Applying K-Means Clustering for User Profiling in Retail: A Department Store Case Study

Jiahao Huang, Pao-Min Tu(✉), Zhicheng Liu, Weisen Song, and Lijie Li

Dongguan University of Technology, Dongguan, Guangdong, China
Jiahao_wong@163.com, paomin.tu@dgut.edu.cn, jackliu32@proton.me, 2405114711@qq.com, 773882712@qq.com

Abstract. In the face of intensifying market competition, department stores are increasingly focused on understanding consumer characteristics and behaviors, as well as evaluating their value. User profiling emerges as a crucial method for comprehending customer needs and preferences, enabling the development of targeted marketing strategies to enhance customer loyalty and improve user experience. This study employs the k-means clustering algorithm for user profiling in department stores. By utilizing the Calinski-Harabasz index and the elbow method, users are grouped based on three features, resulting in optimal clustering and the division of users into four distinct clusters. Each cluster represents a unique user profile, reflecting diverse characteristics and behaviors. User profiling facilitates the understanding of target customer segments, thereby enabling the implementation of effective personalized marketing strategies. Additionally, it promotes the integration of online and offline experiences and facilitates the prediction of future demand trends. The advancements in big data and artificial intelligence technologies make user profiling an essential tool in the retail industry.

Keywords: k-means · user profiling · Calinski-Harabasz index · department store

1 Introduction

Department stores play a vital role in the retail industry, facing increasingly fierce market competition. Given the diverse consumer groups and wide product range, understanding consumer characteristics, behaviors, and assessing their value is of utmost importance. In order to adapt to evolving market demands, department stores need to delve into consumer needs and preferences, formulate targeted marketing strategies, and enhance customer loyalty and user experience. User profiling, which involves detailed descriptions and categorizations of users, aids businesses in comprehending user characteristics, needs, and behaviors.

In the context of department stores, constructing user profiles holds particular significance. Firstly, it helps businesses identify distinct user groups and develop personalized marketing strategies to improve promotional effectiveness. Secondly, user profiling assists in identifying high-value users who contribute the most to the department

S. H. B. D. M. Zailani et al. (Eds.): ICMSEM 2023, 259, pp. 1718–1725, 2024.
https://doi.org/10.2991/978-94-6463-256-9_175

store, enabling the implementation of strategies to enhance their loyalty and purchase frequency, ultimately fostering sustainable business growth.

This study utilizes Python as a tool, leveraging its excellent data processing and analysis capabilities, to perform various tasks, including data cleaning, feature extraction, and cluster analysis on user data. The k-means clustering method is employed to cluster department store users, and user profiles are constructed for each cluster to gain a deeper understanding of user characteristics, behaviors, and value assessment. The research objectives are divided into two parts: first, understanding the clustering status of department store users through clustering analysis, and second, constructing user profiles for each cluster as the basis for precise marketing and recommendation systems in future department store endeavors.

2 Related Work

2.1 K-Means

The k-means clustering algorithm is a widely used unsupervised learning technique that partitions a dataset into K distinct clusters [1–3]. Each cluster exhibits similar features, with high similarity among samples within the cluster and low similarity between clusters. The algorithm aims to determine the optimal cluster partition by minimizing the distance between each sample point and its corresponding cluster center. The k-means algorithm involves the following steps:

1.Initialization: Select K initial cluster center points, which can be chosen randomly or manually specified.

2.Assignment: Calculate the distance between each sample point and each cluster center point and assign the sample to the nearest cluster. The Euclidean distance is commonly used for this calculation.

3.Update cluster centers: Calculate the mean of all sample points within each cluster and use the mean as the new cluster center.

