Research Article

A Precision Marketing Strategy of e-Commerce Platform Based on Consumer Behavior Analysis in the Era of Big Data

Di Zhang¹,² and Minghao Huang ¹

¹Graduate School of Management of Technology, Pukyong National University, Busan 612022, Republic of Korea
²College of Economics and Management, Hulunbuir College, Inner Mongolia, Hulunbuir 021008, China

Correspondence should be addressed to Minghao Huang; yyz@key.edu.pl

Received 31 May 2022; Accepted 13 July 2022; Published 17 August 2022

Academic Editor: Wenlong Hang

Copyright © 2022 Di Zhang and Minghao Huang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to develop a more efficient and accurate marketing strategy for consumers’ purchase behavior, this paper establishes a user value model by modeling and learning the user historical data of e-commerce enterprises. The improved K-means algorithm is used to cluster the purchase behavior of users, and the customer value matrix is constructed from two dimensions of consumption frequency and average consumption amount. Finally, e-commerce users are classified into four categories by marking points. The test results show that the improved K-means algorithm is stable and efficient, and the analysis of user clustering characteristics is helpful to develop more accurate marketing strategies.

1. Introduction

Due to the improvement of storage capacity of computer technology, the amount of data produced by people has also increased exponentially. In this context, the traditional business marketing model is more and more unable to adapt to the complex and changeable social environment, especially the e-commerce industry relying on the Internet. In addition, the number of customers is large, which is becoming more and more difficult to carry out marketing activities for customers. Moreover, with the continuous improvement of online shopping system, consumers’ consumption psychology and consumption form are changing, and the traditional marketing method has been far from meeting the needs of today’s consumers [1, 2]. At this stage, many e-commerce companies do not have a clear target customers in their marketing that all users are publicized and promoted in a large quantity and full coverage way, which will make users receive a large number of messages without any difference. In the long run, it may lead to the user’s resistance psychology and cause the loss of target users. This is not only a waste of resources and capital but also lead to the lack of quantitative evaluation method for the effect of marketing, so it is a general trend to introduce precision marketing based on big data technology into e-commerce industry [3, 4].

Data mining technology is popular in foreign countries in recent years, which is a trend to apply data mining technology to precision marketing. Zhu and Gao [5] studied the method of customer segmentation based on data mining. Qiu Jin et al. [6] tried to reexamine precision marketing from the perspective of the Internet. He proposed that Internet technology can be used to accurately locate customers in the implementation of precision marketing, carry out targeted marketing activities for different users, and use precision marketing to achieve the growth of enterprise benefits. Wu and Yao [7] studied the application of data mining technology in marketing activities. He pointed out that data mining technology can provide more accurate information for enterprises and has certain guiding significance for enterprises to specify marketing strategies, so as to improve the profits of enterprises. Zhao [8] and Sophie et al. [9] pointed out that under the background of big data, enterprises should pay attention to data collection when conducting marketing activities because the role of data is crucial in the era of big data and predicted the development
of data mining in the field of precision marketing in the future, while Yu [10] studied the marketing strategy based on the record data of users on the Internet in the context of big data.

The current research on the construction of e-commerce only focuses on the optimization of data processing, and few attentions are paid to the optimization of data mining processes, such as the construction and selection of data features. The traditional prediction method of user’s purchase behavior is mainly nonmachine prediction, which uses the accumulation of professional knowledge in related fields to predict the future behavior of users. The disadvantage of this method is that the subjectivity is too strong and the prediction accuracy is not high [11, 12]. In addition, the user classification obtained by machine learning lacks the corresponding marketing strategy research. Therefore, this paper establishes a new user value model by modeling and learning the user history data of e-commerce enterprises, so as to accurately locate the target user group of the enterprise, which is helpful for enterprises to achieve the goal of precision marketing. Second, through the feature construction of user historical data, the import information of each feature is obtained by using the algorithm, and the main features are reasonably explained, which can provide more effective guidance for enterprises to formulate marketing strategies for different users.

