A Clustering Algorithm for Cross-border E-commerce Customer Segmentation

[1]SHULING YANG, [2]YAN HOU
[1]Foreign Language School of Jilin Normal University, Siping CHINA
[2]Management School of Jilin Normal University, Siping, CHINA

Abstract—With the deepening of reform and opening up, cross-border e-commerce has made great progress and plays a very important role in today's society. Cross-border e-commerce is not only a place for commodity trading, but also a key channel for information communication when commodities are traded. Clustering analysis is one of the common technologies in the field of data mining, and it has its unique advantages in the application of customer segmentation. Firstly, this paper improves the selection of the initial clustering center of K-means clustering algorithm. Aiming at the defects of the existing literature, such as long time-consuming algorithm and poor accuracy when calculating the corresponding sample points for multiple maximum density parameter values as the initial clustering center, an improved scheme based on quadratic density is proposed and applied to customer value segmentation. The research shows that the improved K-means clustering algorithm significantly improves the quality of clustering, thus improving the effectiveness and pertinence of cross-border e-commerce marketing activities.

Keywords: K-means clustering algorithm; Cross-border e-commerce; Customer segmentation

Received: May 29, 2021. Revised: March 19, 2022. Accepted: April 27, 2022. Published: May 20, 2022.

1. Introduction

Cross-border e-commerce refers specifically to cross-border electronic commerce platform enterprises, which includes both third-party cross-border e-commerce platforms and self-built cross-border e-commerce platforms. In the cross-border electronic commerce transaction, cross-border e-commerce is the network hub of transaction activities, which is not only the medium of commodity display and browsing, but also the place where commodities are traded, and plays the role of bridging the supply and consumption of commodities [1]. Cross-border e-commerce will be very cumbersome and complicated when trading commodities, which also leads to many different operation modes of cross-border e-commerce. How to operate cross-border e-commerce reasonably and effectively to better serve people in the future is a problem that the relevant departments of cross-border e-commerce need to focus on [2].

In recent years, data mining clustering algorithm has become the most important means to achieve the goal of customer segmentation. Among them, K-means algorithm is the most widely used one. Many scholars have improved it and applied it [3-4], or made in-depth performance comparison analysis between K-means algorithm and other clustering analysis methods [5]. Literature [6] puts forward a new method of initial clustering center, which can get a better initial clustering center without setting a threshold, but it needs to scan the data set to be measured several times and calculate the distance of the corresponding data, which leads to greater computational complexity than other algorithms. Literature [7] proposed an algorithm for optimizing the initial center of K-means algorithm. When calculating the density of objects, this algorithm adopted the density-sensitive similarity measure and generated the initial cluster center of samples. Literature [8] proposed a new effective clustering function. In order to reduce the influence of outliers on the clustering results of the algorithm, the weighted K-means method is used to improve the traditional algorithm and get the clustering center. Compared with the traditional K-means, its clustering results are more effective. Literature [9] proposed an improved K-means algorithm based on genetic algorithm, and adopted a customer behavior segmentation model based on k-means algorithm to segment customers. Literature [10] uses fuzzy C-means clustering algorithm as the method of customer clustering. It provides a q uantitative basis for the feature analysis of customer groups, and obtains satisfactory customer clustering results. On the basis of comprehensive analysis of grid clustering algorithm and K-means clustering algorithm, literature [11] proposed an algorithm based on minimum clustering unit. In order to optimize the improper selection of initial points in K-means clustering algorithm, the classification results are greatly affected.

As a key component of customer relationship management, customer segmentation has gradually become an important premise for enterprises to apply customer relationship management. Through customer segmentation, enterprises can not only better identify different needs of different customers for enterprises, but also provide different services to different customers, thus improving customer satisfaction and loyalty. You can also find potential valuable customers in the customer group and enhance the competitiveness of enterprises. The main work of this paper is to improve the K-means clustering algorithm and integrate a single clustering algorithm with the idea of ensemble learning, and then apply the clustering ensemble algorithm to customer segmentation of cross-border e-commerce.