4.Repeat steps 2 and 3 until the cluster center points no longer change or reach a predetermined number of iterations. Each iteration involves recalculating cluster center points and reassigning sample points, leading to continuous optimization of the cluster partition. The final clustering result is chosen based on the minimum total error (loss function), which represents the sum of distances between each sample point and its corresponding cluster center point. The loss function can be calculated using the following formula:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, \mu_i)^2 \tag{1}$$

where K represents the number of clusters, C_i represents the i-th cluster, $d(x, \mu_i)$ represents the distance between sample x and the center point of the i-th cluster. By minimizing the total error J, the cluster partition is optimized to maximize the similarity within clusters and minimize the similarity between clusters. The Calinski-Harabasz index [4, 5] is often used to evaluate the effectiveness of clustering, while the elbow method [6, 7] is employed to determine the optimal number of clusters, K.

Let's consider an example with four points: (1,1), (2,1), (4,3), and (5,4). Our objective is to divide this data into two clusters by choosing K = 2.

Initialization: To begin, we select two data points as the initial cluster centers. Let's go with (1,1) and (5,4). Therefore, we have c1 = (1,1) and c2 = (5,4).

Assignment: Next, we assign each data point to the cluster center that is closest to it. In this case, (1,1) and (2,1) are closer to c_1, while (4,3) and (5,4) are closer to c_2. Consequently, we obtain two clusters: $S_1 = \{(1,1), (2,1)\}$ and $S_2 = \{(4,3), (5,4)\}$.

Update cluster centers: We then calculate the new center for each cluster, which is determined as the average of all points within that cluster. Thus, the updated $c_1 = [(1 + 2)/2, (1 + 1)/2] = (1.5, 1)$, and the updated $c_2 = [(4 + 5)/2, (3 + 4)/2] = (4.5, 3.5)$.

Repeat steps 2 and 3: The loss function can be calculated as $[(1 - 1.5)^2 + (1 - 1)^2] + [(2 - 1.5)^2 + (1 - 1)^2] + [(4 - 4.5)^2 + (3 - 3.5)^2] + [(5 - 4.5)^2 + (4 - 3.5)^2] = 1.5$. We reassign the data points to the nearest cluster center and update the cluster centers again. If the cluster centers no longer change (or the change is negligible) or if we have reached a predetermined number of iterations, we stop the process.

In this particular example, performing one additional iteration would not result in any changes to the cluster centers, signaling that we can conclude the algorithm. The final outcome consists of two clusters: $S_1 = \{(1,1), (2,1)\}$ and $S_2 = \{(4,3), (5,4)\}$, with their respective cluster centers as $c_1 = (1.5, 1)$ and $c_2 = (4.5, 3.5)$.

2.2 User Profiling

User profiling is the process of extracting, integrating and identifying the keyword based information to generate a structured profile and then visualizing the knowledge out of these findings [8–10]. This approach transforms abstract user groups into concrete entities with explicit features and behaviors, aiding businesses in understanding and serving their customers accurately.

User profiles are constructed based on multidimensional user information derived from various sources, including user behavior data from websites, apps, social media, as well as offline consumption data. The goal of user profiling is to depict authentic and comprehensive user characteristics and behavioral patterns. By understanding users' actual needs through user profiling, businesses can identify differences among users and develop personalized marketing strategies and improve the quality of products or services accordingly.

User profiling also helps identify high-value users, who contribute the most value to the company. By conducting in-depth analysis of this user segment, companies can implement effective strategies to enhance user loyalty and increase consumption frequency, thereby achieving sustainable business growth. User profiling is an essential tool for businesses in the era of big data, playing a crucial role in enhancing user experience, fostering customer loyalty, and improving competitiveness.

3 Method

3.1 Dataset

The dataset used in this study comprises a member information table and a sales transaction table. The member information table includes basic information of registered members in the department store, such as member ID, date of birth, gender, and registration time. It consists of 194,760 entries and 4 features. The sales transaction table records sales data in the department store, including both member and non-member transactions, with 1,893,532 entries and 12 features such as member ID, product code, and product price. The data spans from January 1, 2015, to January 3, 2018.

