ICEB 2021 Proceedings (Nanjing, China)

International Conference on Electronic Business (ICEB)

Winter 12-3-2021

Customer Segmentation Using Real Transactional Data in E-Commerce Platform: A Case of Online Fashion Bags Shop

Zhouzhou Yan

Yang Zhao

Yan, Z.Z. & Zhao, Y. (2021). Customer segmentation using real transactional data in e-commerce platform: A case of online fashion bags shop. *In Proceedings of The International Conference on Electronic Business, Volume 21* (pp. 90-99). ICEB'21, Nanjing, China, December 3-7, 2021.

Customer Segmentation Using Real Transactional Data in E-Commerce Platform: A Case of Online Fashion Bags Shop

Zhouzhou Yan [1]
Yang Zhao [2,*]

*Corresponding author
[1] Master Student, Wuhan University, Wuhan, China, Yanzhouzhou@whu.edu.cn
[2] Professor, Wuhan University, Wuhan, China, yangzhao_0813@whu.edu.cn

ABSTRACT

Customer segmentation has been widely used in different businesses and plays important rules in customer service. How to get a suitable segmentation based on the real transactional data to fully mining the hidden customer information in the massive data is still a challenge in current e-commerce platforms. This paper develops a customer segmentation model for online shops and uses the real data from a fashion bag store as a case. This paper firstly conducts a data preprocessing to select the main customer features, then it constructs a segmentation model based on the Fuzzy C-Means algorithm, and finally accomplishes a customer prediction model using a probabilistic neural network to estimate new customer's customer type. The results show that the customer samples are classified into three types, and the prediction accuracy is more than 90%. After that, this paper demonstrates the typical features of each type of customer and compares the new group features with the prior VIP groups. The ANOVA analysis test results show that the new groups have more significant differences than prior VIP groups, which means more effective segmentation results.

Keywords: customer segmentation, Fuzzy C-Means, e-commerce platform, neural network

INTRODUCTION

Facing more complexity and competition in today's fast-moving e-commence, retailers need to capture customer behavior and improve customer satisfaction. Customer segmentation plays an important rule since consumers' buying behavior is not homogenous. Customer segmentation is dividing all customers into distinct groups that share similar characteristics, such as demographics, interests, patterns, or locations (Zhou, Wei & Xu, 2021). Customer segmentation has been widely used in different businesses such as telecom companies (Alkhayrat, Aljnidi & Aljoumaa, 2020) and the bank industry (Jouzdani et al., 2020). It has a positive effect on e-commerce (Li, Deolalikar & Pradhan, 2015). For sellers in e-commerce platforms, customer segmentation helps them to analyze the change of customer behavior of buying and returning, improve the choice of supplied products and personalize after-sales service; For customers, customer segmentation fulfills customers' needs to be more accurately positioned and improves customers' satisfaction with providing better shopping experience; For e-commerce platform managers, customer segmentation system can improve the organization of customer information in the platform and implement the accurate of information services.

Customer segmentation results depend on the data source for feature selection and clustering method (Song et al., 2021). The most commonly used procedure to obtain the customer data consists of filling in questionnaires (Ortigosa, Carro & Quiroga, 2014) and asking the consumers (Rohm & Swaminathan, 2004). Traditional customer segmentation often identifies customers as different level VIPs (very important person). Lee and Lee (2020) tested the difference between regular customers and VIP customers of Internet Banking, and the results indicate that regular customers and VIP customers differently perceive service quality that is relevant to customer satisfaction. Magatef and Tomalieh (2015) pointed out after charging an upfront Fee for VIPs. The VIPs are relieved of inconveniences that could impede future purchases. However, it is not the most effective customer loyalty groups compare with other groups such as tier groups where customers extract long-time value. Both the questionnaire data and VIP segmentation approach often fail due to rapidly changing e-markets.

