Research on E-commerce Enterprise Customer Segmentation Based on Cluster Analysis-Taking Jingdong Century Trading Co., Ltd as an Example

Yang Liu (yl931111@163.com)
Chongqing University

Research

Keywords: e-commerce, cluster analysis, customer segmentation, marketing strategy, differentiated management

Posted Date: January 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-148260/v1

Research Article

Research on E-commerce Enterprise Customer Segmentation Based on Cluster Analysis —Taking Jingdong Century Trading Co., Ltd as an example

Yang Liu[1]

1 Chongqing University, Chongqing 613000, China

Correspondence should be addressed to Yang Liu：yl931111@163.com

Research on E-commerce Enterprise Customer Segmentation Based on the Cluster Analysis

—Taking Jingdong Century Trading Co., Ltd as an example

Abstract: With the rapid development of many Internet e-commerce companies represented by Alibaba, e.g., Jingdong Mall and Pinduoduo in China, the competition, particularly among them for more customers, has intensified. However, the resources of an enterprise are limited, and it is difficult to serve all customers at the same time. Therefore, it is necessary to effectively classify customers and formulate different marketing strategies based on different customer attributes to serve different types of customers in order to maximize the interests of the enterprise. Based on the research results of the customer segmentation and the cluster analysis by domestic and foreign research scholars, this paper selects the consumption data of the company's customers based on the actual situation of the Chinese e-commerce company, Jingdong Mall, and uses the K-means clustering algorithm, Kohonen neural network clustering algorithm and the two-step clustering algorithm, as well as the improved K-means clustering algorithm, and other four clustering methods respectively conducted cluster analysis on the company's customers. Combined with the actual situation of the company, the graphical method and various test methods were used to test the clustering effect. The results show that as compared to the other three algorithms, the two-step clustering algorithm exhibits a better performance in the actual situation and the theoretical test. Then, the customers are divided into five categories, each category is separately analyzed, the characteristics and differences of each category of customers are determined, and different marketing strategies are formulated for the customer groups with different characteristics to provide suggestions for enterprises to differentiate marketing and management in order to achieve the maximum benefits with cost savings.

Keywords: e-commerce, cluster analysis, customer segmentation, marketing strategy, differentiated management

Introduction

With the increasing saturation in the Chinese market and the diversification of the needs of Chinese customers, the competition between Chinese e-commerce companies is intensifying. In order to improve the competitiveness of enterprises, the enterprises need to quickly adapt to the rapidly changing market demands and continue to take the corresponding measures in order to attract customers. Therefore, accurate market segmentation and differentiated marketing strategies are the difficulties that enterprise marketing must face at present. Customer classification is a prerequisite for market segmentation and target marketing. Therefore, how to effectively use data mining methods to segment customers is considered a very popular and important research topic in data mining applications [1]. Whether enterprises want to attract new customers, retain old customers, and identify bad customers and high-quality customers must be based on the customer

cluster analysis. The actual role of the customer clustering analysis in enterprises is mainly reflected in a previous study [2]:

1. It can help enterprises determine the characteristics of customers in order to provide them with personalized services.

2. One can discover the characteristics of customers who buy a certain product so that this product can be promoted to those customers who also have these characteristics but have not bought it, so they can cross-sell.

3. It can help find the characteristics of lost customers and take the targeted measures before these customers with similar characteristics have yet to be lost.

4. It can help identify the potential customer groups and help companies improve the success rate of market activities and acquire new customers.

The customer segmentation method was first proposed by Wendell Smith. He believes that the market has homogeneity and limited resources, so companies must operate on the basis of a limited number of customers and limited resources. In a specific mode, the enterprise functions based on the attributes of customers. Many factors, such as behavior, demand, preference, and value classify customers, and then provide distinctive products and services. According to the external attributes of customers, this classification method is considered the most intuitive, and the required data can also be easily obtained [3].

In his papers, Lazer proposed to use a customer's lifestyle as an analytical basis for identifying and judging valuable users. He emphasized the systematicity of lifestyle, but did not make a valuable statement and a specific comment on the connotation of lifestyle demonstration [4]. Wells and Tigert studied the connotation of lifestyle and proposed to use activity, interest, and opinion (AIO) to express lifestyle [5]. In 1994, Hughes proposed the RFM model to describe the characteristics of customers and classify them by using three variables: the length of time since the customer's last consumption, the number of consumption F, and the amount of consumption M [6]. In order to improve the shortcomings of the RFM analysis, Marcus proposed to use the average consumption amount A and the consumption number F to construct a two-dimensional vector customer value matrix model in order to classify customers [7]. In addition, Haley proposed a profit-based subdivision method that uses elements with causality to classify customers. The specific classification methods include the fitting analysis, factor analysis, cluster analysis, and artificial neural network [8].

Chinese research scholars have also examined this issue in response to local Chinese companies. Liu Yingzi analyzed the popular customer segmentation methods, especially in terms of dimension, time, and customer relevance, and compared their differences and explored their internal relevance [9]. Second, the main content of the dimension and subdivision technology is to maintain the customer and the profitability of the enterprise. However, the objective of the enterprise is not to lose customers but to pursue the classification of customers because customers of different qualities contribute to different values to the enterprise. Customer segmentation is considered more convenient for enterprise management and marketing. Liu Yi, Wan Difang, and Zhang Peng believe that the first step to retain customers is to implement customer segmentation. According to different contributions of different customer segments to the profit of the enterprise, targeted marketing strategies should be developed according to different customers, and different purchase behaviors should be clustered [10]. Xia Weili and Wang Qingsong believe that in the process of enterprise marketing, the most valuable customers can often provide the maximum

profit to the enterprise, and customer segmentation can be used in this process. Combined with some other marketing theories, a customer value prediction model is proposed, and a set of the evaluation system is also established according to each customer's different behavior characteristics for different evaluation and judgment [11]. Ma Huimin, Yin Hanbin, and Xiao Wei started with each specific customer, discussed in detail different performances of different customers, and discussed the true contribution of a customer to the enterprise according to the model. Finally, they provided a real case for the specific analysis and judgment and discussed the customer value [12]. Chen Qi believes that customer segmentation is an important part of enterprise management and one of the important information that enterprises rely on while making decisions. There are a large number of research studies on the consumer behavior data of the telecommunication industry. The customer segmentation model is used to analyze the behavior and the specific value of the telecommunication customers, extract the data from the database, and then partition the data in order to successfully cluster the telecommunication customers with a certain value [13]. Feng Dengguo, Zhang Min, and Li Hao summarized some key customer segmentation technologies and their latest developments. However, they also pointed out some big data concerns about personal and corporate privacy, expressing concerns about this issue as well as data. The outlook of the industry is optimistic, considering the fact that the data industry will bring the third technological revolution [14].

Since customer relationship management has been developed for a certain period of time, there have been some research studies on the application of clustering in customer management. For example, the study in [15] proposed the application of the fuzzy C-means clustering algorithm in the credit card market in Taiwan and analyzed the results. The work in [16] used the fuzzy C-means clustering algorithm in the decision model of bank management. The report in [17] used the fuzzy clustering algorithm to analyze the telecommunication market. Chuan [18] discussed the application of the fuzzy clustering algorithm in the customer management in the financial industry, and the basic idea was similar to that of calculating an example of the mean clustering algorithm in this article. The company's white paper mentions that the K-means algorithm and the system clustering method can be used for market segmentation [19]. In addition, some research scholars have proposed genetic algorithms and neural networks to deal with the clustering of market data.

At present, there are primarily the following customer segmentation methods:

1. Life stage breakdown: The subdivision method with the customer life stage as the dimension connotation strengthens the logical connection between demographic characteristics and customer needs. Similarly, the segmentation based on demographics translates the division of customer life stages. In practice, two methods are used [20]: the first is the PeopleUK Method, which divides the customer's life course into eight development stages, and each stage is divided into different sub-stages, i.e., a total of 46 categories. The second is Claritas, the PriZm method developed by the company, which divides all customers in 16 large customer groups into 60 different groups, covering lifestyle, four life stages, and four income levels. At this point, the center of gravity of the demographic breakdown begins to shift toward lifestyle.

2. Lifestyle breakdown: The study of lifestyle dimensions also stems from a hypothesis—the more you know about customers, the more effective marketing you can formulate for them. Lazer (1963) first proposed to identify and segment customers based on lifestyle [21]. Lifestyle is considered a systematic concept, embodied in various ways and derived and developed from social life mechanisms. Despite emphasizing the systematic nature of lifestyle, it did not regulate

its connotation until Wells and Tigert (1971) proposed to use AIO, namely, activity, interest, and opinion, to express lifestyle [22]. Subsequently, Plummer (1974) used demographic characteristics to enrich the lifestyle, and the connotation of this dimension expanded into four dimensions [23].

3. Breakdown behavior: The RFM analysis proposed by Hughes (1994) uses three behavioral variables for describing and distinguishing customers [24]. In order to avoid the shortcomings of the RFM analysis, Marcus (1998) proposed to construct a two-dimensional customer value matrix model by using the number of purchases F and the average purchase amount A in order to modify the RFM method [25].

