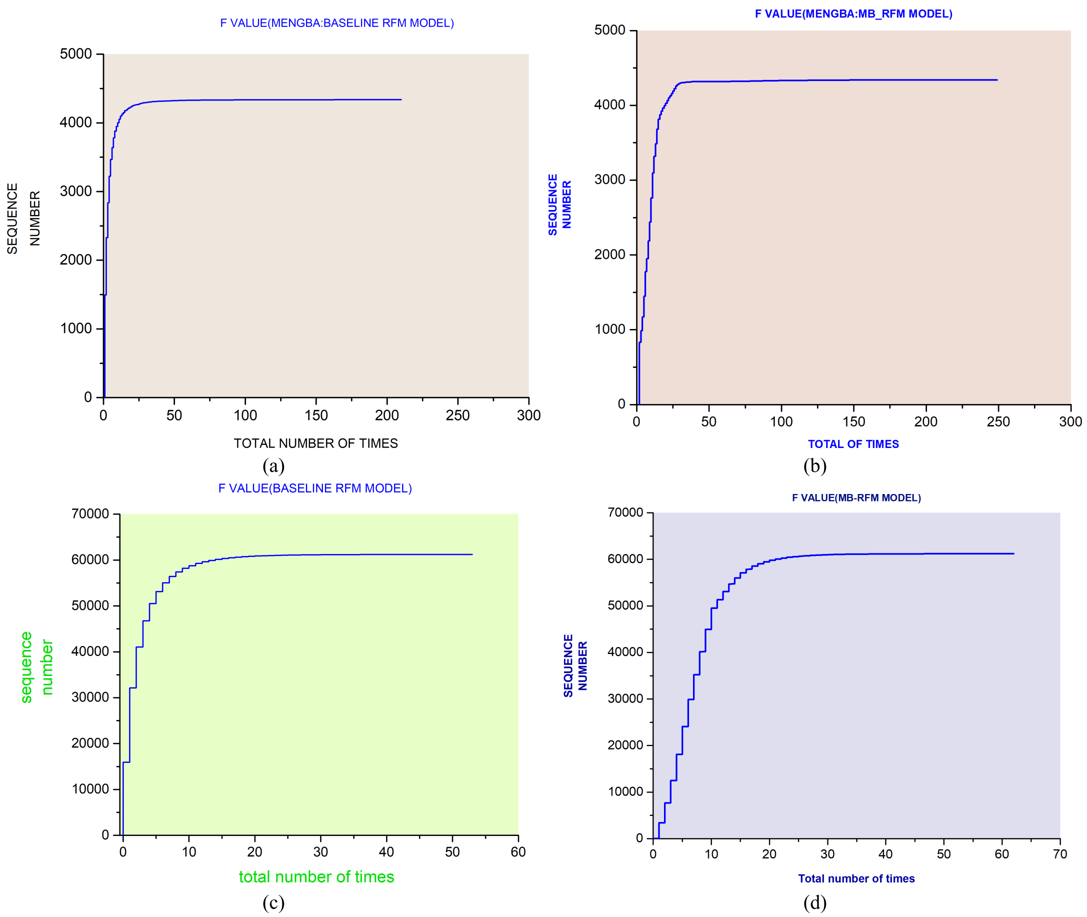


FIGURE 8. (a) Mengba dataset: F value based on Baseline RFM. (b) Yunji dataset: F value based on Baseline RFM. (c) Mengba dataset: F value based on *MB-RFM* model. (d) Yunji dataset: F value based on *MB-RFM* model.