In e-commerce, fortunately, there are large real-time transaction databases containing customer data about demographics and buying or return information. On the premise of protecting customers' privacy, these databases can be used to segment the customer and benefit customer relationship management. As an example, retailers can find those people who like to return goods and take actions to improve their satisfaction and reduce customer returns (Cullinane & Cullinane, 2021), which has brought great economic losses to e-commerce platforms, commodity suppliers, logistics carriers, and consumers (Lysenko-Ryba & Zimon, 2021). For these massive datasets, machine learning algorithms, including the clustering method (Sivaguru & Punniyamoorthy, 2020) and neural networks, are appropriate tools to explore customer value effectively. Though many segmentations algorithm have been presented, their effectiveness depends on detailed data sources of customers. As an example, the customers and their transaction data of a small online fashion clothing shop are very different from those of Amazon. So segmentation methods are needed to design for special online shops.

This paper constructs a customer segmentation model and a customer prediction model for an online fashion bag shop with a clustering algorithm and neural network technology. The real data from more than 15000 customers of the shop over one year are tested. This paper first standardizes customer data, sets and makes correlation analysis to select those main feature variables. After that, the Fuzzy C-Means clustering algorithm is utilized to cluster Customer features. Then probabilistic neural network algorithms are used to build the customer prediction model between customer features and customer types. The accuracy of this model is estimated by test samples. Then this paper employed ANOVA analysis to reveal that the new types have more significant differences than the prior groups based on the VIP system.

RELATED WORKS

Customer Segmentation

Customer segmentation will split the customer base into smaller groups using a variety of unique customer characteristics to help business people. Different feature data have been employed for use segmentation. A common model is based on three variables of RFM (Recency, Frequency, and Monetary), where Recency shows the time since the last customer purchase transaction, frequency indicates the number of purchase transactions in a time, while monetary shows the value of the purchase (Marisa et al., 2019). Christy and Umamakeswari developed a scoring method to evaluate scores of RFM. Some researchers try to develop an RFM model and add some parameters to these three parameters. Khajvand et al. (2011) proposed an extended RFM analysis method with one additional parameter, Count Item. The results of these approaches show that adding count Item as a new parameter to the RFM method makes no difference to clustering result when calculating the Customer Lifetime Value. Cheng and Chen (2009) derived an augmented RFM model, called the RFMTC model (RFM plus Time since first purchase and Churn probability), using the Bernoulli sequence in probability theory. Amine, Bouikhalene, and Lbibb (2015) considered the customer's relation length (L), named LRFM, where L is defined as the number of time periods (such as days) from the first purchase to the last purchase in the database to evaluate the customer's loyalty. In the ride-sharing market, Li et al. (2018) used the length of the membership (L) and replaced the original consumption amount(M) with the two indicators, such as the travel distance(M) and the average value of the discount coefficient (D) enjoyed by the customers in a certain period of time, established a reasonable customer value evaluation model called LRFMD. K-LRFMD to conduct the clustering analysis.

In fact, more online attribute data besides RFM parameters can also be employed for customer characters. Rohm and Swaminathan (2004) presented that online consumer behavior involves various stages of the consumer decision-making process, including introducing problems, information searches, evaluating, and results. In each stage, consumer behavior is greatly influenced by the development of digital technologies, social and cultural changes, and the impact of the closest people on their opinions. Rohm identified four types of online food buyers with different shopping behaviors: convenience shoppers, variety seekers, balanced buyers, and store-oriented shoppers, for example, the convenience of buyers is to save time on online shopping, but various searchers want to have something new when choosing a brand, product or store; this different behavior leads to different choices. In their Analysis, the most common attributes used are location, age, sex, income, lifestyle, and previous purchase behavior. Ortigosa et al. (2014) got the data from some Spanish-speaker customers on Facebook and assumed that customers with similar personalities would show common behavioral patterns when interacting through virtual social networks. They trained a classifier using different machine-learning techniques with the purpose of looking for interaction patterns that predict the customers' personality traits. Inspired by the context, in this research, we employ real online data, including demographic, purchasing data, and return data for segmentation.