4. Interest breakdown: The profit segmentation was first proposed by Haley (1963). It uses causal factors rather than descriptive factors for identifying the market [26]. Its advantage from the traditional segmentation methods lies in its behavior through customer appearance, attitude, and motivation to gain the real benefits. Vriens and co-workers (1996) summarized the commonly used methods in the fitting analysis and expressed the characteristics of these techniques from the following three aspects: segmentation type, segmentation program, and optimization criteria [27]. Kim and Mueller (1978) summarized the use of the basic theory of customer segmentation by applying the factor analysis, including identifying those structures or "factors" that can explain the implicit relationship between a series of variables, hypothesis testing about the structure of the variables, summarizing the initial large number of variables with a few derived variables, and determining the number of dimensions of the variable to be expressed [28].

In this article, e-commerce enterprise customers are primarily subdivided based on cluster analysis methods. According to the real data of a large-scale e-commerce company in the China-Jingdong Mall (already listed in NASDAQ), the K-means clustering and two-step clustering, Kohonen neural clustering and improved K-means algorithm and other four methods are used to classify Jingdong Mall customers. At the same time, the graphic method and the internal inspection method are used to evaluate the classification effects considering different methods. Finally, the classification results are compared and analyzed, the most appropriate classification method is determined, the corresponding customer characteristics and the customer value for the classification effect are analyzed and summarized, and different marketing strategies are proposed in order to provide support to the company's subsequent operation strategy.

I method

The process of clustering is defined as the process of increasing the similarity of members in the same group and decreasing the similarity of members in different groups. It is known as an unsupervised learning process. It is very different from the process of classification. The biggest difference is that the characteristics based on the classification are known before clustering begins, and clustering is like a data preprocessing process. The process of clustering is used in many fields, such as machine learning, biology, and marketing. It can also be used as a preliminary processing step for other data mining algorithms. Generally speaking, the cluster analysis can include the following methods: hierarchical method, partition method, grid method, and model method [29].

1. Define the similarity of data

The key role of the cluster analysis is to describe the distance between sample points and classify them according to the distance between samples. In this paper, the standard Euclidean distance metric is used.

The distance between $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ is calculated as follows:

$$d(x,y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The Euclidean distance between two points on the two-dimensional plane is determined as follows:

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

The Euclidean distance between two points in the three-dimensional space is measured as follows:

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

If there are two vectors, then the Euclidean distance between them is expressed as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2}$$

2. K-means

The K-means algorithm is one of the most popular algorithms in clustering algorithms at present. It has many advantages, such as fast convergence speed, relatively simple algorithm, and is found suitable for large datasets. The basic concept of the algorithm is described below.

Assume that there are *n* data objects in the dataset and the number of clusters given in advance is *k*, then *k* initial cluster centers are randomly selected in the dataset, the distance from other samples in the dataset to the initial center is calculated, and according to the distance, the data samples are grouped into the initial point closest to them as a class. Next, the new cluster center is calculated, the calculation of the distance from other data points to the cluster center is repeated, and the previous steps are repeated. If the cluster center is twice within the specified threshold, then the clustering ends or the next iteration is performed. Eventually, the clustering center converges around a fixed value, and no big changes are observed. This is one of the bases for measuring the end of the algorithm.

From the above algorithm process, we can draw the following conclusion: the clustering effect of the clustering algorithm partly depends on the selection of the initial point of clustering. If the initial cluster centers are selected differently, the results of clustering may also be very different; there is no fixed method for the selection of cluster centers based on previous experience, and the experience of people facing uncertain problems is very often lacking. Therefore, the selection of clustering centers has become a burning topic in recent years.

The K-means algorithm has the advantage of fast operation, and it is also simple and easy to understand. It has become a very popular algorithm at this stage, but it also has some significant disadvantages, which are as follows:

①The initial clustering center is difficult to select. As mentioned above, the quality of the clustering center directly depends on the clustering effect, and the clustering center is difficult to understand without experience.

②The number of clusters must be specified in advance when using this algorithm, which is often difficult to do.

③There is a requirement for the shape of the data, and it is sensitive to noise. The algorithm

does not use datasets with too large sample differences, and it is found to be very sensitive to the sample noise. If there are too many noises and outliers in the sample, it may not be considered suitable to use the K-means algorithm to solve the clustering problem [30].

3. Optimization of the K-means algorithm: Kohonen network clustering

The concept of the self-organizing feature map was proposed by T. Kohone in 1981, who developed the corresponding neural network model. Because the model imitates the self-organizing process of the brain in the organism, it is often used to imitate external information in the organization, which is called the process of automatic formation of concepts [31].

The clustering process of the Kohonen network uses its self-organization, spontaneous, unsupervised completion, mapping and one-dimensional sequences to two-dimensional sequences, which can adjust the parameters between each neuron. The clustering process can be decomposed into a process of adjusting parameters between competitions among neurons and a learning process between competing neurons.

Neuron competition learning process:

Once a vector is input, through the comparison between the input vector and the randomized weights, we can determine that the neuron whose weight vector is closest to the input vector is marked as the winning neuron, which is called the image of the input neuron. If the input vectors are the same, then the images in the competitive layer are also the same [32].

Competitive layer neuron's side feedback process:

There is also a mechanism of adjusting the weight of the neurons in the competition layer, which is known as the side feedback process and follows the below-described rules:

1) Excite side feedback to neighboring neurons around the winning neuron.

2) Display consistent side feedback around the winning neuron for neurons that are further away.

In the above process, we can find that in the competitive layer, around the winning neuron, the neighboring neurons are mutually stimulated, the farther neurons are mutually suppressed, and the farthest neurons are weakly mutually excited.

Once the above process is completed, a clustering area forms near each winning neuron. The objective of this process is to always retain the neuron vectors on the competitive layer approaching the input vectors so that the input vectors with similar characteristics are clustered together. In this way, an autonomous and unsupervised clustering process is realized.

The algorithm can be divided into the following steps:

A. Assign a small random value to each neuron in the network.

B. Suppose the input vector in the network is denoted by

$$X = [X_0, X_1, ..., X_{\text{n-1}}]$$

The weight vector of the neuron becomes

$$W_i = [W_0, W_1, ..., W_{N-1}]$$

Among them,

$$I_i = \sum_{j=o}^{N-1} \left\| X_i - W_{ij} \right\|, \forall i$$

with $I_c = \min\{I_i\}$. Then, the neuron corresponding to c is known as the winning neuron.

C. Determine the weight adjustment formula as follows:

$$W_{ij}(t+1) = W_{ij}(t) + a[X_i - W_{ij}(t)]$$

D. Enter the next vector and repeat the process.

In this way, after continuous training and adjustment, the weight of the links between each neuron will be continuously optimized so that the similarity between the vectors is mapped with the neuron, and a strong homogeneity is observed between the neurons. Different types of neurons show strong heterogeneity. In order to keep homogenous neurons close together, it is necessary to continuously reduce the neighborhood radius and the learning rate during the training process.

The advantages of this algorithm are that it demonstrates very good biological characteristics and can automatically complete the clustering of the input vector. The main reason to resolve the shortcomings of the K-means algorithm is that it does not need to specify the number of clusters in advance and can use it first. In order to determine the approximate number of clusters, the K-means algorithm is used for clustering in order to avoid inaccurate clustering caused by blindly selecting the number of clusters.

4. Two-step clustering

The two-step clustering method is a clustering algorithm that has emerged recently. As compared to other traditional clustering algorithms, it shows the following significant advantages:

①Both the categorical variable clustering and the continuous variable clustering exhibit good performance.

②The number of clusters can be automatically determined.

③This Method can manage large data files.

The two-step clustering algorithm primarily uses the measure of likelihood clustering to cluster categorical variables and continuous variables, so it must meet certain assumptions: continuous variables need to meet normal distribution, whereas discrete variables need to meet polynomial distribution, but in the actual operation, the two-step clustering is a very robust model, so the requirements for assumptions are not very strict. The two-step clustering algorithm is primarily divided into the following two steps:

A. Before clustering, we must first establish a clustering feature tree. Generally, the first data in the dataset is placed on the leaf node of the tree root, that is, the root node. This node contains all the variable information related to this data. In the following steps, each subsequent record calculates the distance measurement as a measure of similarity of the distance. Next, they are added to an existing node or a new node is formed according to the similarity. If a node contains data that is needed, it means the information on this node is similar, and the records on this node can be extracted [33].

B. On the basis of the pre-clustering, the hierarchical clustering is generally used to merge small categories into larger clusters. That is by integrating the data on each node; this method can produce a clustering effect on a group of different orange genus. Generally, when determining the best clustering scheme, AIC or BIC can be used for comparison and selection.

5. Evaluation index

In many clustering algorithms, many indicators are used to evaluate the quality of clustering, both objective and subjective. In this article, the more commonly used aggregation and separation contours, namely, the Silhouette coefficient, are used as indicators for evaluating the clustering effect.

The calculation method follows the below-described steps:

①Calculate the average distance $a(i)$ of the sample data i to other samples in the same cluster. If the index is smaller, it means that the sample is more reasonable to be classified as the cluster.

②Calculate the average distance of the sample data i to all sample points of other clusters. If there are k clusters and sample i is in the first cluster, then $b_i = \min(b_{i1}, b_{i2}, \ldots, b_{ik})$.

③The contour coefficient of sample i is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

④Calculate the average value of all samples, where $s(i)$ is the Silhouette coefficient.

⑤The value range is $[-1,1]$. If the Silhouette coefficient is closer to 1, it means that the clustering effect is better.