3.2 Procedure

1. Data preprocessing: Clean the data in both datasets and merge them.
2. Building member features: Select member records, consolidate purchase records for each member, and construct 12 labels representing various aspects such as age, age group, membership duration, gender, point level, number of purchases, purchase frequency, total expenditure, consumption level, value attribute, time since last purchase, and shopping preferences. These labels encompass basic user features, business-related features, and interest-related features.
3. 3.Perform k-means clustering analysis on user age, number of purchases, and expenditure. Evaluate the optimal clustering results using Calinski-Harabasz index and the elbow method.
4. 4.Visualize the clustering results using the WordCloud library.

4 Results

4.1 Clustering Results

In this study, the k-means algorithm was applied to perform clustering analysis on three features of department store users: age, number of purchases, and expenditure. The Calinski-Harabasz index (Fig. 1) and the elbow method (Fig. 2) were employed to determine the optimal number of clusters, which was found to be 4 (Fig. 3). The resulting clusters consisted of 13,404 users in Cluster 0, 15,016 users in Cluster 1, 230 users in Cluster 2, and 2,651 users in Cluster 3 (Table 1). Cluster 2 exhibited the highest number of purchases and expenditure, followed by Cluster 3. Cluster 0 and Cluster 1 had relatively lower numbers of purchases and expenditure, with the main difference being the age distribution. Cluster 0 comprised older members, while Cluster 1 consisted of relatively younger members.

4.2 User Profiles of Each Cluster

User profiles were constructed by selecting one user from each cluster and visualizing them using the WordCloud library. Figure 4 depicts a member from Cluster 0 with the user code “ae32c400,” aged 52, and a member for only 34 days. They made a single

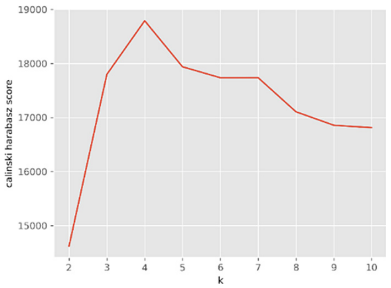

Fig. 1. Calinski-Harabasz index

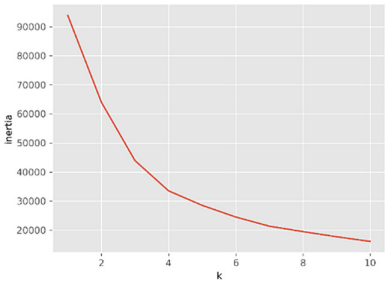

Fig. 2. Elbow method

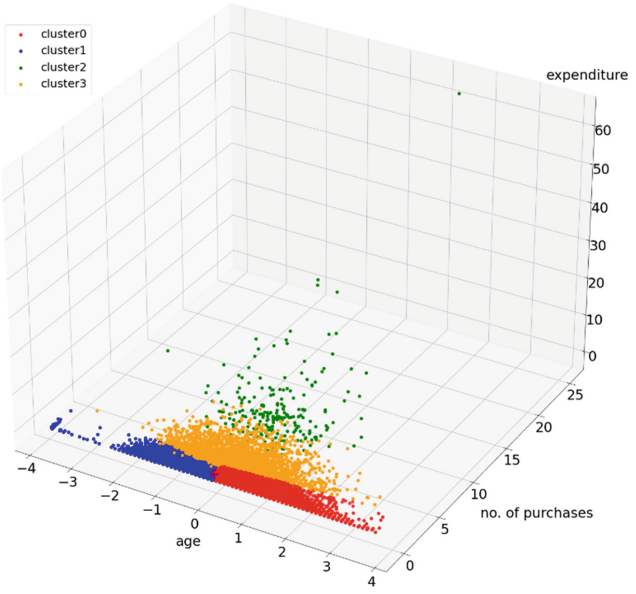

Fig. 3. Clustering result

Table 1. Coordinates of each cluster center

Cluster	No. of users	Age	No. of Purchases	Expenditure
Cluster 0	13,404	0.8536	− 0.2202	− 0.1489
Cluster 1	15,016	− 0.8062	− 0.2641	− 0.2131
Cluster 2	230	0.3722	5.9170	7.7893
Cluster 3	2,651	0.2184	2.0961	1.2843