2. General Method and Process of Customer Segmentation

Generally speaking, customer segmentation can be carried out according to the following three customer attributes [12]:

(1)External attribute

For example, the geographical distribution of customers, the products owned by customers, and the organizational ownership of customers-enterprise users, individual users, government users, etc. This kind of stratification is usually the simplest and most intuitive, but at the same time it is also a relatively extensive classification. We still don't know which customers contribute more to the enterprise and those customers contribute less to the enterprise at every customer level.

(2)Intrinsic attribute

Intrinsic attributes are attributes determined by the internal factors of customers, such as gender, age, beliefs, hobbies, income, family members, credit, personality and value orientation, etc.

(3)Characteristics of consumption behavior

According to the consumer behavior, we can master the real consumer habits and tendencies of customers, and usually get ideal results in practice. However, classification according to consumption behavior also has its limitations. It can only be applied to existing customers. For potential customers, because consumption behavior has not yet started, of course, classification is impossible.

Customer segmentation can generally be divided into five steps (as shown in Figure 1):

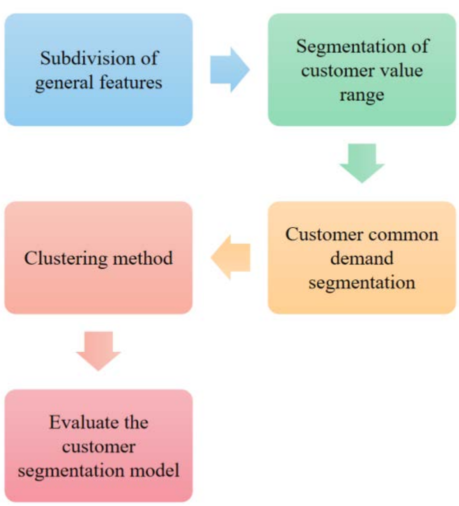

Figure 1 Customer segmentation steps

(1)Subdivision of general characteristics of customers

In order to classify customers according to these characteristics, the main factors that should be considered are: regional characteristics, such as urban or rural areas, urban scale and urban economic development level; Education background, such as age, sex, education level, nature of work unit, position or level; Psychological factors, such as personality characteristics, moral development level, etc.

(2)Customer value segmentation

Customers' contribution to enterprises is different according to their own consumption level. Therefore, after customers are subdivided according to their general characteristics, they should be divided into several grades according to their contribution to the enterprise, such as high-quality customers, potential customers, general customers, small customers and blacklist customers, etc.

(3)Customer common demand segmentation

On the basis of the first two steps of subdivision, select the high-quality customers and potential high-quality customers in the enterprise as the target. Analyze the demand characteristics of all kinds of customers, and formulate enterprise strategies under the guidance of customer demand, and finally provide personalized products and services for each customer group.

(4)Select a clustering method suitable for enterprise data characteristics

Clustering algorithm is an unsupervised learning algorithm. When using clustering technology to subdivide customers, we should choose the appropriate algorithm according to the needs of enterprises, the characteristics of customers and the collected data, so as to mine and discover the true distribution of data.

(5)Evaluate the customer segmentation model

The purpose of customer segmentation model is to divide customers into different clusters according to their various characteristics. According to the needs of enterprises, customers in the same cluster should have similar contribution and consumption tendency, while customers in different clusters should try to be different in these aspects. These characteristics can be measured according to the mean and variance of customer attributes.

3. Research Method

3.1 Clustering Analysis Algorithm

Clustering is a main technology in data mining. The process of grouping a set of objects into multiple classes composed of similar objects is called clustering [13-14]. After grouping, the objects in the same class are similar, but the objects in different classes are different. At the same time, cluster analysis is often used as the first step of data mining, which preprocesses the data, and then uses other algorithms to further analyze the obtained classes. Clustering algorithm can be divided into partition method, hierarchical method, density-based method, grid-based method and model-based method.

K-means algorithm is one of the clustering algorithms based on partition. It uses an iterative climbing method to discover clusters and cluster centers from unlabeled data sets. Its purpose is to divide N samples into M clusters, so that the sum of squares of errors between the data samples in each cluster and the mean value of this cluster is the smallest.

The work flow of K-means clustering algorithm is: firstly, randomly select k samples as initial clustering centers, then calculate the distance between each sample and the clustering center, classify the samples into the class where the nearest clustering center is located, and recalculate the adjusted clustering center of the new class, repeating this process until the clustering centers of the adjacent two times have not changed, at this time, the adjustment of samples ends and the algorithm has converged.