Clustering Techniques

There are two basic types of clustering methods, hierarchical and non-hierarchical. Hierarchical clustering is based on cluster analysis which builds a hierarchy of data points as they move into a cluster or out of it. Strategies for this algorithm generally fall into two categories: agglomerative and divisive. The main advantage of Hierarchical clustering is that the output is in the form of a hierarchy (dendrogram), which tells us exactly at which point the clusters merged or split. Hence it is easy to choose and decide the number of clusters that we wish to take by looking at the dendrogram. However, for a large number of observations, its computational speed is very low as compared to the nonhierarchical methods of clustering.

Another most commonly used algorithm is called k-means due to the fact that the letter k represents the number of clusters chosen. First, an initial set of means is defined, and then subsequent classification is based on their distances to the centers. Next, the clusters' mean is re-computed again, and then reclassification is done. This is repeated until cluster means don't change much between successive iterations. Kanavos et al. (2018) proposed a parallel K-Means algorithm based on MapReduce framework for Large Scale Product Recommendation of Supermarket Ware Based on Customer Behavior Analysis, and they utilize a supermarket database and an additional database from Amazon, both containing information about customers' purchases. Marisa et al. (2019) obtained Customer Lifetime Value (CLV) in two customer segments. Grouping uses the K-Means Clustering method based on the LRFM model. Ortigosa et al. (2014) used a two-stage method for clustering. In the first stage, the self-organizing maps method is used to determine the best number of clusters, and in the second stage, the k means the method is applied. Seven hundred thirty customers of the biggest online retailers specialized in electronics and home applications are segmented into nine clusters. However, Tripathi, Bhardwaj & Poovammal (2018) pointed out two major problems in cluster analysis. One is the difficulty to select the number k of clusters, this algorithm does not have the optimal number of clusters, and another major downside is that it depends upon the initial cluster centers. Hassan, Shah, Othman, and Hassan, pointed that FCM is more superior to K means by producing balanced clusters with the random distribution manner. Kaile Zhou and Yang (2020) also compared the effect of cluster size distribution on k-means and fuzzy c-means (FCM) clustering. The results demonstrate that FCM has a stronger uniform effect than k-means clustering and is suitable for massive data. Therefore, this research uses the FCM algorithm for customer segmentation.

Customer Prediction Model

Ortigosa et al. (2014) presented the techniques such as naïve Bayes, K-nearest neighbors, classification trees, and association rules were able to analyze the dataset and have similar performance, they chose classification trees and worked with the Weka data mining toolbox to analyze Facebook data and Predict customer personality.

Stachl et al. (2020) used both linear and nonlinear regression models (elastic net, random forest) to predict personality and examined the extent to which individuals' Big Five personality dimensions can be predicted on the basis of six different classes. They collected data from 624 volunteers over 30 consecutive days (25,347,089 logging events) and evaluated the models using a (nested) cross-validated approach. The results are similar to the customers' digital footprints from their smartphones.

Kanavos et al. (2018) implemented the Map-Reduce model to process the large datasets of a supermarket from Amazon using a distributed and parallel algorithm. They targeted classifying customers according to their consuming behavior and consequently recommended new products more likely to be purchased by them. Results showed that the proposed method predicts with high accuracy the purchases of the supermarket.

Gu et al. (2020) presented that probabilistic neural network (PNN) shows high efficiency when solving pattern recognition and prediction problem since the calculation results are universally reliable. Mohanty and Palo (2020) recommended the PNN model for children's emotion reorganization due to its accuracy level with the statistical features. Therefore, this research uses the PNN network to predict the unclassified customer.