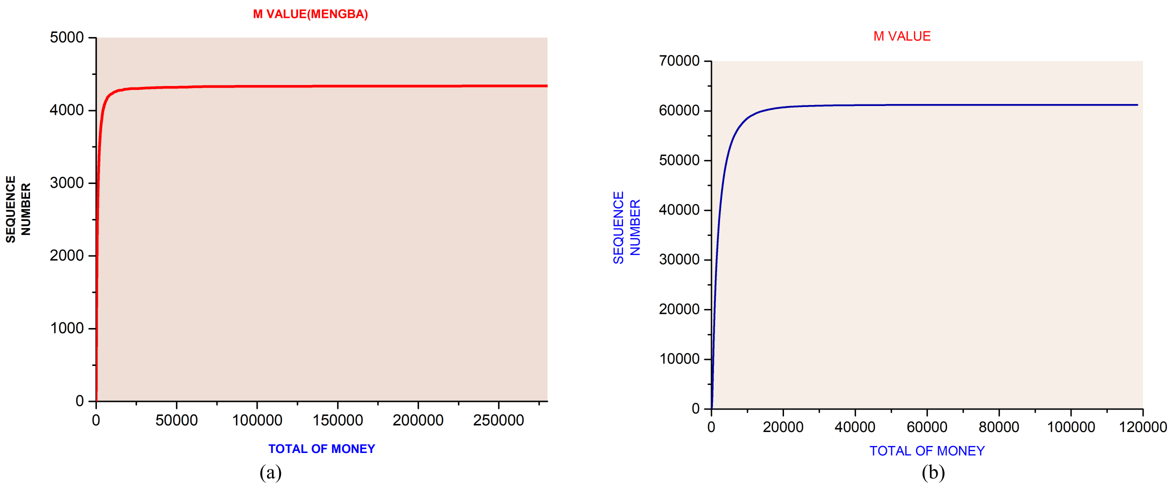
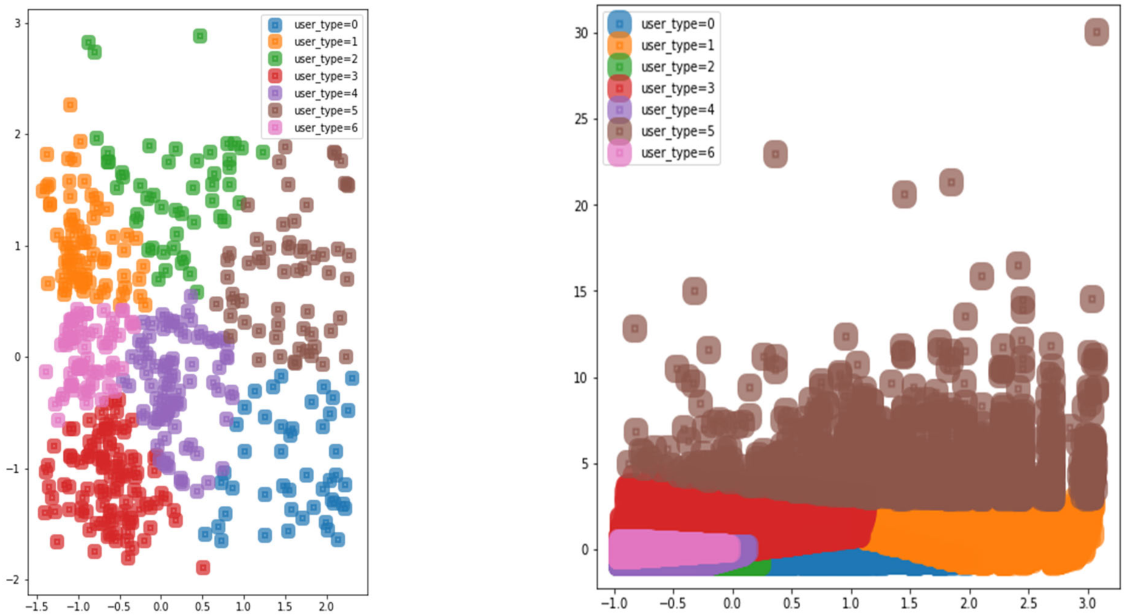


FIGURE 9. (a) Value of M for the Mengba datasets. (b) Value of M for (a) the Yunji datasets.