Let $p \in C_i$, C_i represent a cluster, p represents the sample point in C_i, c_i represents the center point (mean value) of C_i, and the difference between p and c_i can be measured by $d(p,c_i)$, where $d(p,c_i)$ represents the Euclidean distance between two points p,c_i.

The quality of C_i can be measured by the objective function, which represents the sum of error squares of all other sample points in the center point c_i and C_i. Generally, there are the following functions:

$$E(I)=\sum_{i=1}^{k}\sum_{p\in c_i} d(p,c_i)^2 \tag{1}$$

In which $E(I)$ represents the sum of error squares of all sample objects in the data set, p represents the sample point, and c_i represents the center point of cluster C_i.

Execution flow of K-means algorithm:

Algorithm: K-means clustering algorithm

Input: data set D, size n, number of generated clusters k.

Output: divided k clusters.

Steps:

(1)Given a data set D with a sample point capacity of n; Given the size of k, randomly select k sample points as $c_j, j=1,2,\cdots,k$ in the initial cluster.

(2)Cluster the data set D for the first time, calculate the distance $d(x_i,c_j), i=1,2,\cdots,n, j=1,2,\cdots,k$ from other sample points that are not the initial center to the center points of each cluster, find out the minimum distance, divide the sample points into the corresponding clusters, and find out the designated k clusters.

(3)According to the obtained clusters, recalculate the mean value of data samples in each cluster, and take it as the new cluster center $c_j, j=1,2,\cdots,k$.

(4)Repeat steps (2) and (3) until the clustering center points of the previous two times do not change or the objective function converges.

(5)Output the obtained k clusters.

Before K-means clustering, not only the value of cluster number k should be given in advance, but also the initial cluster center should be given. Therefore, the initial value has a great influence on the clustering results. If the initial value is not selected properly, the clustering results will have great errors. Therefore, the K-means algorithm has a strong dependence on the initial clustering center. If the selected initial value is outlier data, the clustering criterion function will be difficult to converge quickly and the clustering result will be unstable, which is also a difficult problem in the K-means algorithm.

3.2 Improvement of K-means Algorithm

In order to achieve better clustering results, we have made corresponding algorithm improvements in view of the shortcomings of k-means algorithm [15-16].

First of all, for the determination of the optimal k value, we use the method of permutation and combination to exhaust all cases, and the specific process is as follows:

Executing k-means algorithm once for all values $(1 \le k \le n)$ of k to obtain each corresponding clustering result data;

Calculating the S value of the distance function after clustering with different k values; Choose the one with the smallest S value from all distance functions, and its corresponding k value is the optimal k value we are looking for.

Aiming at the defect that the initial center point of k-means algorithm is blindly selected, we use a new distance function to select the initial clustering center point according to the distribution characteristics of sample data, so as to improve the clustering speed and effect [17].

Set a data set Q, in which there are n data samples, the final number of clusters is k, and the data variable is p. The distance function between any two data points is defined as:

$$d(x,y)=\sqrt{(x_1,y_2)^2+(x_2,y_2)^2+\cdots+(x_p,y_p)^2} \tag{2}$$

In which: the minimum distance between data variable x and data set Q is defined as:

$$d_{\min}(x,Q)=\min(d(x,y)), y\in Q \tag{3}$$

The maximum distance between data x and data set Q is defined as:

$$d_{\max}(x,Q)=\max(d(x,y)), y\in Q \tag{4}$$

The restriction conditions for data y to be included in data set Q are defined as follows:

$$l=\frac{\max\limits_{1\le x,y\le n}(d(x,y))-\min\limits_{1\le x,y\le n}(d(i,j))}{k} \tag{5}$$

The specific steps of improving the k-means algorithm are as follows:

(1)Enter the data set and initialize the parameters.

(2)Run the iterative process.

1)Run the iterative process and get k clustering results. If $HS=1$ goes to 3), otherwise proceed to the next step.

2)Check whether the clustering result converges, if so, calculate $Sil(k)$ and mark $HS=1$, and go to 4); Otherwise turn back to the previous step.