Statistics and Characteristics of Customer Groups

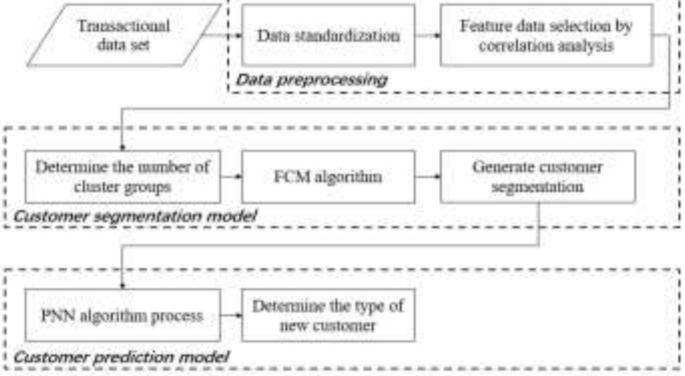
An obvious disadvantage of Neural Network algorithms is that there is a lack of interpretability of the results, i.e., they do not provide the weights and statistical importance of the input variables in the prediction, so in this paper, we try to show the statistical properties of the different groups.

Magatef and Tomalieh (2015) applied the ANOVA analysis on data collected for marketing segmentation and provided some inferences. The result indicated that some variables are insignificant at the confidence level while others are statistically significant as they have high probabilities. Rohm and Swaminathan (2004) employed univariate ANOVA and chi-square tests to identify the behaviors of different facebook customer clusters. They examined several variables (e.g., age, household size, per capita income, gender, and product type) that potentially influence retail shopping and online purchasing behavior of the clusters. The results revealed no differences in age, income, and household size across the four online shopping types. The results also indicated significant differences in purchase behavior across clusters for the following product classes: books and magazines software, etc.

For our data, the test clearly shows the data have unequal variances. So ANOVA based on Games-Howell in SPSS is used to compare all possible combinations of group differences since this method does not require the groups to have equal standard deviations.

RESEARCH MODEL

This paper constructs a customer segmentation model and a customer prediction model with correlation analysis, clustering algorithm, and neural network technology. The data of online shopping and returns from customers of e-commerce platforms are used in the models to profile customers. The Customer segmentation model is constructed with Fuzzy C-Means clustering algorithm, while the customer prediction model is realized by the PNN network. A specific flow chart of the two models is shown in figure 1.


Transactional data set
Data standardization
Feature data selection by correlation analysis
Data preprocessing
Determine the number of cluster groups
FCM algorithm
Generate customer segmentation
Customer segmentation model
PNN algorithm process
Determine the type of new customer
Customer prediction model

Source: This study.
Figure 1: The flow chart of the Customer segmentation model and user prediction model.

In this section, some key steps in the flow chart of the Customer segmentation models are described as follows.

Data Preprocessing 1: Standardize Data Formats

The raw data/attributes of online customers used for customer segmentation models are classified into three formats, including numerical format, ordinal format, and categorical format (Jayarathna, Patra & Shipman, 2015). In order to weaken the impact of different dimensions better, each format of data is standardized by the corresponding method. For numerical format, this paper applies the Min-Max standardization method to standardize the data. For customer features of ordinal format, this study sorts the specific features into n grades according to certain rules, and then assign different sizes of value to features of the different customer according to the order, then use the equation (1) to standardize ordinal features, where x_i is the normalized value of an ordinal feature, i is the rank order of the ordinal feature value and n is the total number of the ordinal feature value formats. $x_i = \frac{i}{n}$ (1)

For category format, this paper designs certain rules to sort the category data and then process the data according to the standardization method of ordinal features. Taking the attribute of customer's location city in the customer data set as an example, the data can be sort into six city levels according to the China city ranking 2020 released by the new first-tier city Research Institute, namely, first-tier city, second-tier city, third-tier city, fourth-tier city, fifth-tier city and not listed city. The cities where customers live are sorted according to the city level, and then the ordinal method is used for standardization.

Data Preprocessing 2: Select Customer Feature Data by Correlation Analysis

The Pearson correlation coefficient and correlation matrix are utilized to study the correlation degree between the customer's raw transaction data and returns data (Abdulhafedh, 2021) and to select those feature variables having a high correlation with the customers' behavior of return, which is cared by the retailer as the customer segmentation features.