II Data and descriptive analysis

1. Data

In this article, the research data is taken from the actual desensitization data of Jingdong Mall users. A total of 15,000 basic consumption data are randomly selected from this database. The data object is Jingdong student users, the time span is one and a half years, and the starting time is June 1, 2018. The end date is December 31, 2019. According to the actual situation, we use the K-means ordinary clustering algorithm, optimized K-means clustering algorithm, neural network clustering algorithm, and two-step clustering algorithm to cluster customers so that the company's operating personnel can more easily understand the composition of customers. It becomes convenient for the company to carry out effective customer identification, precision marketing, and other business activities and strategies. Due to numerous consumption data, strong data privacy, and company secrets, the 15,000 data selected contain basic consumption data, such as customer number, purchase amount (the actual number of payments after discounts are removed), the number of purchases during the time period, and several other variables such as the number of days since the last purchase are due.

Before modeling, it is necessary to check the quality of the data. The first point to verify is whether the data is missing. After verifying the data, a total of 372 missing values are observed in the overall data. Average interpolation and linear interpolation are found while dealing with the missing values. However, as the amount of data is too large, one can simply delete the missing value and then cross-check the outliers. In terms of sales, due to a large amount of data and the actual situation, a huge difference is observed in the purchase amount; it can be ignored. In terms of the number of purchases, considering the actual situation, the data with the number of purchases greater than 400 is deleted. In terms of the time interval, the data greater than 400 is deleted considering the actual situation. Once the data processing is completed, the subsequent analysis processing is performed, and the following details are collected: no missing values are found, there is no data that does not conform to common sense, and no impossible outliers are found.

2. Descriptive analysis

Once the preliminary data preprocessing is completed, the data integration work is performed.

Several fields, such as customer number, sales, cumulative purchases, and the distance between the last consumption and the deadline, related to customer consumption, are selected in the database. During the process, it was found that the sales volume and the number of purchases are most relevant, so the average purchase amount is recalculated, that is, the value of a customer's total sales divided by the cumulative number of purchases so that the customer number and the cumulative purchases can really enter the model. From the average purchase amount and several fields from the purchase time interval, in order to remove the sensitivity of business data, the customer number was recorded, and the purchase amount was retransformed by the same amount. Once the data sorting is completed, in order to have a general understanding of the data of several fields, Table 1 shows a descriptive statistical analysis of the cumulative number of purchases, average purchase amount, and purchase interval.

Table 1. Descriptive analysis.

	N	Max	Min	Mean	Std	Median	Skewness	Kurtosis
Cumulative purchases	15000	109	1	31.08	23.21	25	0.79	-0.4
Average consumption	15000	536	1	153.35	113.48	126	0.75	-0.44
time interval	15000	300	10	106.82	86.27	75	0.76	-0.75
Valid N	15000							

From the data shown in Table 1, we can obtain general data characteristics. The difference between the maximum value and the minimum value of the customer's cumulative purchases reaches 100, which indicates that the customer loyalty is polarized, the average value is not much different from the median, and the absolute value of the skewness does not exceed 1, suggesting that the data generally shows a bilateral symmetry and average. A similar situation is observed for secondary consumption. The data is very different and displays a symmetrical structure. The time interval data can also explain this situation. The kurtosis of the three-field data is found to be close to the normal distribution.

III Results and discussion

1. K-means customer clustering

As discussed above, when using the K-means clustering, one requires achieving the specified number of clusters. As we know very little about the data, the following steps are used to perform the K-means clustering:

①Specify the number of clusters of 3, 4, 5, 6, and 7, respectively, in order to achieve a better clustering effect.

②After specifying the number of clusters, as the random initial point of the algorithm is uncertain, we randomly select the initial point 50 times at every instance and select the one with the smallest Silhouette coefficient among the 50 times.

③Compare the clustering effects of 1–5 cluster numbers.

The clustering effect is presented in Table 2.

Table 2. Clustering quality.

Number of clusters	3	4	5	6	7
Cluster quality	0.7	0.68	0.67	0.62	0.57

In order to understand the classification effect more intuitively, the three-dimensional

rendering after classification is shown in Figure 1.

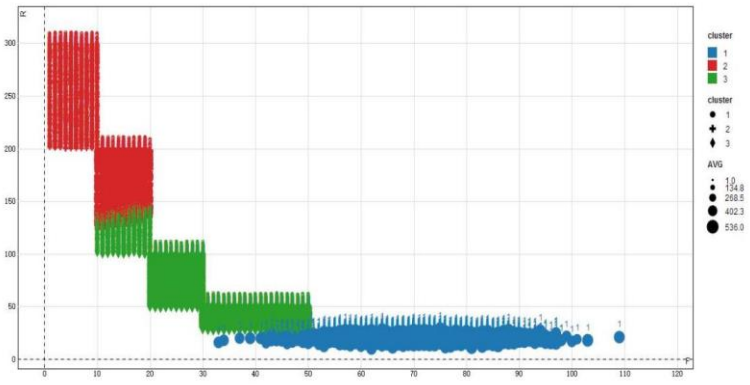
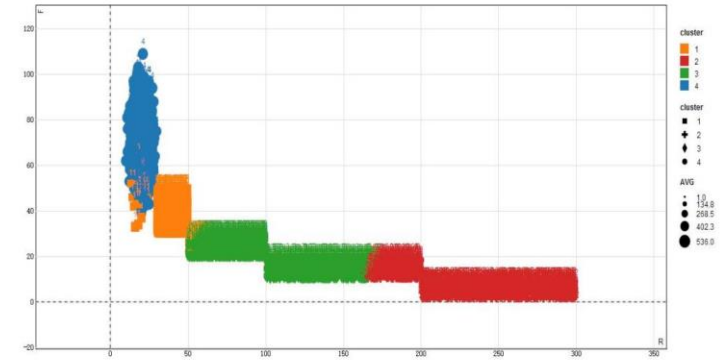


Figure 1. Clustering effect.

Figure 1 shows the three-dimensional rendering when the number of clusters is 3. From the figure, it can be clearly seen that the data is roughly divided into five blocks, and the number of clusters selected three types, which is failed to clearly separate the data. As shown in the classification, the first type is the blue category that is located in the area below the image, which is basically separated from the overall data, and the second type is the green category, which is located in the middle area. The data in the middle area obviously does not belong to one category. The third category is similar to the above situation.

We further try to cluster the data into four categories, as shown in Figure 2.



:

Figure 2. Clustering effect.

When the data is clustered into four categories, the effect is obviously much better than the three categories. The blue category, the red category, and the orange category are completely separated out, and they are highly consistent with the original data block. The green classification effect shown in the figure is not considered ideal. There are also cases where the original data blocks do not match.

Next, we try to divide the data into five categories as shown in Figure 3.

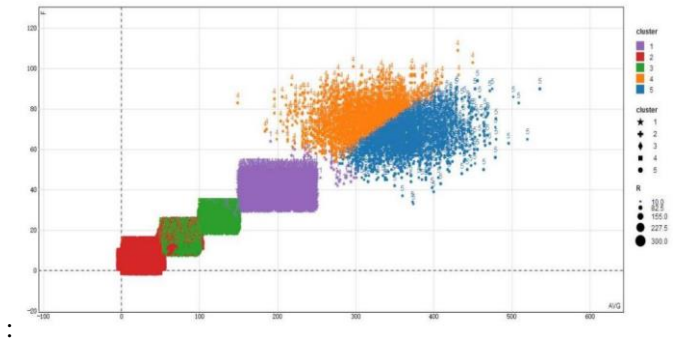


Figure 3. Clustering effect.

The effect of clustering into five categories is roughly the same as that of three categories, and the data located on the upper right of the image cannot be completely separated.

In summary, combined with the Silhooette coefficient and the three-dimensional aggregation graph comprehensive evaluation, it is observed that the clustering effect is the best when the number of clusters is four, and the data in the real data cluster can be basically separated.

In this process, the average value of the three indicators in each category and the percentage of each category in the total are shown in Table 3.

Table 3. Clustering review.

Category 1	Category 2	Category 3	Category 4
AVG	AVG	AVG	AVG
199.34	37.03	104.20	341.79
F	F	F	F
39.76	7.70	20.80	70.23
R	R	R	R
40.07	234.14	100.04	20.00

2. Two-Step customer clustering

As discussed above, when the two-step clustering is performed, two main parameters are considered. One is the measurement of the sample data distance, which includes the logarithmic similarity measurement, and the other is the Euclidean distance measurement. There are generally two clustering criteria: Schwarzer Bay The Yers criterion, and the Akaike information criterion, which are known as the BIC criterion and the AIC criterion. The Silhouette coefficients obtained by combining these two criteria are shown in Table 4.

Table 4. Clustering quality.

		Distance type	
Clustering criterion		Logarithmic similarity	Euclidean
	BIC	0.7	0.7
	AIC	0.7	0.7

As can be seen from Table 4, the Silhouette coefficients are the same for all. Next, we compare the three-dimensional clustering effect map to determine which clustering effect is more ideal.

When using the Euclidean measurement as a distance measurement standard, the number of clusters in the model is 3 for all types. Figures 4 and 5 show the three-dimensional clustering diagram for all types.

