HighTech and Innovation Journal

ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 4, No. 2, June, 2023

Research on Customer Relationship Segmentation of Apparel Retail Industry through Data Mining

Ning Zhu [1*]

[1] Art College, Wuxi Taihu University, Wuxi, Jiangsu 214000, China.

Received 13 February 2023; Revised 03 May 2023; Accepted 11 May 2023; Published 01 June 2023

Abstract

Objectives: This paper aims to segment customers in the apparel retail industry using data mining techniques. Methods: First, a customer segmentation model was constructed, and then the K-means algorithm was used to classify customers based on indicators from the model. The classification effectiveness was enhanced by introducing indicator feature weights. A case study was also conducted. Findings: When the value of K was 4, the K-means algorithm achieved the best classification performance. The improved K-means algorithm outperformed the traditional K-means algorithm in terms of classification effectiveness. The improved K-means algorithm categorized customers into premium customers, important customers, regular customers, and churned customers. Different marketing suggestions were proposed to manufacturers. Novelty: The novelty of this article lies in the introduction of feature weights for indicators, which allows for a distinction between their importance and improves classification effectiveness.

Keywords: Data Mining; Apparel Retailing; Customer Relationship Segmentation; Cluster Analysis; Recency; Frequency; Monetary Model.

1. Introduction

The chain industry, which has higher visibility, a stronger capital chain, a more stable supply chain, and more advanced management concepts compared to the retail industry, is the main competitor of the retail industry. Therefore, in order to attract and retain more customers, the retail industry needs to segment customer relationships and provide clearer and more targeted marketing programs for different customer groups [1]. Relevant literature on customer relationship segmentation includes the following. Götze & Brunner [2] carried out a hierarchical cluster analysis using seven scales and identified six distinct consumer groups that encompassed all types of consumers, ranging from uncompromising meat eaters to health-conscious meat eaters. Abbasimehr & Bahrini [3] represented each customer behavior as a time series of recency, frequency, and monetary (RFM) variables and then used a time series clustering algorithm for customer segmentation. The results showed that the best clustering model for grocery retailers could be achieved using hierarchical clustering with a complexity-invariant distance measure.

Sun & Liang [4] collected the experiences of dried fruit consumption from 1,160 participants in China through an online survey and categorized them into consumer groups. The study results showed that these participants could be divided into three age groups, and there were significant differences in the socio-demographic distribution among the three age groups of consumers. Singh et al. [5] proposed customer segmentation based on demographic characteristics such as gender, age, and spending score and performed a factual analysis of the dataset. The comparison of different

* Corresponding author: rtzn33@163.com; zhun@wxu.edu.cn

http://dx.doi.org/10.28991/HIJ-2023-04-02-05

classification algorithms showed that multilayer perceptron was superior to Nave Bayes and regression analysis with an accuracy of 98.33%.

Hasheminejad & Khorrami [6] used two clustering algorithms, i.e., K-means and CPSOII, to segment customers. By analyzing the dataset, they found that, compared to K-means, the advantage of CPSOII was that it could determine the number of clusters automatically. In previous studies, clustering analysis was conducted using the features of classified objects. However, this paper used clustering analysis algorithms from data mining techniques to segment customers in the apparel retail industry. Furthermore, feature weights were introduced for indicators to improve classification algorithm performance.

2. The Establishment of Customer Relationship Segmentation Model

2.1. Customer Relationship Segmentation Concept

Customer segmentation refers to constructing models based on different customer needs or preferences in order to divide customers into different groups. There are significant differences between these customers. Generally, customer segmentation can be based on three aspects: (1) external attributes such as location and organizational affiliation; (2) internal attributes such as the customer's age, income, hobbies, and credit rating; (3) consumer behavior such as frequency and amount of consumption. Retail companies can develop better marketing strategies through customer segmentation to provide personalized services that attract and retain more customers while improving the possibility of their long-term operation.

2.2. RFM Model

The RFM model [7] was originally used in direct response marketing. This model segments customers into three dimensions: R, F, and M, respectively. R is the customer's consumption interval, F is the customer's purchase frequency, and M is the customer's consumption amount. If all three values are high, it indicates that these customers have a short consumption interval, high consumption frequency, and strong consumption power; thus, they are considered high-value customers. If two of the values are high, it suggests that these customers are general-value customers. If only one of the values is high, it implies that these customers need to be retained; if none of the three values are high, it proves that these customers are potential customers [8–11]. This paper used the RFM model to establish the customer segmentation index system for the apparel retail industry based on customers' consumption behavior, as shown in Table 1.