The correlation matrix R can be written as equation (2), where $\rho_{ij}\,(i,j = 1,2 \ldots n)$ is the correlation coefficient of attributes x_i and x_j.

$$R = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix} \qquad (2)$$

The representative attributes with a high correlation coefficient in the correlation matrix are selected as the features in the customer segmentation and to construct the prediction model between customer types and customers' shopping behaviors.

The Customer Segmentation Model

Customers can be classified into different types by attribute similarity calculation. In this paper, the FCM algorithm is used to cluster.

In the FCM algorithm, a membership degree u_{ij} between 0 and 1 is used to represent the similarity degree between the i-th cluster group and the j-th customer sample, which meets the following constraint condition:

$$\sum_{i=1}^{c} u_{ij} = 1 \text{,} j = 1, \ldots, n \qquad (3)$$

A membership matrix U with the size of $n \times c$ is composed of n customer samples and c clustering groups. The objective function of the FCM algorithm is shown in equation (4).

$$J(U \text{,} c_1 \text{,} \ldots \text{,} c_c) = \sum_{i=1}^{c} J_i = \sum_{i=2}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \qquad (4)$$

where c_i is the cluster center of the i-th cluster group, d_{ij} is the Euclidean distance between the cluster center of the i-th cluster group and the j-th customer feature sample. m is the membership factor, and the default value is 2. Therefore, by minimizing the objective function J , the best cluster numbers c and u for every sample and their group are obtained.

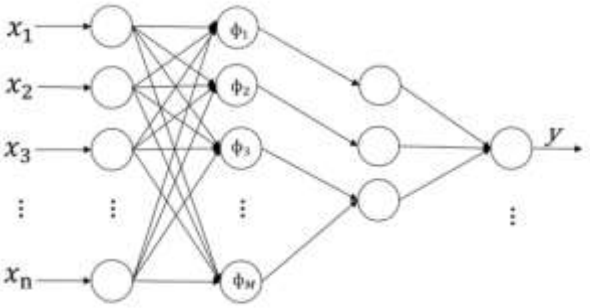
The effect of the clustering algorithm is closely related to the rationality of input parameters. However, the rationality of the selected cluster number c cannot be fully considered in manual selection. Therefore, we use partition coefficient V_{pc} and partition entropy V_{pe} to measure the clustering effect when the cluster number c value is different.

The Customer Prediction Model

After customer segmentation based on the features of purchase and return behavior, we can use a neural network to establish a nonlinear prediction model so that when unlabeled customers appear in the e-commerce platform stores, we can use the model to predict which types/groups of customer segmentation the unlabeled customer may be according to the features data recorded in the stores.

The prediction model based on Probabilistic Neural Network (PNN) is used in this paper. PNN is a neural network designed by the principle of statistics. It belongs to RBF neural network. It has considered both density function estimation and Bayesian decision theory. PNN has a simple structure, easy training process, and fast speed of obtaining results. It is very suitable for real-time processing data and has a wide range of applications (Satheesh & Nagaraj, 2021). It can achieve arbitrary nonlinear approximation.

In this paper, the PNN network structure is with four layers is employed. The customer samples are received by the input layer and then go through the mode layer and the summation layer. Finally, the classification results are output from the output layer. The network structure is shown in figure 2.



Source: This study.

Figure 2: The structure of the PNN neural network.

The process of training and testing the PNN neural network algorithm is as follows:

Step 1: The features data of the customer samples which have been profiled are selected as the input variables x_i , and the corresponding clustering results y of these customer samples are used as the output variables.

Step 2: According to a certain proportion, the customer samples are classified into two categories: the training samples used to train the neural network structure and the test samples used to test the accuracy of the neural network. More details will be introduced in section 4.