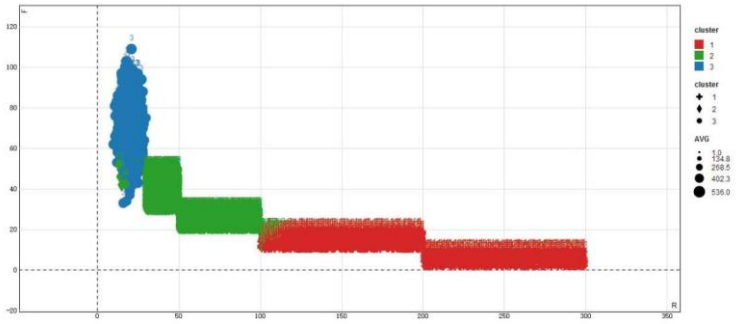


Figure 4. Clustering effect.

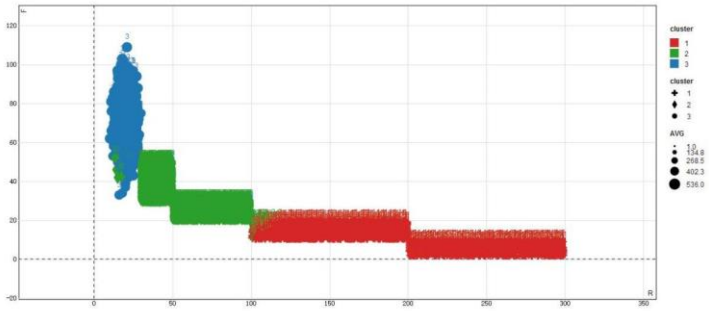


Figure 5. Clustering effect.

From the above images, it can be seen that when clustered into three categories, the data on the upper left can be basically separated, whereas the clustering effect on the lower right is very poor, which indicates that the clustering effect is not achieved.

Figures 6 and 7 show the three-dimensional clustering effect diagram by using the logarithmic similarity distance measurement.

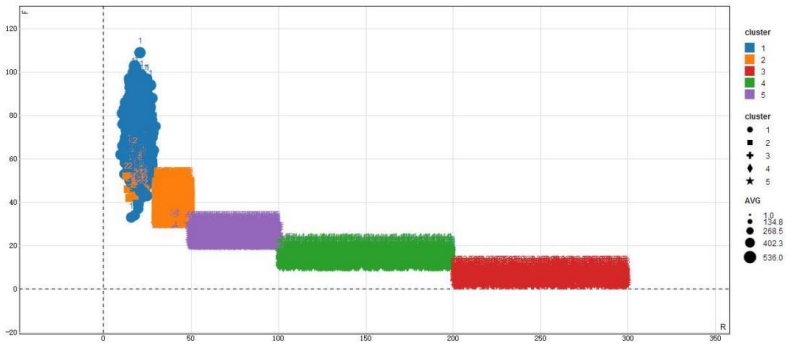


Figure 6. Clustering effect.

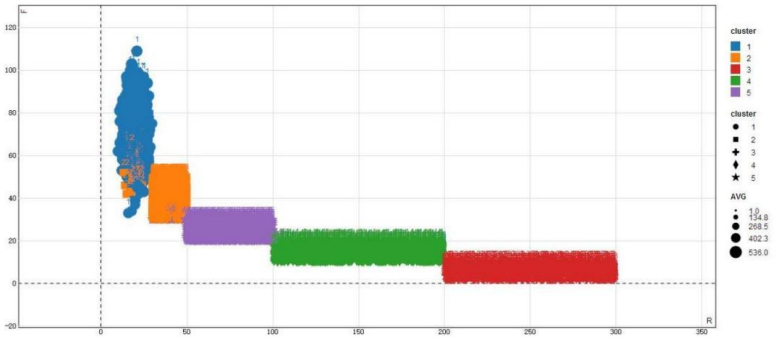


Figure 7. Clustering effect.

From the above images, it can be seen that when the logarithmic similarity is used as the measurement distance, five types of clustering are observed. The four types of data on the lower right are completely separated, and only a small part of the data on the upper left overlaps. Therefore, the clustering effect is considered ideal.

In Table 5, the logarithmic similarity is used as the distance measurement value, and the BIC criterion is used as the clustering criterion. The average value of the three indicators in each category and the percentage of the total for each category are shown in the Table

Table 5. Clustering review.

Clustering	Category 1	Category 2	Category 3	Category 4	Category 5
	AVG	AVG	AVG	AVG	AVG
	341.75	201.87	25.72	74.67	125.35
	F	F	F	F	F
	70.19	40.29	5.49	14.98	25.01
	R	R	R	R	R
	20	39.53	249.81	149.91	74.76

3. Kohonen customer clustering results

Kohonen is a process that imitates the activity of biological neurons. It can automatically train the weights according to the input data in order to achieve the clustering effect. The process does not need to use parameters in advance. The training process is roughly divided into two stages.

In the first stage, the data distribution is collected. In this stage, the model uses a relatively large neighborhood radius and the initial learning rate. The objective is to capture the rough pattern of the data.

The second stage is called the weight training stage. In this stage, a smaller neighborhood radius and the initial learning rate are used in order to find more subtle distributions and clusters in the data.

The setting of the neighborhood radius and the initial learning rate is realized and explored many times, and generally it is not achieved overnight. After selecting a relatively suitable domain radius and a suitable initial learning rate for the data distribution, after earlier explorations, the following steps are followed:

①Select a relatively large radius and the learning rate for the first time so that a setting may not conform to the data distribution and decrease in the next parameter setting.

②Next, according to the setting of the first step, reduce the size of the parameters in sequence by 0.05 each time and observe the clustering results.

According to the above method, the clustering results are as follows.

For the first time, we set the neighborhood radius of the first stage to 2 and the initial learning rate to 0.3. The neighborhood radius of the second stage was set to 1, and the initial learning rate was set to 0.25. The results show that the Silhouette coefficient is 0.3, the clustering quality is very poor, and it is almost difficult to identify different categories based on the clustering results.

For the second time, we set the neighborhood radius of the first stage to 2 and the initial learning rate to 0.25, and the neighborhood radius of the second stage to 1 and the initial learning rate to 0.2. The results show that that data is clustered into 47 categories, and the ratio is much better than the last time, but the effect is still not satisfactory.

By analogy, the clustering is continuously tried, and the results are as follows.

Table 6. Clustering results.

First stage initial learning rate	Second stage initial learning rate	Cluster quality	Number of clusters
0.3	0.25	0.3	34
0.25	0.2	0.3	18
0.2	0.15	0.4	9
15	0.1	0.3	17
0.1	0.05	0.4	26
0.05	0	0.3	40

From Table 6, it can be observed that the best cluster quality is 0.4, and the number of clusters is grouped into nine categories. Although on the premise of trying different parameters in 6 categories, the Kohonen clustering still cannot effectively separate the data, and the clustering effect is found to be poor.

At the same time, we observed that in the Kohonen clustering the number of clusters with the best clustering effect is 9, which is much larger than the number of clusters of other methods. This clustering method is too delicate, resulting in too many customer classifications. Therefore, it is concluded that too much chaos is not conducive for the enterprise to assemble valuable customers in order to implement precise marketing for them, so the Kohonen clustering is not found very effective in this kind of situation.

4. Improved K-means clustering attempt

From the above discussion, it can be inferred that the traditional K-means algorithm shows the clustering uncertainty and large errors caused by the randomness of the initial point selection. The number of clusters must be given in advance, and it consumes a lot of memory. First, we use the Kohonen network clustering to briefly describe the general situation of the dataset, find the initial clustering center and the number of possible clusters in the dataset, and then we use K-means to continue clustering, avoiding the blind selection of the initial cluster. For the uncertainty due to the number of points and clusters, The below-listed steps are followed:

①By using the Kohonen network clustering, first determine the initial clustering center and the number of clusters.

②Fix the clustering center before the K-means clustering and check in turn whether each data point is classified into a certain category.

③If the sample points are classified into one category, ignore the sample points, otherwise calculate the distance from the sample point to the center of each category, find the minimum distance, and then classify it into that category.

④Repeat the previous steps (steps 2 and 3) until the cluster center no longer changes.

According to the above algorithm, we first select the best one from the Kohonen network cluster, that is, the one with the number of clusters of 9 is used in the K-means cluster as discussed above. The overall structure does not match and the classification is too delicate, so this method is not recommended.

IV General discussion

Based on the results of the above three clustering methods, the Kohonen clustering is the least preferred. If the quality of the model clustering is used, the K-means algorithm aggregates and separates the contours when the number of clusters is 3, 4, and 5. The order of quality is 0.7,

0.68, and 0.67. The number of clusters obtained by the two-step clustering algorithm is 5, and the model's cohesion and separation contour is 0.7, which shows that the two types of algorithms perform similarly in terms of the model quality, and we cannot determine which one is better or worse. If we look at the number of classifications, the number of classifications obtained by the K-means algorithm is 3, 4, and 5, and the number of clusters obtained by the two-step clustering algorithm is grouped into five categories. In the Kohonen algorithm, the clusters are separated and the number of clusters is too scattered, which indicates that this method is not conducive to customer segmentation based on the clustering results. According to the classification effect chart, the Kohonen algorithm is reflected in the figure because of the large number of classifications. In this case, each data cluster is difficult to identify, and the clustering effect is not found satisfactory. When the number of K-means clusters is set to 3, we can draw the following conclusion based on the clustering effect diagram: the blue data cluster at the bottom right of the image exhibits a better classification effect. It is basically separated from other classes, and the red and green data clusters at the upper right of the image show very poor classification, thus indicating a serious data overlap phenomenon. When the K-means algorithm cluster number is specified as 4, the following conclusions can be drawn from the figure based on the clustering effect: the data cluster on the upper left of the image is basically separated, and the clustering effect is better, but the clustering effect on the lower right of the image is not good. The red and green data blocks overlap, and the orange data blocks are basically separated from the dataset, but the overall clustering effect is better than the last time. When the number of clusters in the K-means algorithm is specified as 5, the following conclusion can be drawn based on the clustering effect diagram: the red data block at the bottom left of the image is basically separated from the dataset, and the classification effect of the green and blue data blocks is also unsatisfactory. The first set of data on the upper right of the image is forcibly divided into two categories, some of which are too far-fetched. The following classification effect chart of the two-step clustering is followed: the data is divided into five categories in the clustering model, each category is marked with a different color, one can find that the classification effect is perfect, each class can be basically separated from the dataset without overlapping, and no overlapping with other datasets and the size of each type of the dataset is basically similar, which indicates a good classification effect from all aspects.