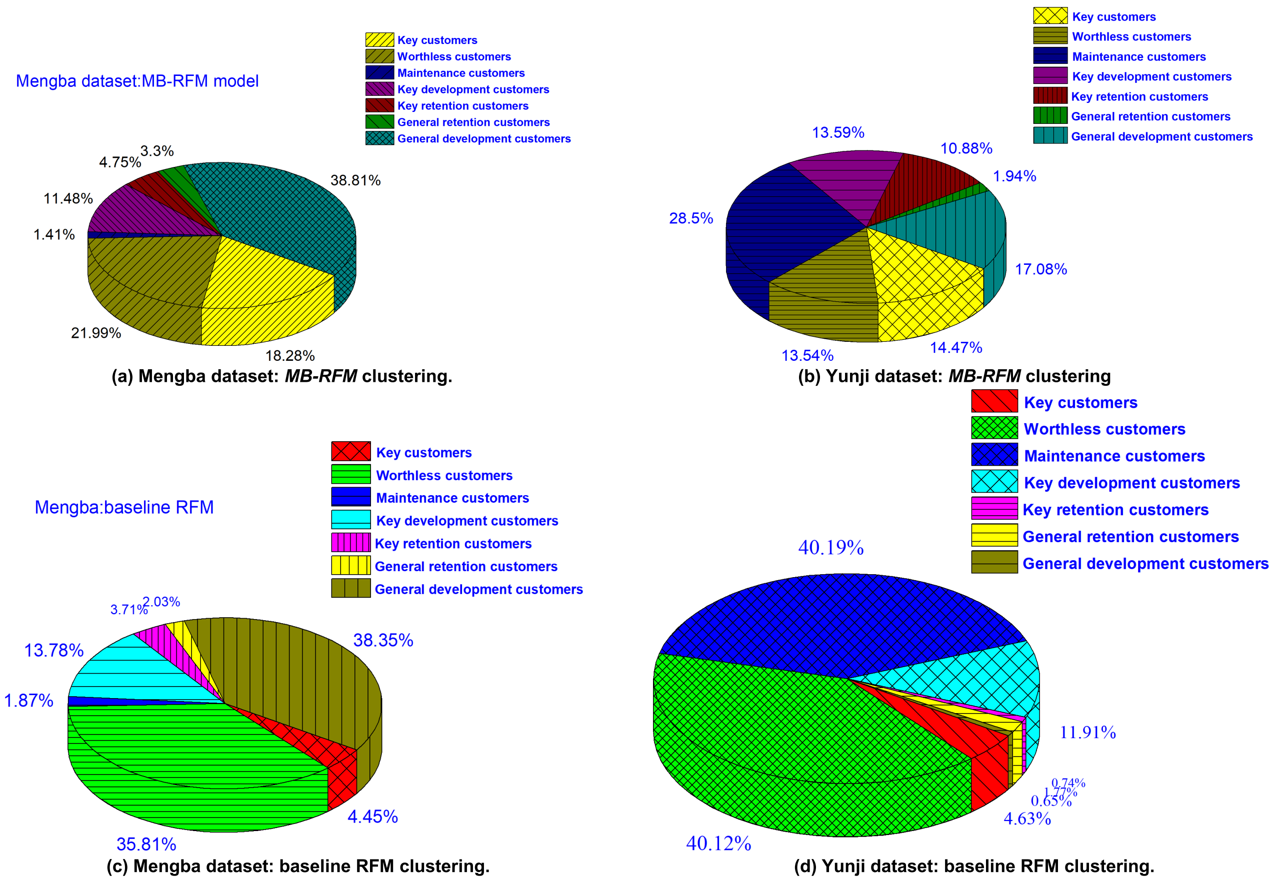
finally, enterprises should improve customer satisfaction and loyalty to ensure that this cluster can be transformed into key customers (T1).

(3) General development customers (T3). Compared to T1 and T2, T3 is characterized by moderate R, high F, and low M Such customers have a high demand for products, but their total consumption is often not high, which can be attributed to their weak consumption power. However, such customers can still bring profits to the company. According to the characteristics of such customers, enterprises should analyze customer requirements and timely recommend appropriate products via text, phone call, or social media platforms.