RESULT OF CUSTOMER SEGMENTATION AND PREDICTION

Collection and Analysis of Customer’s Data

This research is in collaboration with a flagship store of women's fashion bags on the Taobao e-commerce platform, and the data comes from the store's background data server, however considering the interesting relationship, these data are processed with data masking techniques to obfuscate some sensitive data. Finally, we extract the transaction data of 15858 customers with return records in one year from the back-end database.

These raw transaction data attributes consist of (1) customer basic information, including customer’s name, customer’s member rank, customer’s location; (2) customer return information, including the number of orders returned, proportion of orders returned, amount of order returned; (3) customer purchase information, including buyer's payment payable, seller's actual payment, the total number of orders, buyer's actual payment amount, price of the purchased product, the delay time of payment, the delay time of receiving goods, the total number of product purchased, the average value of product category and product number in a single purchase, discount rate of orders, whether the customer leaves a product review, etc.

Selection of Customer Features

Firstly, the study standardizes the numerical attributes, ordinal attributes, and categorical attributes of the original customer information, transforms them into a data matrix, and calculates the correlation matrix as table 1.

Table 1: Correlation matrix table of customer attributes.

	Member rank	City level	Average value of buyer's payment payable	Maximum value of buyer's payment payable	Minimum value of buyer's payment payable	···	Delivery delay time	discount rate of orders
Member rank	1.00	-0.01	0.02	0.04	0.00	···	0.00	0.01
City level	-0.01	1.00	0.02	0.03	-0.01	···	-0.01	0.01

Average value of buyer's payment payable	0.02	0.02	1.00	0.96	0.97	⋯	0.04	0.25
Maximum value of buyer's payment payable	0.04	0.03	0.96	1.00	0.86	⋯	0.03	0.28
Minimum value of buyer's payment payable	0.00	-0.01	0.97	0.86	1.00	⋯	0.04	0.21
⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮
Delivery delay time	0.00	-0.01	0.04	0.03	0.04	⋯	1.00	0.07
Discount rate of orders	0.01	0.01	0.25	0.28	0.21		0.07	1.00

Source: This study.

From the correlation matrix *R*, it can be observed that the correlation between the number of returns and the total number of orders is high, with a correlation coefficient of 0.60. In addition, the negative correlation between the return proportion and the cumulative consumption amount, the total number of transaction orders, and the number of successful transactions are high, with the correlation coefficients of -0.70, -0.41, -0.80, respectively. The positive correlation between the return proportion and the number of return orders is high, with a correlation coefficient of 0.45. The correlation between the return amount and the total number of transaction orders, as well as the correlation between the return amount and the number of return orders, is relatively high. The correlation coefficients are 0.51 and 0.79.

Considering the correlation coefficients and the representativity of customer features, the study selects 14 main feature variables to construct customer segmentation, including basic customer features, return features, and purchase features which have high correlations with return features (e.g., customer's member level, customer's location, number of order returned, the proportion of order returned, cumulative return amount, order payment amount, cumulative consumption amount, price of the product purchased, a quantity of purchased product, type of purchased product, count of consumption order, whether to participate in pre-sale, payment delay time, Proportion of order discounts, etc.).

Result of the Customer Segmentations Model

Using Bezdek's clustering validity function-partition coefficient and partition entropy (Tekin, Kaya & Cebi, 2021), we can know that 3 is the best number of cluster groups.

Then FCM algorithm is used to cluster the customer return features. The clustering result shows that in 15858 customer data samples, the number of customer samples contained in type 1 (c_1), type 2 (c_2), type 3 (c_3) is 7922, 4840, 3096 respectively, accounting for 50%, 30%, 20% of the total number of samples.

In order to compare the feature variables of different cluster groups, the average values of customer basic feature variables, return feature variables, and purchase variables of customer samples in each cluster group are listed in table 2.

Table 2: The result of customer sample classification.