In order to more accurately compare the differences between the clustering models, we use some distance-based indicators for comparison and collect some information about the distance from the various clustering models to compare the advantages and disadvantages of these models. These evaluation criteria include the number of clusters, the average distance between clusters, average distance within clusters, sum of the squared distances within clusters, DUNN, CH, and other indicators.

Table 7. Summary of the clustering index.

	K-means (K = 3)	K-means (K = 4)	K-means (K = 5)	Two-step clustering	Kohonen
Number of clusters	3	4	5	5	9
Outliers	0	0	0	0	0
Average distance between classes	2.738	2.927	2.49	2.742	3.67
Average distance within class	1.64	1.527	1.579	1.376	2.73
Sum of squares within distance	21736.72	2237.49	1980	1756.82	2988.42
CH	849.23	1038.64	1498.71	1827.25	624.37

As can be seen from Table 7, the average distance between classes is the maximum for Kohonen, which is 3.67, and the minimum for the K-means (K = 5), which is 2.49. The average distance between the two-step clustering is in the middle, whereas the average distance within the class is the minimum for the two-step clustering. For the sum of squares, it can be seen that the two-step clustering value is the smallest, whereas the Kohonen value is the largest, which indicates that the clustering effect is not good. For the CH index, the situation is the same, the two-step clustering value is the largest, and the Kohonen value is the smallest, which indicates that the two-step clustering effect is better, but the Kohonen clustering effect is not good.

Explanation of CH indicators:

This index describes the tightness by calculating the value of the intra-class dispersion matrix and the separation degree by determining the value of the inter-class dispersion matrix. The CH index is expressed as follows:

$$CH(\mathrm{k}) = \frac{tr B(\mathrm{k}) / (\mathrm{k}-1)}{tr W(\mathrm{k}) / (\mathrm{n}-k)}$$

where n is the number of clusters, k is the number of samples in the current class, tr $B(k)$ is the trace of the inter-class dispersion matrix, and tr $W(k)$ is the trace of the intra-class dispersion matrix.

From the above formula, it can be observed that the larger the CH, the closer the classes to each other. Conversely, the more dispersed the classes and the larger the CH index, the better the clustering effect.

Based on the above discussion, we can draw the following conclusions:

①Based on the number of clusters, the two-step clustering clustered out five categories, each of, which is of moderate size and convenient for practical commercial promotion.

②From the perspective of the clustering quality, the two-step clustering is quite good in terms of various indicators, which is found to be consistent with a good clustering model.

③Based on the clustering effect diagram, the data clusters grouped by the two-step clustering are clearly defined, which are easy to identify, and the original dataset is well distinguished.

In summary, as compared to the other two methods, the two-step clustering method exhibits the best clustering effect, and the two-step clustering method has the following obvious advantages in practical applications:

①This Method can be applied to large-scale problems.

②This Method can be applied to classification problems.

③In this Method, there is no need to specify the number of clusters in advance, but it can also automatically realize the clustering function according to the characteristics of the dataset.

Considering the above advantages, the two-step clustering model demonstrates a better effect in solving the problem of customer segmentation in Jingdong Mall.

V Conclusion

By the establishment and analysis of the previous model, we can find that the two-step clustering method demonstrates the best clustering effect. Therefore, we select the two-step clustering method as the customer segmentation clustering model. Finally, customers are divided into five categories in this method. Their basic numerical characteristics are presented in Table 8.

Table 8. Model overview.

	Number of customers	Average of average consumption	Average of consumption	Average of the most recent time interval
First-class customers	19.70%	341.75	70.19	20
The second category of customers	20.10%	201.87	40.29	39.53
The third category of customers	20%	25.72	5.49	249.81
Fourth category of customers	19.80%	74.67	14.98	149.91
The fifth category of customers	20.30%	195.35	15.01	74.76

From Table 8, we can draw the following conclusions:

①The proportion of the five types of customers is relatively average. The minimum proportion belongs to the first category customers, which is 19.70%, the largest proportion belongs to the fifth category customers, which is 20.30%, and the proportions of the remaining categories are around 20%. The appearance of customer inclination helps companies to conduct precise marketing based on the classification results.

②The first types of customers are the most eye-catching among the five types of customers. It accounts for a relatively small proportion of the total number of customers, the highest average consumption fee, the most frequent consumption, and the shortest consumption time interval. Therefore, they are the most valuable to companies. Most of these customers also contribute to the enterprise. Based on the frequency of consumption and the final consumption time, these types of customers demonstrate a strong willingness to spend and are also the most loyal to the company. They should be considered as more valuable old users. Based on the average of the average consumption of these types of customers, the average value of consumption is also the largest, which indicates that they have a strong purchasing power, so they should be considered the most valuable users of the enterprise, and hence the enterprise should focus on them, provide special personnel to serve them, and always contact such customers at any time. In addition, the best resources of the enterprise should communicate with them, and their good relationship with the enterprise should be maintained in order to obtain more lasting and reliable benefits.

③Let's again discuss the characteristics of the third types of customers. It accounts for 20% of the total customers. The average of average consumption is 25.72 Yuan, which is the lowest among the five types of customers. The average number of consumption times is 5.49, and it is also among the five types of customers. The lowest and the average value of the most recent time interval is 249.81, which is the highest among the five types of customers and shows that they account for the lowest value of all customers, so the enterprises can appropriately reduce their investment in them and put more resources in other customers.

④Let's again discuss the characteristics of the fourth types of customers. The characteristics of these types of customers are also very obvious. The proportion of these customers is 19.80%, which is at the intermediate level. The average of average consumption is 74.67 Yuan, which is only higher than the average value of the third type of customers. It is 14.98 times, which is only slightly higher than the third type of customers, and the average value of the most recent time interval is 149.91 times, which is only slightly lower than the third type of customers. Overall, these types of customers are not valuable, but they are better than the third types of customers.

Based on the value of consumption, they consume less. The possible reason may be that they are not loyal enough to the company, and the product is not targeted to them. The company should resolve these issues and increase activities to attract such customers and increase their average consumption. On average, although the amount of each purchase is not large, it can also bring value to the enterprise. Therefore, the targeted products should be recommended to them, and the enterprise should try to retain such customers. Considering the most recent time interval, the situation is not optimistic. They have not visited the store to buy goods for more than four months, and there is a risk of customer churn. The enterprise should appropriately conduct telephonic interviews and door-to-door sales to attract such customers and prevent customer churn.

⑤The second type of customers accounted for 20.10% of the total customers, and all the characteristics are considered to be relatively good. There is great hope to develop them into high-quality customers. Based on the average consumption, they spend 201.87 Yuan on average. The overall customer is upstream, which indicates that these customers have a strong purchasing power and a strong willingness to buy. Therefore, enterprises should put more resources on such customers, issue coupons to them, and bundle promotions and other activities to improve their purchases. Based on the average purchases, these customers also show great value, but the purchases are not very high. The possible reason may be that they are not loyal to the company or there are more competitive companies. Therefore, enterprises should investigate the reasons and continue to introduce new products to attract such customers in order to increase their purchase frequency.

⑥The main characteristics of the fifth type of customers are the average of the average consumption and the average number of consumption times. Based on the numerical values, we can find that the average of average consumption is 195.35 Yuan, which is upstream among all customers. These types of customers show a strong purchasing power and a strong willingness to buy, but it is only 15.1 times as compared to the number of consumptions. This is not very good among all customers, which indicates that these types of customers are not very loyal to the enterprise, but they only occasionally buy. They rarely visit again to buy the previous product. In this situation, companies should do user service work such as timely follow-up on user usage, survey user satisfaction, provide users with value-added services, and appropriate after the user purchases the product. Additional products and other services increase user stickiness and loyalty.

Abbreviations

This article does not contain any abbreviations.

Availability of data and materials

The data that supports the findings of this study are available in the supplementary material of this article.

Competing interests

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work; there is no professional or other personal interest of any nature or kind in any product.

funding statement

This article is funded by the commercial project of Jingdong Century Trading Co., Ltd.

Authors' contributions

The article was completed by Yang Liu.

Acknowledgments

The authors would like to thank Jingdong Century Trading Co., Ltd., for assisting in project framing, data collection, and managerial support for this research project.

References

[1]HUNGS-Y,YENDC,WANGH-Y. Applying data mining micromanagement. Expert Systems with Applications,2006,31(3):515–524.

[2]TSUEN 一 HOHSU.An Applieation of Fuzzy Clustering in Group-Positioning Analysis [J].Proe.Natl.Sei,Coune.ROC(C),2000(10):157 一 167P.