2. Precision Marketing Based on Data Mining

2.1. Data Mining of Consumer Behavior. The core of data mining is to analyze the characteristics of consumer behavior, and manage and research the consumer groups with common characteristics based on the consumer behavior characteristics obtained by data mining [13]. In the process of analyzing consumer behavior characteristics, the application of data mining technology follows the following process as shown in Figure 1:

1. Clear the subject and main research methods, combined with the needs of business development, determine the research objectives
2. According to the data collection, we choose the appropriate consumer behavior index system
3. Data preparation includes data selection and data cleaning, because there are so many consumer behaviors that can be studied, it is necessary to select the consumer behavior model according to the availability of data
4. According to different data mining objectives and the characteristics of data itself, different mining algorithms can be used and improved on the basis of the original algorithm
5. Analyze the data to prepare for the establishment of the model
6. Establish the model, analyze each index data in the model, and finally get the research results
7. According to the results of data mining, on the one hand, it is necessary to put forward specific strategies to guide accurate marketing, on the other hand, according to the feedback of consumers on the use of data mining results, continue to adjust and optimize data mining analysis methods.

2.2. Problem Description. This paper establishes a new user value model by modeling and learning the user history data of e-commerce enterprises, so as to accurately locate the target user group of the enterprise, which is helpful for enterprises to achieve the goal of precision marketing.

The traditional user value model is RFM model. While the three indicators of RFM model, R (user’s recent consumption time), F (user’s consumption frequency), and M (user’s monetary) are not applicable in e-commerce industry. Moreover, there is a natural problem in the RFM model, that is, there may be correlation between F index and M index. If a user consumes more frequently, the sum of his consumption amount will be correspondingly high. However, in the traditional model, the importance of each indicator is the same, so the explanatory power of indicators in explaining user value is not strong. Therefore, this paper establishes a new DQMC model. Among them:

- $D$ is the difference between the last shopping and the average interval between shopping;
- $Q$ is the purchase quantity of the user in the latest period of time;
- $M$ is the average monetary of users in the latest period;
- $C$ is the purchase conversion rate of users in the latest period of time.

In addition, the traditional RFM model divides the user into several categories manually and sets an interval for each category. If the value of a user falls into a certain category, the user belongs to a certain category. This classification method is greatly influenced by human subjective factors. For the same data, different users may get different classification results, which is difficult to have a standard to judge whether the model is good. In view of the above shortcomings, this paper introduces clustering algorithm to classify users. The clustering algorithm can avoid the influence of subjective factors on the results and has an intuitive and stable evaluation index for the model. The most important thing is that the stability of the model is strong. That is, for the same dataset, the same results can be obtained every time the model runs. The user classification process of e-commerce industry is shown in Figure 2:

1. Collect transaction data related to users in e-commerce industry, preprocess the data, remove the abnormal data in the dataset, and then standardize the data in the dataset after preprocessing.
2. Calculate the values of $D$, $Q$, $M$, and $C$ by data, and the value matrix of users is formed.
3. Calculate the importance of each index in the matrix by principal component analysis as the weight value of the index.
4. The clustering algorithm is used to cluster the matrix, and the corresponding user categories are obtained.
5. Use the weight $w_i$ value calculated in the third step to calculate the value of each category and each user.
(6) Use user value to verify clustering effect.

(7) For each category, according to the characteristics of users, specific marketing strategies are proposed to achieve precision marketing.

### 3. Establishment of the Clustering Model

#### 3.1. Data Preprocessing

There are two kinds of datasets used in this paper, one is user comment dataset and the other is user behavior dataset. The attribute values of these two datasets are described in Table 1.

According to the attributes of the dataset, the preprocessing steps are missing value processing, data standardization, and feature selection.