3)Check whether the cluster center meets the convergence condition, if it converges, get k clusters and calculate $Sil_{\max}$, if it is $Sil(k) < Sil(k-1)$, then $H=H+1$; $H=0$ when $Sil(k) > Sil_{\max}$.

4)Check whether it is $H > k_1/2$; Whether k is 2; And check whether the number of cycles meets the termination condition, if so, go to 5), otherwise, go back to 1).

5)Test the number of optimal clusters corresponding to $Sil_{\max}$. If it is 2, calculate the $Hartigan$ index and compare it.

6)Output cluster number and cluster center.

(3)The output cluster number k and cluster center initialize k-means algorithm.

(4)Run k-means algorithm to get the final clustering result.

According to the above steps, draw the flow chart of the improved k-means algorithm, as shown in Figure 2 below:

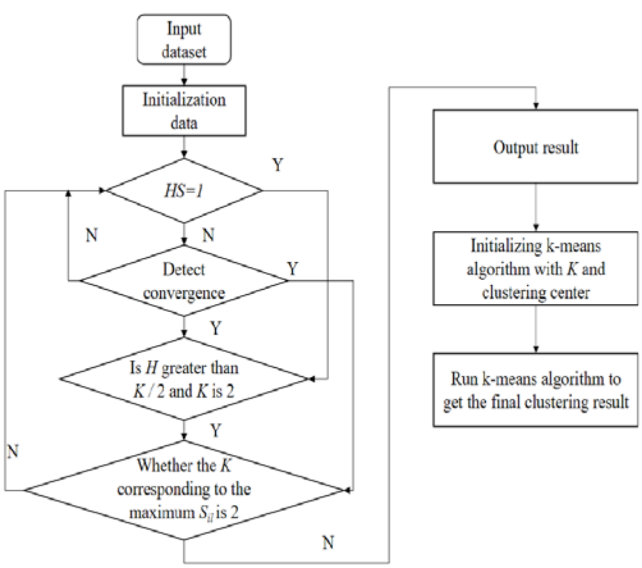

Figure 2 Flow chart of improved k-means algorithm

3.3 Application of Improved k-means Algorithm in Cross-border E-commerce Customer Segmentation

3.3.1Data acquisition

Today, with the vigorous development of e-commerce, various e-commerce websites have accumulated a large amount of data information, such as customer information, sales information, commodity information, etc. These data play a vital role in the formulation of sales strategies or promotion strategies of enterprises [18-19]. How to obtain effective data of customer segmentation is also the key to our success.

The data used in this paper comes from a co smetics e-commerce website. Because there are a lot of complex data in the data source, we need to extract the used data from the database, including customer information table, commodity information table, customer order table, etc.

3.3.2Data preprocessing

The original data usually has defects such as nonstandard, repetitive and incomplete, etc. The original data can be repaired by data cleaning, such as missing value processing, exclusion of isolated points, deletion of noise data, etc., so as to make the data consistent as much as possible.

In the enterprise database, when customers register their accounts, some options are optional, and some options may be sensitive information, which will cause customers to be reluctant to fill in information, leaving a lot of missing values in the database. Therefore, before data analysis, we must first deal with the missing values. Commonly used methods to deal with missing values include: manual processing, estimated filling and so on.

Noise refers to repeated, wrong and incomplete data in data. Error data can be detected by statistical principles [20]. Basically, the two positive and negative standard deviations of data exceeding the mean value can be regarded as noise data. Incomplete data is data with incomplete information. For example, the commonly used language information left by some customers is not complete enough, which will also become potential noise data. As for the duplicate data, simply put, it i s the data of repeated information. A customer's consumption behavior is recorded as the same twice, which will definitely have a wrong influence on the related analysis.

For each customer record with many dimensions, its attribute types are not consistent, some are numeric, some are Boolean, and some are text [21]. In order to facilitate the later correct calculation, it is necessary to transform the source records to meet the data mining standards. Here, the attribute values are normalized and mapped to the [0,1] interval, and the calculation formula is as follows:

$$A_i^* = \frac{A_i - \min(A_i)}{\max(A_i) - \min(A_i)} \tag{6}$$

In the formula, A_i represents the i th attribute value of data A, $\min(A_i), \max(A_i)$ represents the minimum and maximum values of the i th attribute in all elements respectively, and A_i^* is the normalized value.