[3]J, L, Castejón;, R, Gilar;, P, Miñano;, M, González;. Latent class cluster analysis in exploring different profiles of gifted and talented students [J]: Latent class cluster analysis in exploring different profiles of gifted and talented students, 2016, (10): 43–50.

[4] LAmir, Torghabeh;, Reza, Rezaee. Electrofacies in gas shale from well log data via cluster analysis: A case study of the Perth Basin, Western Australia [J]. Open Geosciences, 2014, (14): 22–26.

[5]Nadezda, Zenina;, Andrejs, Romanovs. Transport Simulation Model Calibration with Two-Step Cluster Analysis Procedure [J]. De Gruyter, 2006, (4): 46–48.

[6]Daniel, Macina;, Zofia, Piwowarska. SBA-15 loaded with iron by various methods as catalyst for DeNOx process [J]. Materials Research Bulletin, 2014, (45): 56–59.

[7]Mika, Koivisto;, Eveliina, Rientamo;. Unconscious vision spots the animal but not the dog: Masked priming of natural scenes [J]. Consciousness and Cognition, 2004, (56): 77–79.

[8]John, K, Mark;, Sophie, Dionne. Utility of standard pharmacopeial and nonpharmacopeial methods in distinguishing folded, unfolded, and process variant forms of interferon α –2[J]. Journal of Pharmaceutical Sciences, 2006, (47): 42–47.

[9] Zhou Qingsong. Application analysis of SAS-WebLogic data mining model [J]. Fujian Computer, 2015, (4): 26–28.

[10] Gao Zhihao, Jiang Jianzhong, Xia Lei. Real-time multimedia data mining model based on hierarchical vector distance [J]. Computer Engineering and Applications, 2007, (3): 14–17.

[11] Xia Weili, Wang Qingsong. Research on customer segmentation and retention strategy based on customer value [J]. Management Science, 2006, (4): 34–37.

[12] Ma Huimin, Yin Hanbin, Xiao Wei. Customer potential value prediction model and subdivision research [J]. Industrial Engineering and Management, 2003, (2): 14–17.

[13] Chen Qi. Application of data mining in the segmentation of telecommunication customers [J]. Science Technology and Engineering, 2009, (16): 37–40.

[14] Feng Dengguo, Zhang Min, Li Hao. Big Data Security and Privacy Protection [J]. Science Technology and Engineering, 2009, (16): 48–52.

[15]TSUEN-HOHSU.An Applieation of Fuzzy Clustering in Group-Positioning Analysis [J].Proe.Natl.Sci,Counc.ROC(C),2000(10):157–167P.

[16] M. Michalopoulos, GD Dounias, N. Thomaidis, G. Tselentis. Deeision Making Using Fuzzy C-means and Inductive Machine Learning for Managing Bank Branches Performance.www.erudit.de/erudit/events/esit99/12623_P.pdf .

[17] Steven Russell, Weldon Lordwick. Fuzzy Clustering in Data Mining For Telco Data base Marketing CamPaigns. New York: Proeeedings of NAFIPS99.1999(6):720–726P.

[18] Fuzzy Clustering for Customer Segmentation in the Finanical Services Industry.http://www.sigmaplus.fr/DataEngine/Tutorial/tableofeont.htm#l.

[19] SPSS technical white paper. SPSS company. www.spss.eom.en.23–25P.

[20]Soper, Suzanne, the evolution of segment ation methods in services: where next [J] .Journal of Financial Services Marketing, 2002, 8:68 –69.

[21]Lazer, William, Life style concept and marketing, toward scientific marketing [M] .St ephen Greyser, ed, Chi cago :American Marketing Assn., 1963:130.

[22]Wells, William, Tigert, Doug, Activities, interests, and opinion [J]. Journal of Advertising Research, August, 1971 ,11:27 –35.

[23] Plummer, Joseph T ., The concept and application of life style segmentation [J] .Journal of Marketing, Jan 1974, 38(1):34.

[24]Hughes,A ., Strategic database marketing: the masterplan for st arting and managing a profitable, customer-based marketing program [A]. Irwin Prof essional .1994.

[25]Marcus, C., A practical yet meaningful approach to customer segmentation [J] .The Journal of Consumer Marketing, 1998, 15(5): 494.

[26]Haley, Russell.I ., Benefit segmentation: a decision-oriented research tool [J] .Journal of Marketing ,1968, 32(7):30 –31.

[27] Vriens, Marco, Wedel, Michel, Wilms, Tom, metric conjoint segmentation method: a monte carlo comparison [J] .Journal of Marketing Research ,Chicago :Feb 1996 ,33(1):73–75.

[28] Kim, J., Mueller, C ., Factor analysis, sage publications [A] .Papers No.13 and 14.

[29] Chu Ruihong, Wang Hongjun, Yang Yan, Li Tianrui. Clustering Ensemble Based on Density Peaks [J]. Acta Automatica Sinica, 2004, (32): 47–49.

[30] Mao Dianhui. Improved Mapopy-based Canopy-Kmeans algorithm [J]. Computer Engineering and Applications, 2012, (48): 51–53.

[31] Lin Zhiyuan, Liu Gangdai, Guo Xian. Application of Kohonen neural network in radar multi-target sorting [J]. Journal of Air Force Engineering University (Natural Science Edition), 2003, (5): 46–50.

[32] Wang Fang, Zhu Han, Li Yunpeng, Liu Yufang. Temperature sensing research of dislocation fiber interference laser spectroscopy combined with BP neural network [J]. Spectroscopy and Spectral Analysis, 2007, (47): 26–27.

[33] Shang Yunlong, Zhang Qi, Cui Naxin, Zhang Chenghui. Research on the variable-order RC equivalent circuit model of lithium-ion battery based on AIC criterion [J]. Journal of Electrotechnics, 2015, (30): 27–29.

Figures

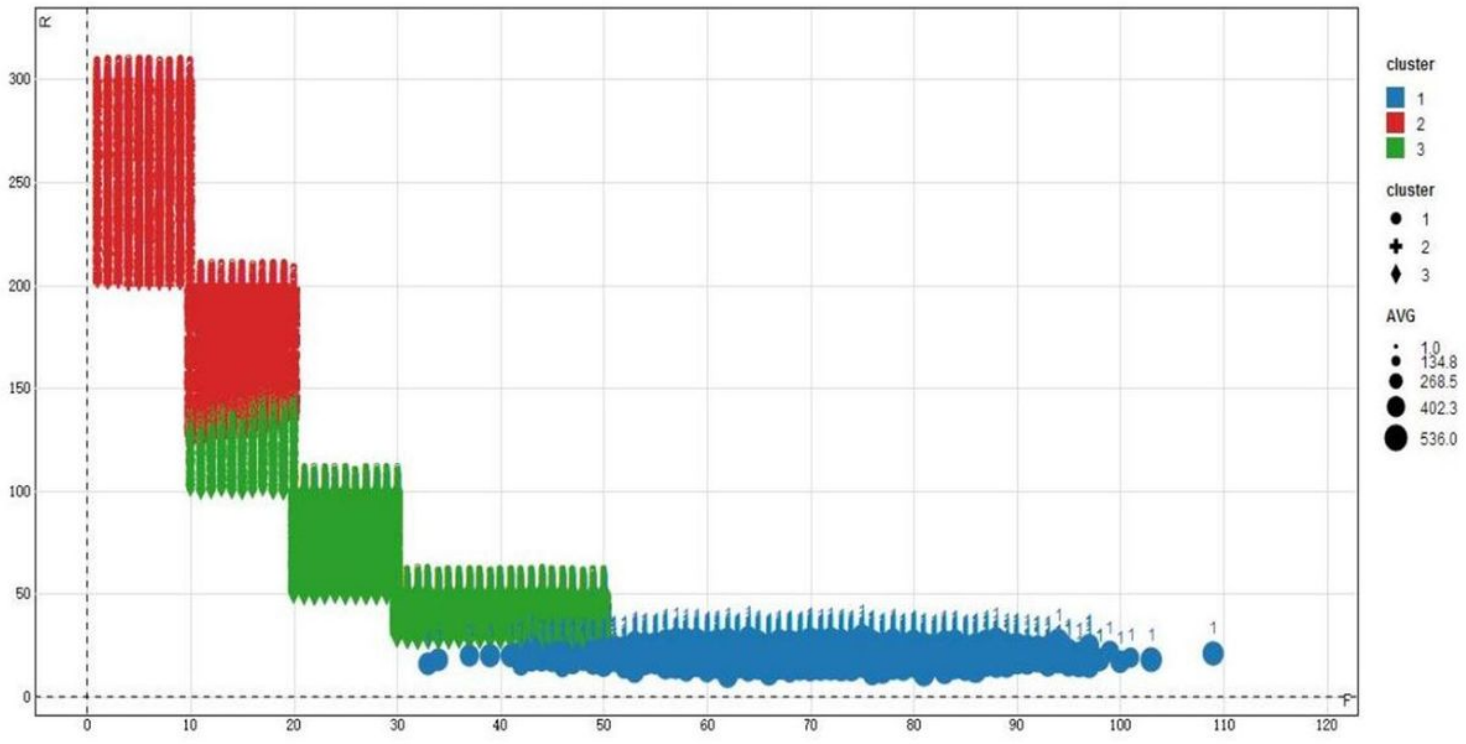


Figure 1

Clustering effect.