##### 3.1.1. Missing Value Processing

For this dataset, we mainly use two methods to deal with missing values. One is deletion, when the number of missing values of an attribute is less than 5%, we directly delete the missing values. The second is the mean interpolation method. If the number of missing values is too large, we will use the mean or mode interpolation for the attribute statistical indicator type.

The simple deletion method is convenient and effective and is suitable for most sample data. When the number of samples increases, the mean interpolation method can better ensure the authenticity of data.

##### 3.1.2. Data Standardization

Because some attributes in the dataset do not have obvious maximum and minimum values, it is difficult to ensure that the standardized dataset retains the information of the original dataset if the min max method is used. Therefore, the normalization method is used to standardize the dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Describe</th>
</tr>
</thead>
<tbody>
<tr>
<td>User id</td>
<td>The identity of the user</td>
</tr>
<tr>
<td>Comment</td>
<td>User comment data on products</td>
</tr>
<tr>
<td>Commodity id</td>
<td>Identification of goods</td>
</tr>
<tr>
<td>Commodity category</td>
<td>Identification of commodity category</td>
</tr>
<tr>
<td>User behavior</td>
<td>Description of the user's behavior, such as</td>
</tr>
<tr>
<td>Time stamp</td>
<td>When the user's behavior occurs</td>
</tr>
<tr>
<td>Amount of money</td>
<td>The amount of money that a user consumes</td>
</tr>
<tr>
<td></td>
<td>when he makes a purchase</td>
</tr>
</tbody>
</table>

---

#### Figure 1: Data mining of consumer behavior.

#### Figure 2: User classification.

---
Z-score method is the most commonly used, which is different from the min max-method above, and the original dataset must be normalized to zone \([0, 1]\). However, the final result is not always in this range. Suppose there is a set of sequences \(\{x_1, x_2, x_3, \ldots, x_n\}\), to normalize this sequence, the formula is as follows:

\[
y_i = \frac{x_i - \bar{x}}{s},
\]

\[\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]

\[s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.
\]

After normalization, the mean value of the sequence becomes 0, and the variance is 1, so the purpose of removing data dimension is achieved.

3.1.3. Feature Selection. In this paper, principal component analysis is used to calculate the weight of each index. Principal component analysis is a method to simplify datasets. It is impossible to obtain a series of linear correlation variables by linear transformation:

(1) Centralize the raw data, which is to make each data point subtract the mean value of its category.

(2) Solve the covariance matrix of the feature. If there are \(n\) features in the original data, the covariance matrix formed is an \(n\)-order matrix. For the four indicators \(D, Q, M,\) and \(C\) in this paper, the covariance matrix formed by these four indicators is as follows:

\[
\text{cov} = \begin{bmatrix}
\text{cov}(D, D) & \text{cov}(D, Q) & \text{cov}(D, M) & \text{cov}(D, C) \\
\text{cov}(Q, D) & \text{cov}(Q, Q) & \text{cov}(Q, M) & \text{cov}(Q, C) \\
\text{cov}(M, D) & \text{cov}(M, Q) & \text{cov}(M, M) & \text{cov}(M, C) \\
\text{cov}(C, D) & \text{cov}(C, Q) & \text{cov}(C, M) & \text{cov}(C, C)
\end{bmatrix}
\]

(2)

where \(\text{cov}\) is the covariance, and if the two variables in brackets are the same, it is the variance of itself.

(3) Eigenvalue decomposition.

In order to solve the eigenvalues and eigenvectors of the above covariance matrix, four eigenvalues will be obtained, and each eigenvalue corresponds to an indicator. The formula is as follows:

\[cw = \lambda w,\]

(3)

where \(W\) represents a matrix composed of eigenvectors and \(\lambda\) represents the column vector composed of eigenvalues.

(4) The corresponding eigenvalue of each index is taken as the weight of each index.