Because of the need to count the final subdivision results, the REFERENCE attribute of customers is replaced by ordinal numbers and recorded from 1, and each number represents a customer.

3.3.3Establishment of customer behavior segmentation model

Another important step before applying subdivision algorithm is to establish subdivision model, that is, which subdivision method can get better subdivision results. After determining the attribute index of subdivision, divide customer groups by subdivision method, and finally extract the group characteristics of each customer group.

In this application, after pre-processing the customer data, a segmentation model is established according to the relevant attributes of the customers, and then the data is integrated and subdivided by using the selective clustering integration algorithm, and different customer groups are divided according to the segmentation results, and then corresponding marketing suggestions are provided by extracting the group characteristics. The specific process of the whole application case is shown in Figure 3:

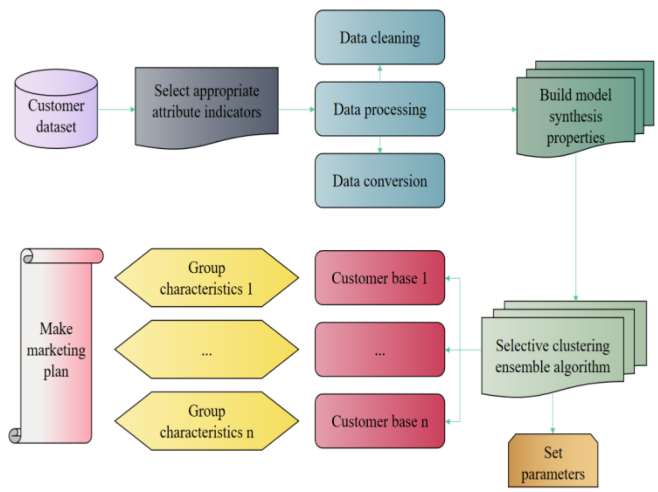

Figure 3 Overall flow chart of subdivision scheme

According to the characteristics of numerous customer data attributes and different data types, customers are divided into six groups according to the customer's Referranc R attribute. At this time, each group only contains the customer ID.

4. Analysis and Discussion

4.1 Performance Analysis of Improved k-means Algorithm

In order to test the feasibility of the optimization algorithm proposed in this paper, this section uses four different data sets on UCI data set for testing. They are customers dataset, Iris dataset, seeds dataset and Wine dataset. These data sets have been specifically classified in UCI data set, so we can accurately and intuitively calculate the accuracy and experimental effect of each clustering algorithm on these data sets.

Compare the initial clustering center selection time, clustering accuracy, overall running time and other aspects, and comprehensively compare the advantages and disadvantages of each algorithm. The experimental environment is Windows 7, 64-bit operating system and eclipse integrated environment.

Because the traditional K-means clustering algorithm selects the initial center point in a random way, there is no comparison here, only the time spent by the OLD-means algorithm and the text improved k-means algorithm in selecting the initial center point.

The results obtained by these two algorithms are all fixed when obtaining the initial center point, but according to the operating conditions of the machine and platform, each algorithm will have slight differences in time every time it runs, so it runs on the experimental platform for 10 times to obtain the average calculation time. As shown in Figure 5:

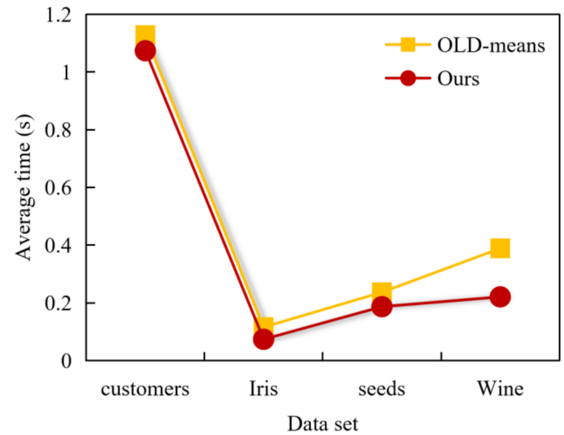

Figure 5 Time comparison of two algorithms in selecting initial center point

As can be seen from Figure 5, for the given four kinds of data sets, the time consumed by the improved k-means algorithm in selecting the initial center point is reduced to varying degrees compared with the OLD-means algorithm. Except for the customers data set, the reduction rate of the other three data sets is obvious.