Note: The values have been normalized

purchase with a total expenditure of 2,625 CNY. Figure 5 represents a member from Cluster 1 with the user code "003f9494," aged 27, and a member for 1,071 days. They also made a single purchase with a total expenditure of 1,220 CNY. Figure 6 represents a member from Cluster 2 with the user code "02ccbbbd," aged 28, and a member for 1,097 days. They made 81 purchases, with a total expenditure of 234,085 CNY. Figure 7 represents a member from Cluster 3 with the user code "ae80ec8b," aged 43, and a member for 1,055 days. They made 10 purchases, with a total expenditure of 65,521 CNY. These user profiles provide clear insights into the characteristics of each cluster, facilitating customer loyalty programs and targeted marketing efforts.

Fig. 4. User profiling of cluster 0

Fig. 5. User profiling of cluster 1

Fig. 6. User profiling of cluster 2

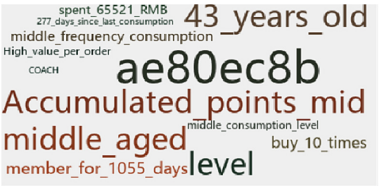

Fig. 7. User profiling of cluster 3

5 Conclusion

The k-means clustering method is employed to effectively cluster department store users, allowing for the construction of user profiles that provide a comprehensive understanding of user characteristics, behaviors, and value assessment. Through the utilization of evaluation measures such as the Calinski-Harabasz index and the elbow method, an optimal clustering outcome is achieved, resulting in the identification of four distinct clusters. Each cluster is accompanied by a user profile that emphasizes significant differences among them.

The practice of user profiling empowers department stores to comprehend target customer segments, enabling the implementation of personalized marketing strategies tailored to individual preferences. Moreover, it fosters the integration of online and offline experiences, creating a seamless customer journey, and facilitates accurate prediction of future demand trends. Given the advancements in big data and artificial intelligence technologies, user profiling has become an indispensable tool within the retail industry, contributing to improved decision-making processes and the delivery of enhanced customer experiences.

References

1. Ahmed M, Seraj R, Islam S.M.S. (2020) The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8): 1295. https://doi.org/10.3390/electronics9081295.
2. Sinaga, K.P., Yang, M.-S. (2020) Unsupervised k-means clustering algorithm. IEEE Access, 8: 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796.
3. Govender, P., Sivakumar, V. (2020) Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1): 40–56. https://doi.org/10.1016/j.apr.2019.09.009.

4. Wang, X., Xu, Y. (2019) An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. IOP Conference Series: Materials Science and Engineering, 569 052024. https://doi.org/10.1088/1757-899X/569/5/052024.
5. Wang, Y., Xu, Y., Gao, T. (2021) Evaluation method of wind turbine group classification based on Calinski Harabasz. 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2). https://doi.org/10.1109/EI252483.2021.9713300.
6. Liu, F., Deng, Y. (2021) Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems 29(5): 986–995. https://doi.org/10.1109/TFUZZ.2020.2966182.
7. Nainggolan, R., Perangin-angin, R., Simarmata, E. Tarigan, A.F. (2019) Improved the performance of the k-means cluster using the sum of squared error (SSE) optimized by using the elbow method. Journal of Physics: Conference Series, 1361 012015. https://doi.org/10.1088/1742-6596/1361/1/012015.
8. Oh, J., Sung, Y. Kim, J., Humayoun, M. Park, Y.-H., Yu, H. (2012) Time-dependent user profiling for TV recommendation. 2012 Second International Conference on Cloud and Green Computing. https://doi.org/10.1109/CGC.2012.119.
9. Kanoje, S., Girase, S., Mukhopadhyay, D. (2014) User profiling trends, techniques and applications. International Journal of Advance Foundation and Research in Computer (IJAFRC), 1(1).
10. Kanoje, S., Mukhopadhyay, D., Girase, S. (2016) User profiling for university recommender system using automatic information retrieval. Procedia Computer Science, 78: 5–12. https://doi.org/10.1016/j.procs.2016.02.002.