3.2. User Clustering Based on Improved K-Means

3.2.1. Value Matrix. Customer segmentation is usually based on the interval, frequency, and amount of consumption behavior. However, in the same time period, the two variables of consumption frequency and total consumption amount have multicollinearity, so this paper transforms the total consumption amount into the average consumption amount. Since the consumption data comes from the same shopping mall, according to the consumer’s membership ID, shopping mall backstage can get the average consumption amount through the summary of consumption records, and the data can be tracked for many times, which is conducive to the merchants to retrieve the average consumption amount in different time periods according to different needs. Moreover, more customer groups can be obtained by customer segmentation from the three dimensions of interval, frequency, and amount. Therefore, two dimensions of consumption frequency and average consumption amount are selected to simplify the result of customer segmentation, so that all customers are dispersed in the two-dimensional matrix of customer value as shown in Figure 3.

Users are clustered according to the above four indexes. User index matrix constructed in this paper regards each row as a point in 4-dimensional space. Because the size of the dataset used in this paper is small, K-means algorithm is adopted after considering the running speed and the difficulty of model interpretation.

Before clustering, the value of each user is calculated as follows:

\[V = w_1 \times D + w_2 \times Q + w_3 \times M + w_4 \times C,\]

(4)

where \((w_1, w_2, w_3, w_4)\) is the eigenvalue obtained by principal component analysis, and \(V\) is the value matrix of user. After clustering, the user value matrix of each category is calculated according to formula (4).

3.2.2. K-Means Clustering. In order to solve the above problem of how to distinguish the types of customers, in this paper, based on the value matrix, the method of changing the initial center point in K-means clustering is proposed. The initial center point is limited in different regions, that is, different types of customers are located in the value matrix, and the categories of customers are distinguished by marking points.

If customers are clustered into \(k\) class, assuming that there are a class in the abscissa and \(b\) class in the ordinate, \(a \times b = k\). Find the maximum coordinate \((x_{\text{max}}, y_{\text{max}})\) and the minimum coordinate \((x_{\text{min}}, y_{\text{min}})\). First, abscissa of the center points of different categories of regions can be obtained according to the following formula:

\[X_{ki} = \frac{x_{\text{max}} - x_{\text{min}}}{a} \cdot i + x_{\text{min}} - \frac{x_{\text{max}} - x_{\text{min}}}{a} \cdot \frac{1}{2} \]

\[= \frac{x_{\text{max}} - x_{\text{min}}}{a} \left( i - \frac{1}{2} \right) + x_{\text{min}}.
\]
Similarly, the obtained ordinates are as follows:

\[ Y_{ij} = \frac{y_{\text{max}} - y_{\text{min}}}{b} \left( j - \frac{1}{2} \right) + y_{\text{min}}. \]  (6)

Therefore, the coordinates of the center point of clustering region \( k_{ij} \) are as follows:

\[ \left( \frac{x_{\text{max}} - x_{\text{min}}}{a}, \left( i - \frac{1}{2} \right) + x_{\text{min}}, \frac{y_{\text{max}} - y_{\text{min}}}{b} \left( j - \frac{1}{2} \right) + y_{\text{min}} \right). \]  (7)

As shown in Figure 4, assuming that \( k = 4 \), then \( a = 2 \) and \( b = 2 \). According to formula (7), four regional centers \( k_{11}, k_{12}, k_{21}, \) and \( k_{22} \) can be obtained.

Then the initial center point of \( K \)-means algorithm is randomly generated in the corresponding region. This can not only ensure the effective implementation, but also distinguish different types of customers according to the positioning.