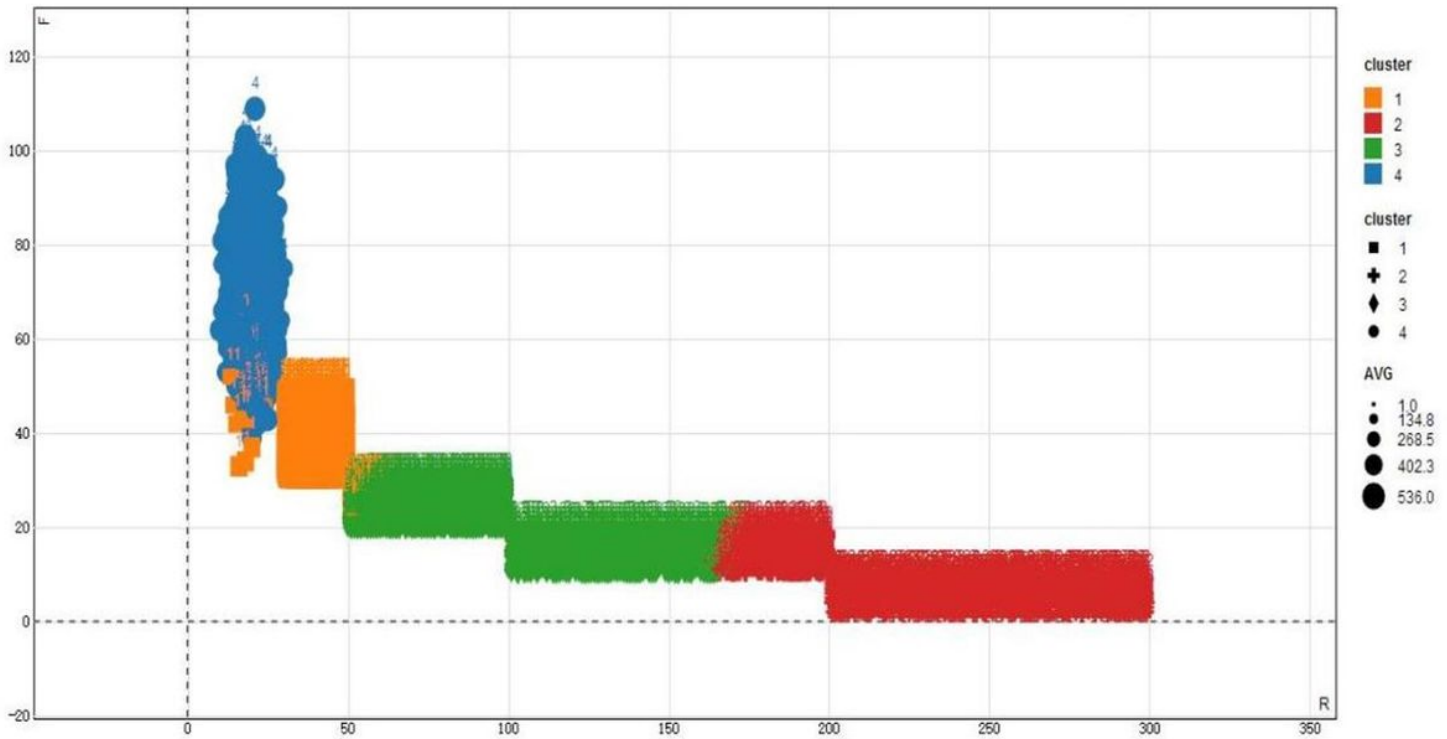


Figure 2

Clustering effect.

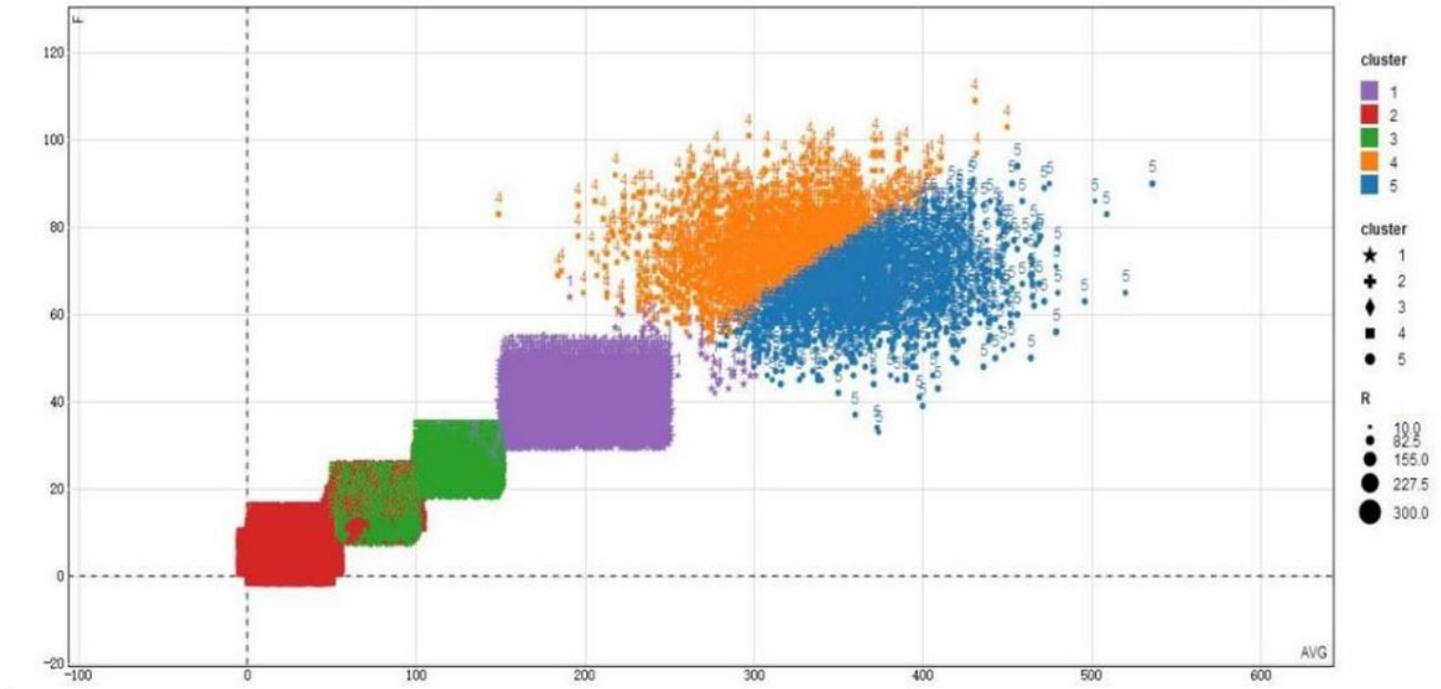


Figure 3

Clustering effect.

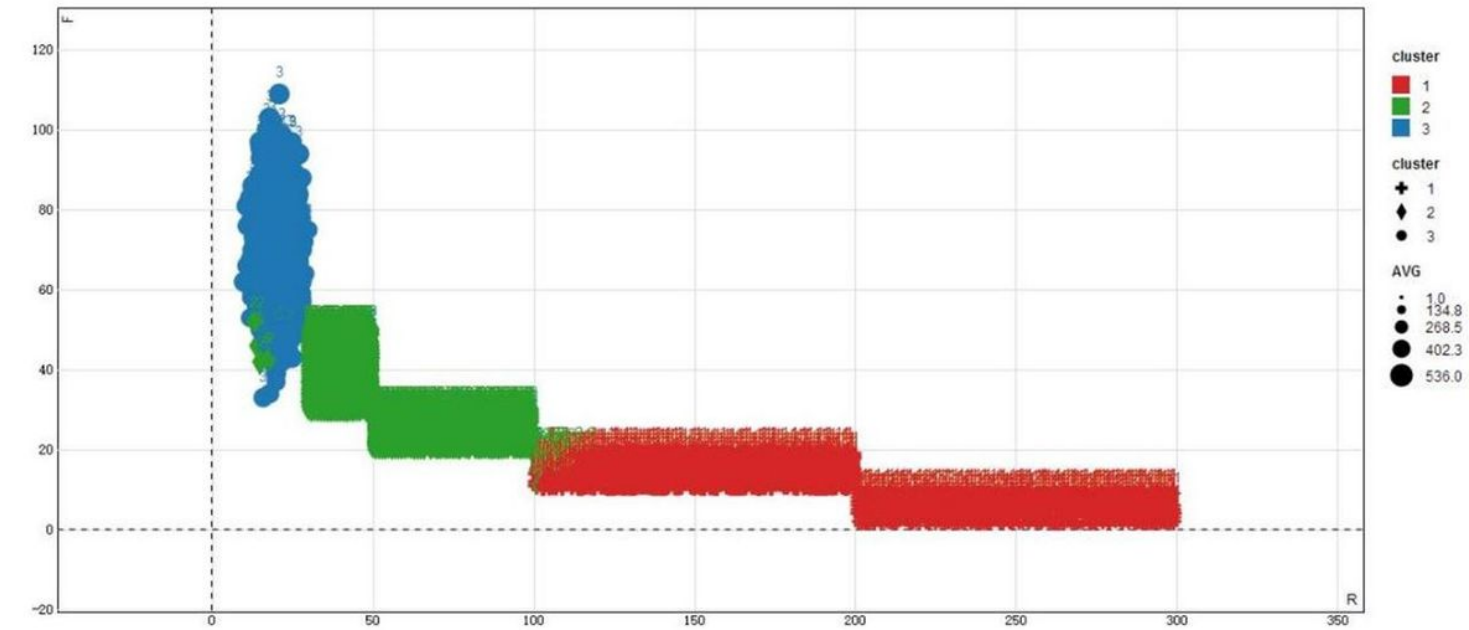


Figure 4

Clustering effect.