(a) (b)

FIGURE 10. **(a) Result of Mengba dataset clustering. (b) Result of Yunji dataset clustering.**



(a) Mengba dataset: *MB-RFM* clustering.

(b) Yunji dataset: *MB-RFM* clustering

(c) Mengba dataset: baseline RFM clustering.

(d) Yunji dataset: baseline RFM clustering.

FIGURE 11. **Customer classification results for two datasets.**

(4) General retention customers (T4). This type of customer is the complete opposite of T1, characterized by low R, F, and M. Such customers last used the application for a short period and not often; moreover, their total consumption is not high. Enterprises can consider offering discounted products to these customers regularly to increase their consumption.

TABLE 5. Trends Of R, F, And M for each cluster.

User segmentation	R	F	M
Key customers	↓	↑	↑
Key development customers	↑	↑	↑
General development customers	↑	↑	↓
General retention customers	↓	↓	↓
Key retention customers	↑	↓	↑
Maintenance customers	↓	↓	↑
Worthless customers	↑	↓	↓

TABLE 6. Results of number: the customer classifications in each category.

Type of customer	Yunji: RFM	Yunji: *MB-RFM*	Mengba: RFM	Mengba: *MB-RFM*
Key customers	2833	8856	193	793
Worthless customers	24550	8288	1554	954
Maintenance customers	24596	17439	81	61
Key development customers	7288	8317	598	498
Key retention customers	453	6659	161	206
General retention customers	1082	1187	88	143
General development customers	396	10452	1664	1684

(5) Key retention customers (T5). T5 has not only approximate cluster sizes but also the characteristics of the *MB-RFM* model, i.e., low R, low F, and high M, across 10.88% of all customers (Figure 11 (b)). Such customers used to make considerable money for the application, but these days, they trade infrequently and show signs of losing out. To solve these problems, enterprises should develop strategies such as strengthening interaction with customers via telephone or social media, designing personalized recommended products according to customer characteristics and requirements, cultivating customer identification with the application, and striving to retain customers and guide them to become key customers (T1).

(6) Maintenance customers (T6) and worthless customers (T7). These types of customers barely purchase products and do not generate much profit for the company; thus, enterprises need not invest too much in them.

V. CONCLUSION

Customer classification can help companies in understanding their customers better to make personalized recommendations and promote their applications. This could enable merchants to appropriately adopt various marketing strategies according to customer characteristics. Compared to traditional RFM models with one user-item interaction behavior, the proposed *MB-RFM* model included multiple user-item interaction behaviors based on an improved SOM neural network, which is beneficial for effective customer segmentation. We obtained transaction records (such as view, add to cart, and purchase) from the data of two local applications in China. Then, we extracted the *MB-RFM* values from the received data and calculated each weight using the superiority chart and entropy value methods. Finally, the SOM model was used to classify customers into seven categories to provide complementary marketing strategies. Type 1 represented the cluster of key customers, which should be closely considered by applications; however, Types 6 and 7 denoted underappreciated clusters that do not necessarily require investment. Type 2 denoted the key development cluster. Moreover, Type 3 included formerly profitable but complicated customers. Type 4 was the complete opposite of Type 1 because these customers rarely bought expensive commodities; applications can consider offering discounted products to these customers regularly. Type 5 was considered a regular, loyal, and valuable cluster. The proposed *MB-RFM* model could analyze and extract user-item multi-behaviors and effectively implement customer classification. Thus, the information obtained in this study can help in developing marketing strategies (such as pricing policies, promotion strategies, and personalized service strategies) for applications to improve the utilization rate of customers and targeted item promotion.