Table 1. Customer relationship segmentation index system in the apparel retail industry

Segmentation Dimensions	Segmentation Indicators
Consumption interval (R)	The proportion of customer consumption proximity [12] among all customers (R1)
	The proportion of a customer's consumption proximity in his/her consumption proximity in the past year (R2)
Purchase frequency (F)	The proportion of a customer’s consumption times among all customers (F1)
	The proportion of a customer's consumption times in his/her total consumption times in the past year (F2)
Consumption amount (M)	The proportion of a customer’s consumption amount [13] among all customers (M1)
	The proportion of a customer's consumption amount in his/her total consumption amount in the past year (M2)

In Table 1, each sub-dimension consists of two sub-indicators. The R1 and R2 indicators in the consumption interval dimension reflect the likelihood of customer churn from both a group and individual perspective. A higher value indicates a greater likelihood of churn. The F1 and F2 indicators in the consumption frequency dimension reflect customer loyalty from both a group and individual perspective [14]. A higher value indicates higher loyalty. The M1 and M2 indicators in the consumption amount dimension reflect customer contribution to businesses from both a group and individual perspective. A higher value indicates higher loyalty.

3. Customer Classification Algorithm based on the RFM Model

The RFM model constructed above can measure the value of retail customers to merchants, and merchants can use the measurement results to develop different marketing strategies for various retail customers [15]. However, due to the large number of customers in the retail industry and the amount of data that needs processing, manual customer segmentation is challenging. Therefore, this article utilizes the K-means algorithm in data mining technology to classify customer data [16]. When classifying customer data using the K-means algorithm, it compares different customers based on segmentation indicators to determine their proximity and groups similar customers together. The issue of initial clustering data and cluster centers can be resolved through multiple clustering iterations, while the importance of feature indicators is represented by weights [17, 18]. The flow of the improved classification algorithm is shown in Figure 1.

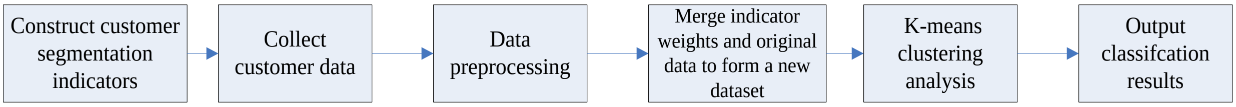
Construct customer segmentation indicators
Collect customer data
Data preprocessing
Merge indicator weights and original data to form a new dataset
K-means clustering analysis
Output classifcation results

Figure 1. The customer classification algorithm based on the RFM model

1- The RFM model is utilized to construct customer segmentation indicators, as shown in Table 1.

2- Customer data is collected based on the constructed segmentation indicators.

3- The collected data is preprocessed by standardizing it.

4- The entropy method [19] was used to assign weights to the indicators. The calculation formulas are as follows:

$$\begin{cases} y_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} \\ e_j = -\frac{\sum_i (y_{ij}\, ln(y_{ij}))}{ln(m)} \\ d_j = 1 - e_j \\ \omega_j = \frac{d_j}{\sum_j d_j} \end{cases} \quad (1)$$

where y_{ij} represents the weight proportion of indicator j of customer i, e_j is the information entropy of indicator j, d_j is the information utility of indicator j, ω_j is the weight of indicator j, and m is the number of customer. After obtaining the weight of each indicator, it is multiplied by the corresponding data to obtain a new dataset [20].

5- The K-means algorithm is used for cluster analysis by randomly selecting K data points as initial centroids [21]. The data is assigned to different clusters based on their Euclidean distance from the centroids. Afterwards, the mean center of each cluster is used as a new centroid, and the process of assigning data points to clusters is repeated until convergence.

6- The classification results are output, and then the features of each kind of data are analyzed.