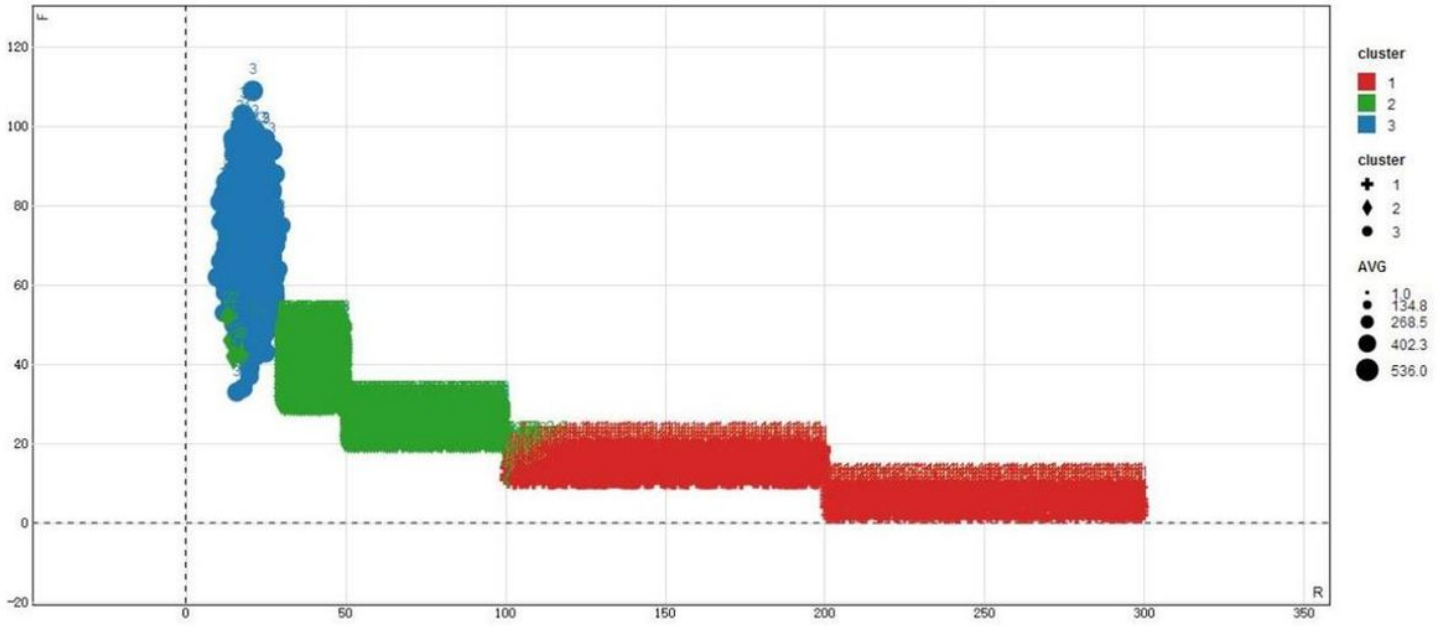


Figure 5

Clustering effect.

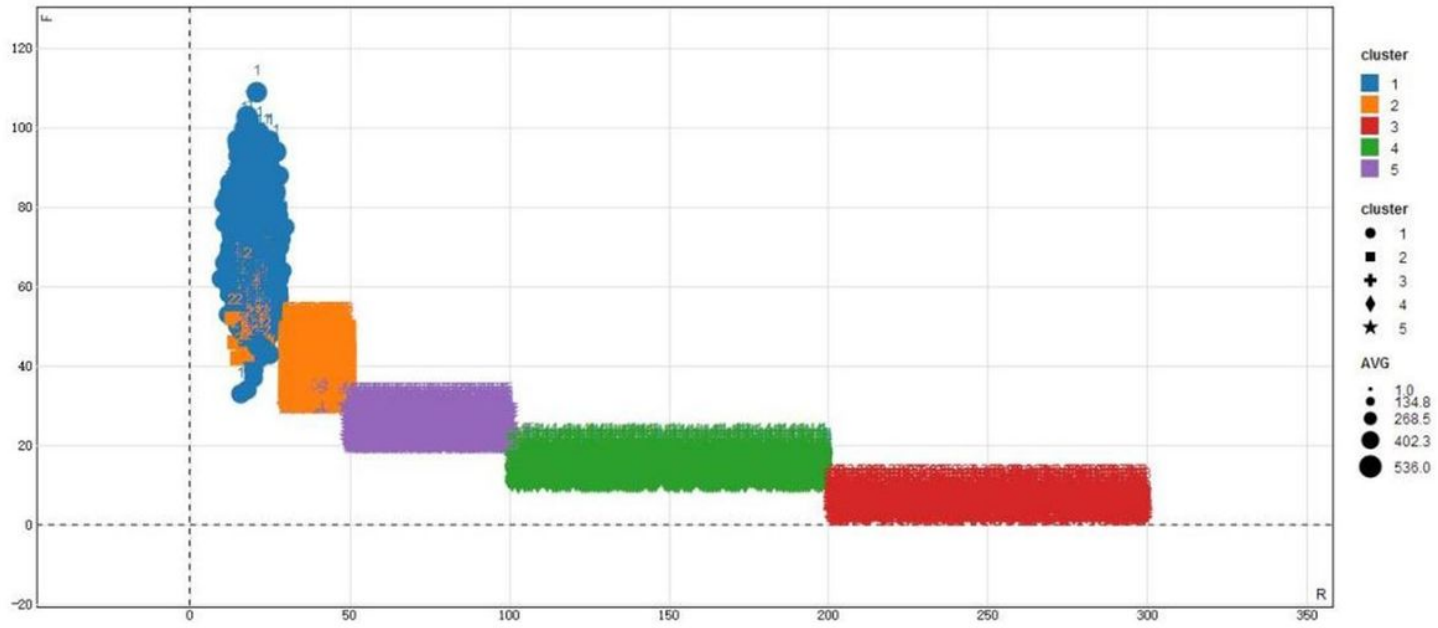


Figure 6

Clustering effect.

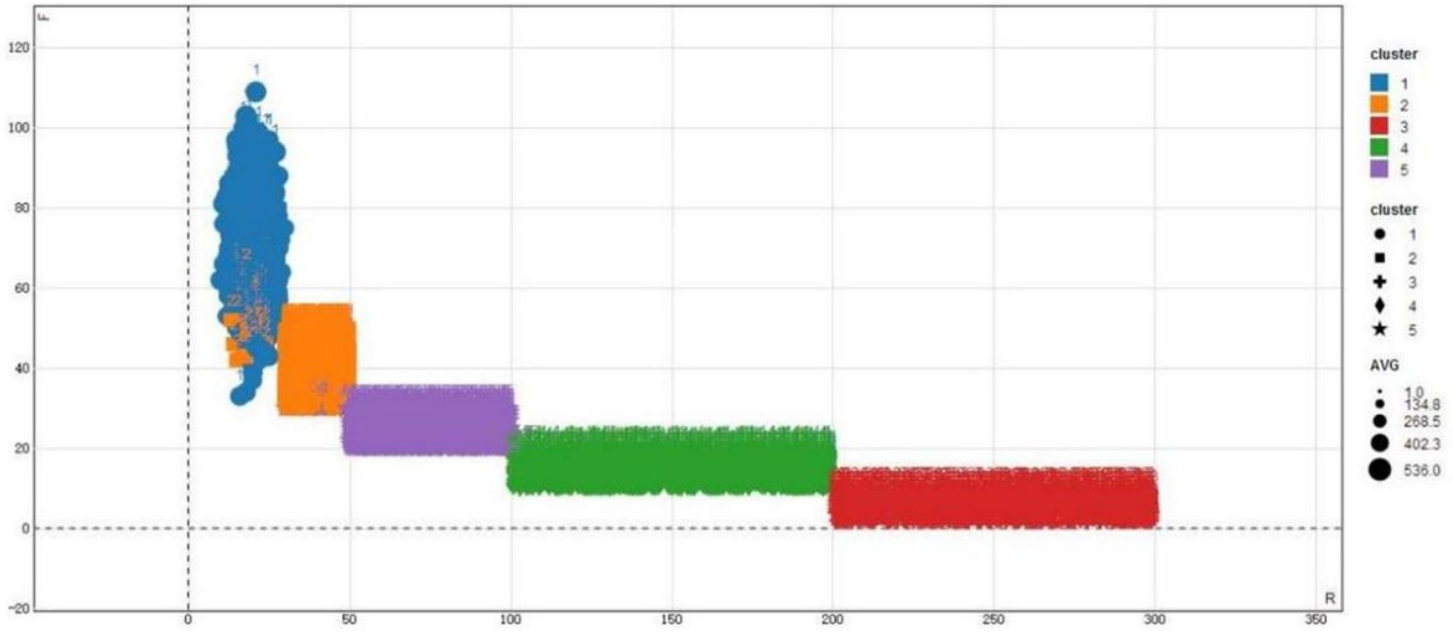


Figure 7

Clustering effect.