ACKNOWLEDGMENT

The authors would like to thank KetengEdit (www.ketengedit.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] (2020). *Report on E-commerce in China*. [Online]. Available: http://www.mofcom.gov.cn/
[2] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using iso-value curves for customer base analysis," *J. Marketing Res.*, vol. 42, no. 4, pp. 415–430, Nov. 2005.
[3] J. T. Wei, S. Y. Lin, and H. H. Wu, "A review of the application of RFM model," *Afr. J. Bus. Manage.*, vol. 4, no. 19, pp. 4199–4206, Dec. 2020.
[4] C. Yan and L. Lu, "Classifying non-life insurance customers based on improved SOM and RFM models," *Data Anal. Knowl. Discovery*, vol. 4, no. 4, pp. 83–90, Jan. 2020.
[5] C. Qian, M. Yang, P. Li, and S. Li, "Application of customer segmentation for electronic toll collection: A case study," *J. Adv. Transp.*, vol. 2018, pp. 1–9, Aug. 2018.
[6] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "Applying data mining and RFM model to analyze customers' values of a veterinary hospital," in *Proc. Int. Symp. Comput., Consum. Control (IS C)*, Jul. 2016, pp. 481–484.
[7] C. X. Wu, M. Y. Bao, and Q. L. Gou, "Application of self-organizing map to classification of the telecommunication company," *Comput. Syst. Appl.*, vol. 19, no. 8, pp. 168–172, Mar. 2010.
[8] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.
[9] Y. Zong and E. Pan, "A SOM-based customer stratification model," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–8, Mar. 2022.
[10] H. G. liu, Y. Xin, and Q. C. Han, "SOM neural network-based, mobile client segmentation study," *J. Microcomputer Appl.*, vol. 34, no. 23, pp. 51–54, Dec. 2015.

[11] R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, "Combining RFM model and clustering techniques for customer value analysis of a company selling online," in *Proc. IEEE/ACS 12th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2015, pp. 1–6.

[12] H. H. Wu, E. C. Chang, and C. F. Lo, "Applying RFM model and K-means method in customer value analysis of an outfitter," in *Global Perspective for Competitive Enterprise, Economy and Ecology*. London, U.K.: Springer, 2009, pp. 665–672.

[13] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, Apr. 2009.

[14] S. Monalisa, P. Nadya, and R. Novita, "Analysis for customer lifetime value categorization with RFM model," *Proc. Comput. Sci.*, vol. 161, pp. 834–840, Jul. 2019.

[15] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Marketing Customer Strategy Manage.*, vol. 19, no. 3, pp. 197–208, Sep. 2012.

[16] R. Vohra, "A using self-organizing maps and K means clustering based on RFM model for customer segmentation in the online retail business," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2020, pp. 484–497.

[17] M. Hosseini and M. Shabani, "New approach to customer segmentation based on changes in customer value," *J. Marketing Anal.*, vol. 3, no. 3, pp. 110–121, Sep. 2015.

[18] I.-C. Yeh, K.-J. Yang, and T.-M. Ting, "Knowledge discovery on RFM model using Bernoulli sequence," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5866–5871, Apr. 2009.

[19] J. Wu, L. Shi, L. Yang, X. Niu, Y. Li, X. Cui, S.-B. Tsai, and Y. Zhang, "User value identification based on improved RFM model and k-means++ algorithm for complex data analysis," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–8, May 2021.

[20] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *J. Retailing Consum. Services*, vol. 61, Jul. 2021, Art. no. 102566.

[21] J. Zhou, J. Wei, and B. Xu, "Customer segmentation by web content mining," *J. Retailing Consum. Services*, vol. 61, Jul. 2021, Art. no. 102588.

[22] L. J. Chen, Y. Song, and H. Wang, "Optimization of supporting scheme for deep foundation pit based on Delphi method and precedence diagram method," *J. Guilin Univ. Technol.*, vol. 37, no. 3, pp. 514–518, Oct. 2017.

[23] C. H. Chen, "A novel multi-criteria decision-making model for building material supplier selection based on entropy-AHP weighted TOPSIS," *Entropy*, vol. 22, no. 2, pp. 259–265, Nov. 2020.