However, Iris data set and Wine data set have more equal density parameter values in the local range, and their value jumps greatly in the overall range. Therefore, the improvement measures in this paper are fully utilized in clustering, which effectively shortens the time for finding the initial center point.

In each dataset, the traditional K-means clustering algorithm is run 10 times, and the highest accuracy, the lowest accuracy and the average accuracy are obtained respectively. However, the results obtained by the OLD-means algorithm and the improved k-means algorithm in this paper are fixed when selecting the initial clustering center, so we only need to run once to get better clustering accuracy. Fig. 6 shows the accuracy of each algorithm on different data sets:

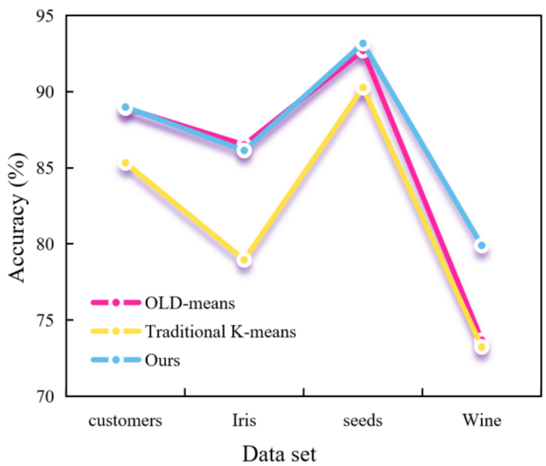

Figure 6 Compare the accuracy of three algorithms

From Figure 6, we can see that the traditional K-means algorithm has a great difference between the highest accuracy and the lowest accuracy, so it has great instability. From Iris, seeds and Wine data sets, we can see that the clustering accuracy of the improved k-means algorithm has a cer tain improvement compared with the OLD-means algorithm, and the improved K-means data set has a lower improvement.

Below, from the perspective of the overall operation of clustering algorithms, we compare the time consumed by the three algorithms on different data sets, run them on the experimental platform for 10 times, and take their respective average values. Fig. 7 shows the overall time consumption of the three algorithms.

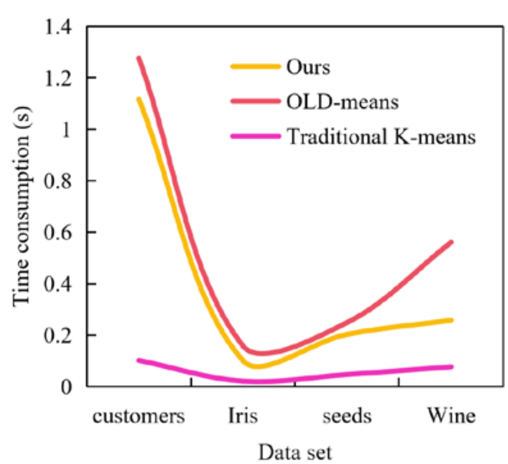

Figure 7 Overall time-consuming of the three algorithms

It can be seen from Figure 7 that the traditional K-means algorithm generates the initial cluster center randomly, so it consumes less time than the other two algorithms, but it brings instability and low accuracy. The latter two algorithms are stable and have good clustering results, although it takes a lot of time to calculate the density parameter values of sample points. At the same time, it can be seen that the overall running time of the improved k-means algorithm is shorter than that of the OLD-means algorithm.

To sum up, by testing the time consumption of initial center point selection, the accuracy of the algorithm and the overall running time consumption of the algorithm, it is shown that the improved k-means algorithm proposed in this paper has better optimization effect.

4.2 Cluster analysis of cross-border e-commerce customer segmentation

According to the customer segmentation results obtained by the improved k-means algorithm and the original data, we can get the consumption characteristics of customers in the following customer categories, as shown in Table 1.

Table 1 Customer consumption characteristics of different customer categories

Customer category	Customer consumption characteristics
Category 1	The average consumption times were 7.24 and the average consumption amount was 982.37
	The average consumption times are 7.631 and the average consumption amount is 956.01
Category 3	The average consumption times are 7.71 and the average consumption amount is 979.25
Category 4	The average consumption times were 7.82 and the average consumption amount was 977.52

Combined with superscript 1, we can get the following conclusions:

The number of customers in category 4 is the largest, which is characterized by less consumption times and less average consumption amount. Combined with relevant information of customers, it can be seen that most customers in this category have low educational background, low income and uneven distribution of age and location.