\( K \)-means clustering algorithm uses Euclidean distance to calculate the distance between each point and the center point for clustering:

1. Find the maximum value of all points \((x_{\text{max}}, y_{\text{max}})\) and minimum \((x_{\text{min}}, y_{\text{min}})\), initialize \( k \) points around \( k_{ij} \) as cluster centroids
2. Each point is assigned to the corresponding cluster according to its Euclidean distance from the centroid
3. Calculate the average value in each cluster as the new cluster centroid, and return to step (2)
4. Judge whether the centroid has changed. If it does not change, the clustering has converged; otherwise, return to step (2)
5. Output the clustering results

### 4. Experiments and Results

#### 4.1. Effectiveness of Clustering Algorithm

**4.1.1. Evaluation Index.** In order to verify the effectiveness of the algorithm in this paper, two standard test datasets, Breast cancer and Iris plants in UCI database, are selected for experiment, and the spectral clustering method commonly used in related research is used to compare with the \( K \)-means algorithm in this paper. Many literature have proposed internal quality evaluation and relative quality evaluation methods for clustering algorithm to evaluate unsupervised clustering results. However, the compactness and separation performance of these evaluation indicators are single, and the measurement effect is one-sided. Therefore, it is necessary to evaluate the clustering results comprehensively by the overall index which can measure the compactness and separation at the same time.

Silhouette coefficient \( s(i) \) considers both compactness and separation, and its calculation is shown in

\[ s(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)}, & a(i) < b(i), \\
0, & a(i) = b(i), \\
\frac{b(i)}{a(i)} - 1, & a(i) > b(i), 
\end{cases} \]  (8)

where \( a_i \) represents the average distance between sample \( i \) and other samples in its class \( C_i \). The smaller \( a_i \) is, the more sample \( i \) should belong to this category. \( b_i \) represents the minimum value in \( b_{ij} \) of the average distance between sample \( i \) and all samples in a certain class \( C_j \). The larger \( b_i \) is, the less sample \( i \) belongs to other classes. Therefore, the value of \( s(i) \) is between \(-1\) and \(1\), and the closer \( s(i) \) is to \(1\), the more reasonable the cluster is.

#### 4.2. Evaluation Results

**4.2.1. Analysis of Cluster Efficiency.** In order to classify e-commerce users more accurately, the clustering situation under different cluster numbers is discussed. The results are shown in Figure 5.

According to the elbow rule, when the number of categories increases, the average distortion degree will be lower and lower, but the improvement effect will be
smaller and smaller, so the obvious inflection point in the graph is the best value of cluster number. According to Figure 5, when the number of clusters is 5, there is an obvious inflection point. The clustering efficiency of the algorithm with different number of clusters is shown in Figures 6 and 7.

With the increase of the number of clusters, the running time of different datasets of $K$-means clustering algorithm increases linearly. After testing multiple groups of datasets, the processing time of improved clustering algorithm for 500 groups of data is maintained at about 0.3 seconds, and the contour coefficient is also maintained between 0.35 and 0.5. According to the definition of contour coefficient, the closer its value is to 1, the more reasonable the clustering is. When the data are divided into four categories, the contour coefficient is the largest. Therefore, it is reasonable to divide users into four categories.

4.2.2. Comparison of Different Algorithms. The operation of different algorithms is shown in Table 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Breast cancer</th>
<th>Iris plants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy rate (%)</td>
<td>Time (ms)</td>
</tr>
<tr>
<td>$K$-means</td>
<td>85.4</td>
<td>19.8</td>
</tr>
<tr>
<td>Spectral clustering</td>
<td>66.7</td>
<td>103.8</td>
</tr>
</tbody>
</table>

It can be seen from Table 2 that first, in terms of clustering accuracy, the accuracy rates of $K$-means on two datasets are 85.4% and 89.3%, respectively, which have good clustering effect. The accuracy rates of spectral clustering are 66.7% and 90%, respectively, and the clustering effect is more unstable than $K$-means; second, the running time of $K$-means on the two datasets is 19.8 ms and 21.9 ms, while that of spectral clustering is 103.8 ms and 118.1 ms, respectively, which can be clearly seen that the running time of spectral clustering is more than five times that of $K$-means algorithm, and the clustering efficiency is low. This is mainly because spectral clustering algorithm involves the similarity matrix between data and cannot show better accuracy and clustering efficiency when the dataset dimension is high. However, the e-commerce platform has large transaction volume, large data scale, and complex dimensions, which requires high clustering efficiency. Therefore, it is more efficient to divide the customer behavior according to the needs of e-commerce.