[24] M. Liu, S. Zhang, and H. E. Yuan, "Classification study of differential telecom users based on SOM neural network," *J. Guangxi Norm Univ.*, vol. 36, pp. 17–24, Mar. 2018.

[25] L. Skanderova, T. Fabian, and I. Zelinka, "Self-adapting self-organizing migrating algorithm," *Swarm Evol. Comput.*, vol. 51, Dec. 2019, Art. no. 100593.

[26] T. Kohonen, "Essentials of the self-organizing map," *Neural Netw.*, vol. 37, pp. 52–65, Jan. 2013.

[27] H. Li and W. Wu, "Construction of 'Chinese national Geography' APP user operation strategy based on RFM model," in *Proc. 2nd Int. Conf. E-Commerce Internet Technol. (ECIT)*, Mar. 2021, pp. 396–399.

[28] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Marketing Customer Strategy Manage.*, vol. 19, no. 3, pp. 197–208, Sep. 2012.

[29] W. Liu, L. Wu, and B. Wang, "Multidimensional sparse fuzzy reasoning method based on weight of precedence chart," *J. Comput. Eng.*, vol. 35, no. 11, pp. 210–215, Nov. 2009.

[30] R. G. Martínez et al., "An RFM model customizable to product catalogues and marketing criteria using fuzzy linguistic models: Case study of a retail business," *Mathematics*, vol. 9, no. 16, pp. 1836–1845, Nov. 2021.

[31] D. L. Olson and B. K. Chae, "Direct marketing decision support through predictive customer response modeling," *Decis. Support Syst.*, vol. 54, no. 1, pp. 443–451, Jan. 2012.

[32] K. Coussement, F. A. M. Van den Bossche, and K. W. De Bock, "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees," *J. Bus. Res.*, vol. 67, no. 1, pp. 2751–2758, Jan. 2014.

[33] P. Anitha and M. M. Patil, "Forecasting of transportation cost for logistics data," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2021, pp. 1–6.

[34] R. Kumar, S. Singh, P. S. Bilga, J. Singh, S. Singh, M.-L. Scutaru, and C. I. Pruncu, "Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review," *J. Mater. Res. Technol.*, vol. 10, pp. 1471–1492, Jan. 2021.

JUAN LIAO is currently an Associate Professor with the School of Big Data Engineering, Kaili University, China. She has presided over or participated in more than ten national, provincial, and municipal projects. She has published more than 20 articles and obtained one authorized invention patent, five utility model patents, and three appearance patents. Her current research interests include development of decision support systems and big data mining technology. She won the Second Prize of Shanghai "Digital Ocean" Excellent Paper and the Second Prize of Kaili University Excellent Scientific Research Achievements. Representative results are under the background of "big data platform establishment and research of e-commerce intelligent decision-making," "the research and application of image processing technology," "big data mining technology in the application of e-commerce needs information resource extraction and research," and "study on time series similarity matching algorithm". These results have been popularized in some enterprises, some achievements have been in the national promotion.

AMAN JANTAN is currently an Associate Professor with the School of Computer Sciences, Universiti Sains Malaysia, Malaysia. He has published more than 100 articles in reputed journals. His research interests include digital forensic, artificial intelligence, malware, intrusion detection systems, computer security, cryptography, and computer and network security. He received national and international recognition in some of his work.

YUNFEI RUAN was born in Shaoyang, Hunan, China, in 1985. He received the B.S. degree from Southwest University, Chongqing, China, in 2009. He is currently an Experimentalist at Kaili University, Guizhou, China. His current research interests include intelligent detection, data analysis, and data mining.

CHANGMIN ZHOU was born in Kaili, Guizhou, China, in 1978. She received the B.S. and M.S. degrees in computer science and technology from Guizhou Normal University, Guizhou, in 2002 and 2008, respectively. She has been an Associate Professor at Kaili University, Guizhou, since 2014. She has published more than 20 articles. Her current research interests include data mining technology, data analysis, and intelligent detection and control.