4. Case Study

4.1. Data Acquisition and Processing

The research data were obtained from the transaction data of apparel retail enterprises that collaborated with schools in 2022. A total of 278,624 transaction data were available, and by aggregating according to customers' names, information for a total of 41,294 customers was obtained. The transaction data included various types of clothes in the store such as dresses, T-shirts, jeans, jackets, casual pants, and other categories. To facilitate better cluster analysis of the dataset, the data were processed as follows [22]. The first step was data cleaning. For some customer information data in the dataset, there were missing and abnormal values, which were deleted to ensure the overall validity of the dataset. The second step was to delete invalid information from the dataset, such as the customer's phone number and shipping address, as they were not pertinent to this study. The third step was to standardize data (Table 2), and the relevant calculation formulas are [23]:

$$\begin{cases} x_{ij}^{'} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} & \text{Positive indicator} \\ x_{ij}^{'} = \frac{min(x_j) - x_{ij}}{max(x_j) - min(x_j)} & \text{Negative indicator} \end{cases} \quad (2)$$

Table 2. The statistical description of different indicator data after standardization

	R1	R2	F1	F2	M1	M2
Maximum value	0.197	0.096	7.097	0.243	0.325	0.436
Minimum value	0.008	-0.032	0.027	-0.135	0.000	-0.134
Mean value	0.075	0.014	0.321	-0.001	0.013	-0.001
Standard deviation	0.041	0.018	0.272	0.014	0.012	0.011

4.2. Experimental Design

- Determine the value of K in the K-means algorithm: the K value was set to 3, 4, 5, 6, and 7, respectively. Various values of K were used to classify the dataset. Then, both average intra-cluster distance and inter-cluster distance for each classification result.

- Compare the improved-K means algorithm with the traditional K-means algorithm: The K value was set to the optimal K value obtained from the previous experiment. The dataset was classified using both traditional K-means and improved-K means algorithms. Then, the quality of the resulting classification results was compared, and the results were also analyzed.

4.3. Experimental Results

The initial setting of the K value affects the classification performance of the K-means algorithm. Table 3 shows the classification effectiveness under the setting of different K values. This paper utilized intra-cluster average distance and inter-cluster average distance to evaluate the algorithm's classification performance. For clustering algorithms, a smaller intra-cluster average distance indicates a higher level of data aggregation within the same class, while a larger inter-cluster average distance suggests greater separation between different classes, thereby indicating better classification performance. The data in Table 3 shows that when the value of K was set to 4, the average distance within each class was minimized and the average distance between clusters was maximized. Therefore, setting K to 4 achieved the optimal classification performance for this dataset.

Table 3. The classification effectiveness under different K values

K value	3	4	5	6	7
Intra-cluster distance	0.012	0.009	0.013	0.019	0.022
Inter-cluster distance	0.111	0.168	0.102	0.097	0.074

The traditional K-means algorithm and the improved K-means algorithm both set the value of K as 4 through previous tests, and the classification results are shown in Table 4. Simply comparing the specific data in the classification results does not reveal the superiority or inferiority of these two algorithms. Therefore, we compared their intra-cluster average distances. From Table 4, it can be observed that the intra-cluster average distance of the traditional K-means algorithm was larger than that of the improved K-means algorithm, indicating that the improved K-means algorithm had better classification performance.

Table 4. The classification results of two K-means algorithms

Classification algorithm	Classification number	Number of people	Average proximity/day	Average transaction frequency/n	Average consumption amount/yuan	Total consumption amount/yuan	Intra-cluster average distance
The traditional K-means algorithm	1	17418	30.7	9	6454.54	98840612	0.091
	2	8351	30.9	28	24536.25	109875846	0.132
	3	8852	56.4	11	9563.36	62354712	0.110
	4	6383	97.3	6	8229.35	35268974	0.122
The improved K-means algorithm	1	6741	35.5	24	20347.87	94758955	0.008
	2	3514	30.7	41	39874.63	57896398	0.021
	3	10601	43.4	12	11845.36	99658741	0.011
	4	20175	49.0	6	5523.35	108942256	0.009

The 2nd group, although having the smallest proportion in terms of population, exhibited the highest per capita consumption and transaction frequency, as well as the smallest consumption proximity. It can be said that this type of group represented the premium customers for retailers. Despite not reaching the same level of per capita consumption as the 2nd group, the 1st group still ranked among important customers. The 3rd group had a relatively average consumption level but possessed a larger number and development potential. Despite having the largest proportion in terms of population, the 4th group provided minimal per capita consumption, i.e., these users only occasionally engaged in transactions with retailers and were easy to lose without causing significant losses.