4.3. User Classification

4.3.1. Experimental Data. The data used in this paper come from the sales data of a clothing store on Taobao platform from November 2020 to June 2021, totaling 57666 items.
Data composition comes from two aspects: user attribute data and sales data. User attribute data include buyer’s nickname, gender, age, consumption grade, city level, college student, behavior type, and so on; sales data include buyer’s nickname, payment date, delivery date, end time, paid amount, purchase quantity, order status, province, city, and so on. In order to analyze the operation of the store more completely, 11,000 comments are collected from users of the store in 2021 through crawler.

4.3.2. Analysis of Clustering Feature. The results of cluster analysis show that it is reasonable to divide users into four categories. The value scores of various types of customers are shown in Figure 8.

It can be seen from the above table that for feature \( D \), it describes whether the user is in the normal shopping time. When the value is negative, it means that the normal shopping cycle has been exceeded. The larger the value, the more normal purchase time remaining. According to the average value of feature \( D \) in the table, users in the five categories are currently in the normal shopping cycle, but the value of the fourth class of users is large, the value of the other three classes of users is small, and the minimum value of \( D \) feature can be seen that the first, the second, and the third classes all have the phenomenon of exceeding the normal shopping cycle, which shows that the fourth class is similar to the fifth category in terms of quality.

For feature \( Q \), it describes the average number of users’ purchases in the statistical period. From the average data of feature \( Q \), we can see that the fourth category has the highest number of purchases, far exceeding the average value of the other three categories. Therefore, the fourth class of users may be loyal users of the store.

For feature \( M \), it describes the average purchase of users in the statistical period, which can be seen from the data in the table above that the values of the third and fourth categories are much larger than those of the other three categories, especially the average and minimum values.

For feature \( C \), it describes the purchase conversion rate of users. From the data in the table above, it can be seen that the purchase conversion rate of the third and fourth types of users is high, which may be loyal users.

4.3.3. Precision Marketing Strategy. By comparing the average value of each type of user index with the average value of all data indicators, it can be seen that for the first type of users, their purchase quantity and conversion rate are all higher than the average value, but the normal shopping cycle and consumption amount are far lower than the average value. This kind of users belong to potential users for stores, and they have no clear purpose in purchasing activities. In addition, this kind of users have insufficient loyalty to stores, so their shopping cycle and consumption amount are small. However, if the correct targeted marketing strategy is adopted for such users, there will be potential for them to become loyal users; for the second type of users, each index is lower than the average value of the index. This kind of users is of little value to the shops and may have lost. Therefore, the first type of users is named lost customers; for the third type of users, except for the \( D \) index is low, the other indicators are basically above the average index. This kind of users is important to the store, but there is a certain risk of loss. Therefore, this type of users is defined as the focus to recover users; for the fourth type of users, the indicators are basically higher than the average value of indicators, indicating that such users are loyal to the store and are high-value users of the store.

5. Conclusion

This paper classifies e-commerce platform customers by data mining technology, calculates the value of each type of customers, and formulates different marketing strategies for different types of customers. A new user value model is established. The improved K-means algorithm is used to cluster the purchase behavior of users, and the customer categories are distinguished by marking points. The test results show that the algorithm is efficient, and when the data are divided into four categories, the contour coefficient is the largest, so it is reasonable to divide users into four categories. Through the cluster analysis of consumer behavior, it can help the e-commerce platform to tap potential customers, maintain high-value customers, and analyze the reasons for customer churn, so as to formulate more accurate marketing strategies.

However, this paper included different categories of users for clustering, without specific statistics of the number of users in each category. This part of the research will be carried out in the follow-up work.

Data Availability

The dataset can be accessed upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.
References