Category 3 customers have the least number, and the average consumption times of these customers are the least, but the average consumption amount is very high. It is observed that these customers have high academic qualifications and high income.

Compared with Category 4, Category 2 customers have fewer customers, but more customers than other types. This kind of customers have more average consumption times and average consumption amount. Most of them are around 30 years old, with average education and average income, and most of them come from second-and third-tier cities.

The average consumption times of Category 1 customers are a little less than that of Category 2 customers, but the consumption amount is very high. Most of them have the characteristics of high education, high income, first-tier cities, etc., and their age is generally 35 to 45 years old. They often come to this website for consumption, and the consumption amount is also very high.

Customer category information, customer quantity ratio and value are shown in Table 2 and Figure 8.

Table 2 Customer category information table

Customer category	Location	Academic degree	Age	Income
Category 1	First-tier city	Well-educated	35-45	High income
	Second and third tier cities	common	30	Common
Category 3	first-tier city	Well-educated	25-35	High income
Category 4	Unevenly distributed	Low education	Unevenly distributed	Low income

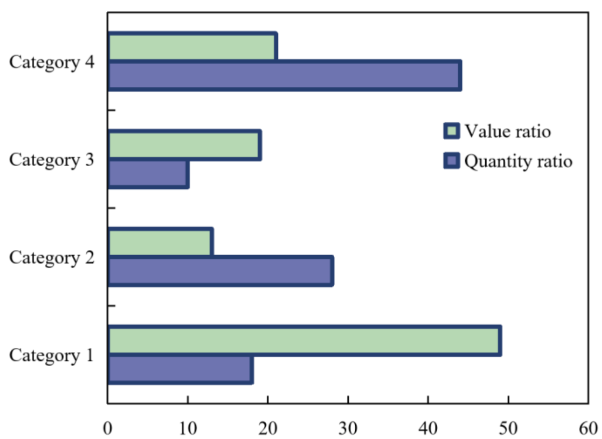

Figure 8 Customer ratio and value ratio

According to the comparison of customer quantity ratio and value ratio in the above table, it can be concluded that:

Category 4 has the largest number of customers, but the created value is not high, so we call it lead customers, and we don't need to invest too much resources in these customers;

Category 3 customers, although the number is the smallest, have created higher value, and they belong to potential customers. Enterprises should try their best to retain and develop these customers closer to L customers;

Category 2 customers account for an average number and create less value. Enterprises should use part of resources to keep such customers close to Category 3 customers.

Category 1 customers are not the largest in number, but they have created nearly 50% of profits. Therefore, this class of customers is called Platinum customers. They have created huge profits for enterprises, and enterprises should focus on maintaining these customers with limited resources.

5. Conclusion

In recent years, with the rapid development of information technology and cross-border electronic commerce, a h uge amount of commercial information has been accumulated in the database of cross-border e-commerce enterprises. If we can extract useful information from the massive and complicated information, make use of it and implement high-precision marketing schemes, we can help enterprises maintain and develop their own resource advantages in the fierce global competition. In this paper, the advantages and disadvantages of clustering analysis algorithm are analyzed, and a cl ustering algorithm based on quadratic density optimization is proposed on the premise of deep understanding of clustering analysis. Based on the improved K-means clustering algorithm, the base clustering algorithm is selected by extracting the sample subset, calculating the standard mutual information value, and the clustering integration algorithm is implemented by using voting rules. This model can obtain more robust and accurate clustering results by further combining multiple clustering fusion results. In some test data sets and the actual project implementation process, the model has achieved satisfactory clustering results.

In this paper, the customer segmentation results are divided into customer segmentation based on customer value and customer segmentation based on customer behavior, and the corresponding marketing strategies are given. However, there is no proposal to combine the two aspects to analyze the evaluation results. Therefore, in the next step, we can study the two aspects together to see if it can provide some help to the enterprise marketing strategy.

Acknowledgments

This work is supported by 2019 Doctoral Research Startup Project of Jilin Normal University(Project No. 2019037).