5. Discussion

By conducting detailed analysis of customers' purchasing behavior, preferences, and needs, businesses can gain a better understanding of the demands of different customer groups and thus provide products and services that better meet their needs. Through data mining techniques, it is possible to delve into the buying behavior, preferences, and demands of clothing retail customers in order to offer them more personalized and high-quality services.

When segmenting customers in the apparel retail industry, multiple factors need to be taken into consideration, such as basic customer information (such as age, gender, and occupation), purchasing behavior, preferences, etc. By analyzing this data, customers can be divided into different segments to better understand their needs and develop corresponding marketing strategies. However, the amount of collected data is massive and it is difficult to gather important information solely through manual methods. This article utilized the K-means algorithm for classifying collected data for customer segmentation. Additionally, weights were introduced in feature indicators during classification to enhance the effectiveness of the algorithm.

In the classification results of the improved K-means algorithm, group 2 emerges as the premium customer segment. These customers exhibit high levels of consumption and demonstrate loyalty towards the manufacturer. When dealing with this type of customer, manufacturers should prioritize delivering personalized services to minimize their churn rate. On the other hand, group 1 represents an important customer segment characterized by strong desires for consumption and a larger quantity compared to group 2. Manufacturers need to employ marketing strategies that stimulate consumption and cultivate customer loyalty. Group 3 are regular customers, whose consumption ability is at a moderate level. The total quantity also falls into the moderate range. For manufacturers, they are considered as potential customers with development prospects overall. Manufacturers can enhance their attractiveness to this type of customer by offering preferential activities. Group 4 are churned customers, which has the highest proportion in terms of quantity but provides the least amount of consumption. Their transaction behavior with manufacturers is mostly accidental. Although they are more likely to churn, their loss results in relatively small economic benefits. Therefore, manufacturers do not need to allocate excessive marketing resources for this type of customer.

6. Conclusion

The article first briefly introduced the concept of customer segmentation and then constructed an RFM model for customer segmentation. The indicators of the RFM model were utilized for classifying customers through the K-means algorithm, and feature weights were introduced to enhance classification effectiveness. A case study was subsequently conducted. The results are as follows: (1) The K-means algorithm achieved the best classification performance when the value of K was 4. (2) The improved K-means algorithm outperformed the traditional K-means algorithm in terms of classification effectiveness. (3) The improved K-means algorithm divided customers into four categories, including premium customers, important customers, regular customers, and churned customers, and different marketing suggestions for each type of customer were provided.

The contribution of this article lies in the utilization of the RFM model for constructing customer segmentation indicators and the application of the K-means algorithm for cluster analysis. By incorporating weights into these indicators, the performance of the classification algorithm is enhanced, providing effective references for the vendor's customer relationship segmentation.

7. Declarations

7.1. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.2. Funding

The author received no financial support for the research, authorship, and/or publication of this article.

7.3. Institutional Review Board Statement

Not applicable.

7.4. Informed Consent Statement

Not applicable.

7.5. Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Abdin, M. S. (2020). A study to identify and profile consumer segments in the mobile telecommunication services market. Indian Journal of Marketing, 50(5–7), 46–60. doi:10.17010/ijom/2020/v50/i5-7/152119.

[2] Götze, F., & Brunner, T. A. (2021). A consumer segmentation study for meat and meat alternatives in Switzerland. Foods, 10(6), 1–14. doi:10.3390/foods10061273.

[3] Abbasimehr, H., & Bahrini, A. (2022). An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. Expert Systems with Applications, 192, 116373 1–116373 11. doi:10.1016/j.eswa.2021.116373.

[4] Sun, Y., & Liang, C. (2021). Effects of determinants of dried fruit purchase intention and the related consumer segmentation on e-commerce in China. British Food Journal, 123(3), 1133–1154. doi:10.1108/BFJ-07-2020-0617.

[5] Singh, K. K., Singh, A., Singh, P., & Singh, N. (2020). Machine Learning based Classification and Segmentation Techniques for CRM: A Customer Analytics. International Journal of Business Forecasting and Marketing Intelligence, 6(2), 99-117. doi:10.1504/ijbfmi.2020.10031824.