References

[1] WU Qiping, WU Chengmao. A fast and robust clustering segmentation algorithm for kernel space graphics[J].caai transactions on intelligent systems, 2019, 014(004):804-811.

[2] Liu Z, Xiang B, Song Y, et al. An Improved Unsupervised Image Segmentation Method Based on Multi-Objective Particle Swarm Optimization Clustering Algorithm[J]. Computer, Materials and Continuum, 2019(2):11.

[3] Li M. Study on the Grouping of Patients with Chronic Infectious Diseases Based on Data Mining[J]. Journal of Biosciences and Medicines, 2019, 07(11):119-135.

[4] Hartini S, Gata W, Kurniawan S, et al. Cosmetics Customer Segmentation and Profile in Indonesia Using Clustering and Classification Algorithm[J]. Journal of Physics Conference Series, 2020, 1641:012001.

[5] Chinta S S. Kernelised Rough Sets Based Clustering Algorithms Fused With Firefly Algorithm for Image Segmentation[J]. International Journal of Fuzzy System Applications, 2019, 8(4):25-38.

[6] Liu F. 3D Block Matching Algorithm in Concealed Image Recognition and E-Commerce Customer Segmentation[J]. IEEE Sensors Journal, 2019, PP(99):1-1.

[7] Nurmalasari, Mukhayaroh A, Marlina S, et al. Implementation of Clustering Algorithm Method for Customer Segmentation[J]. Journal of Computational and Theoretical Nanoscience, 2020, 17(2):1388-1395.

[8] Audrey, Guo. Four Keys of Cross-border E-commerce After Orders[J]. 2021(2015-2):12-13.

[9] Long V T. Research on the Influence of Transportation Services Quality on Purchasing Intention of Customer in E-Commerce - Evidence from Purchasing Intention of Vietnamese Consumer in Cosmetic Industry[J]. International Journal of Social Science and Education Research, 2020, 3(5):45-53.

[10] Zhang Z, Y Wang, Yang J. Deep quantised portrait matting[J]. IET Computer Vision, 2020, 14(6):339-349.

[11] Zhang Shi. Application of SOM-K-Means Cluster Algorithm in Customer Segmentation of Retail Bank SOM-K-Means[J]. Journal of panzhihua university, 2019, 036(005):66-70.

[12] Hu F, Chen H, Wang X. An Intuitionistic Kernel-Based Fuzzy C-Means Clustering Algorithm With Local Information for Power Equipment Image Segmentation[J]. IEEE Access, 2020, 8:4500-4514.

[13] Brust A F, Payton E J, Hobbs T J, et al. Application of the Maximum Flow–Minimum Cut Algorithm to Segmentation and Clustering of Materials Datasets[J]. Microscopy and Microanalysis, 2019, 25(4):924-941.

[14] Singh H. Improving Customer Segmentation in E-Commerce using Predictive Neural Network[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(2):2326-2331.

[15] Lei X, Ouyang H, Xu L. Kernel-Distance-Based Intuitionistic Fuzzy c-Means Clustering Algorithm and Its Application[J]. Pattern Recognition and Image Analysis, 2019, 29(4):592-597.

[16] Hasan F J, Mia M J. Association Rules and Clustering on Sparse Data of a Leading Online Retailer[J]. International Journal of Computer Science and Information Security, 2019, 17(4):112-116.

[17] Zhao H H, Luo X C, Ma R, et al. An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables[J]. IEEE Access, 2021, PP(99):1-1.

[18] Xie Zhongyang. Application and exploration of Python-based clustering method in e-commerce customer segmentation[J]. Digital Technology and Application, 2019, 37(03):230-231.

[19] Deng Xiaoyi, Jin Chun, Higuchi Ryoyuki, etc. KSP hybrid clustering algorithm for customer segmentation in mobile commerce[J]. 2021(2011-4):54-61.

[20] Wang Jian, Yin Lifei, Xu Jiadong. Research and application of LNG customer segmentation based on cluster analysis[J]. Chemical Management, 2019(6): 2.

[21] Huang Xuejin, Xie Chunxia. Research on the STP strategy of Chongqing's cross-border e-commerce retail export[J]. Jiangsu Business Forum, 2020(6): 4.