[6] Hasheminejad, S. M. H., & Khorrami, M. (2020). Clustering of bank customers based on lifetime value using data mining methods. Intelligent Decision Technologies, 14(4), 507–515. doi:10.3233/IDT-190176.

[7] Paul, L., & Radha Ramanan, T. (2019). An RFM and CLV analysis for customer retention and customer relationship management of a logistics firm. International Journal of Applied Management Science, 11(4), 333–351. doi:10.1504/IJAMS.2019.103713.

[8] Hung, S. L., Kao, C. Y., & Huang, J. W. (2022). Constrained K-means and genetic algorithm-based approaches for optimal placement of wireless structural health monitoring sensors. Civil Engineering Journal, 8(12), 2675-2692. doi:10.28991/CEJ-2022-08-12-01.

[9] Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent Fuzzy Systems, 38(5), 6159–6173. doi:10.3233/jifs-179698.

[10] Doğuç, Ö. (2022). Data Mining Applications in Banking Sector While Preserving Customer Privacy. Emerging Science Journal, 6(6), 1444-1454. doi:10.28991/ESJ-2022-06-06-014.

[11] Deif, A., & Vivek, T. (2022). Understanding AI Application Dynamics in Oil and Gas Supply Chain Management and Development: A Location Perspective. HighTech and Innovation Journal, 3, 1-14. doi:10.28991/HIJ-SP2022-03-01.

[12] Alghamdi, A. (2023). A Hybrid Method for Customer Segmentation in Saudi Arabia Restaurants Using Clustering, Neural Networks and Optimization Learning Techniques. Arabian Journal for Science and Engineering, 48(2), 2021–2039. doi:10.1007/s13369-022-07091-y.

[13] Sivaguru, M., & Punniyamoorthy, M. (2020). Performance-enhanced rough k-means clustering algorithm. Soft Computing, 25(2), 1595–1616. doi:10.1007/s00500-020-05247-2.

[14] Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2021). Cluster Analysis for mixed data: An application to credit risk evaluation. Socio-Economic Planning Sciences, 73, 100850. doi:10.1016/j.seps.2020.100850.

[15] Long, V. T. (2020). Research on the influence of transportation services quality on purchasing intention of customer in e-commerce-evidence from purchasing intention of Vietnamese consumer in cosmetic industry. International Journal of Social Science and Education Research, 3(5), 45-53.

[16] Liu, Y., & Chen, C. (2022). Improved RFM Model for Customer Segmentation Using Hybrid Meta-heuristic Algorithm in Medical IoT Applications. International Journal on Artificial Intelligence Tools, 31(1), 22500099. doi:10.1142/S0218213022500099.

[17] Tripopsakul, S., & Puriwat, W. (2022). Understanding the Impact of ESG on Brand Trust and Customer Engagement. Journal of Human, Earth, and Future, 3(4), 430-440. doi:10.28991/HEF-2022-03-04-03.

[18] Tam, P. T., Son, D. M., Tan, T. Le, & Ha, H. (2021). Data Driven Customer Segmentation for Vietnamese SMEs in Big Data Era. Macro Management & Public Policies, 3(2), 33–43. doi:10.30564/mmpp.v3i2.3553.

[19] Bhatnagar, A., & Bhatia, M. (2021). A Sentiment Analysis Based Approach for Customer Segmentation. Recent Patents on Engineering, 16(2), 32–42. doi:10.2174/1872212115666210122161605.

[20] Heidari, A., Imani, D. M., & Khalilzadeh, M. (2021). A hub location model in the sustainable supply chain considering customer segmentation. Journal of Engineering, Design and Technology, 19(6), 1387–1420. doi:10.1108/JEDT-07-2020-0279.

[21] Koca, O. B. (2021). Determining customer segmentation and behaviour models with database marketing and machine learning. Pressacademia, 8(2), 89–111. doi:10.17261/pressacademia.2021.1409.

[22] Bezerra, G. C. L., de Souza, E. M., & Correia, A. R. (2021). Passenger expectations and airport service quality: Exploring customer segmentation. Transportation Research Record, 2675(10), 604–615. doi:10.1177/03611981211011992.

[23] Muhal, H., & Jain, H. (2021). Two-Stage Customer Segmentation Using K-Means Clustering and Artificial Neural Network. International Research Journal of Engineering and Technology, 8(3), 485–490.