Clustering group c_i				Type1	Type2	Type3
Number of samples				7922	4840	3096
Customer features	Customer basic features	Member rank		1.04	1.06	1.04
		City level		3.30	3.36	3.33
	Customer return features	Count of return orders		1.00	1.09	2.34
		Proportion of return orders		0.50	0.29	0.59
		Cumulative return amount		126.83	144.19	304.36
	Customer purchase features	Features of consumption amount	Order payment amount/yuan	124.24	129.24	129.35
			Accumulated consumption amount / yuan	124.98	350.05	218.06
			Price of the product purchased / yuan	112.55	110.49	110.77
		Features of consumer products	Product number in a single purchase	1.13	1.26	1.20
			Product kind in a single purchase	1.12	1.20	1.19
		Features of consumption frequency	Total number of consumption order	2.01	3.93	4.08
		Features of pay in advance	Rate of participation in the pre-sale	0.14	0.15	0.16
		Features of paying habit	Payment delay time	51.19	69.53	76.43

		Features of price-sensitive	Proportion of order discounts	0.03	0.01	0.01

Source: This study.

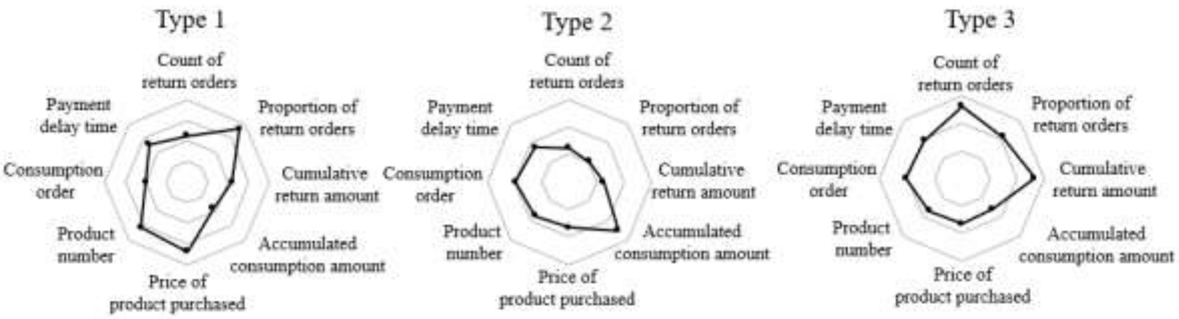
Result of the Customer Prediction Model

After building the customer segmentation based on customer features data, we use a neural network to build a nonlinear prediction model between customer features and customer types and train the neural network. The prediction model based on the purchase features of profiled customers can be used to predict the type of new customers. In the experiment, these features, including cumulative consumption amount, the total number of consumption orders, product number in a single purchase, product kind in a single purchase, are selected as input variables. Customer type is used as an output variable to construct the probabilistic neural network.

In the experiment, the group of return feature data and the group of purchase feature data contain 15000 profiled customers, respectively. Among 15000 customer samples, 90% of customer samples are used as training samples, and 10% of customer samples are used as test samples. The correctness of the trained PNN will be tested with test samples. The test results of the neural network show that the prediction accuracy based on purchase features is more than 90%.

DIFFERENCE ANALYSIS AMONG THREE CUSTOMER GROUPS

According to the Customer segmentation results in table 2, the main features of three types of customers are analyzed and discussed in this section. In this paper, these three types of customers are defined as impulsive customers, economic customers, and cautious customers.



Source: This study.

Figure 3: Typical features of three customer types.

Customer Type 1: Impulsive Customer

In this paper, type 1 customers are defined as impulsive customers who have the minimum payment delay time among the three types; the reason could be they do not want to spend time shopping. The customer samples of type 1 account for 50% of the total samples. It can be seen from the radar chart of eight dimensions in figure 3 that its typical features are fewer consumption orders, fewer return orders, lower cumulative consumption amount, and cumulative return amount than the other two types, but the proportion of return orders is higher, the average price of purchased products is higher, the quantity of purchased products is less, and the payment delay time is less than other two customer types.

According to their fewer consumption orders, fewer return orders, lower cumulative consumption amount, and cumulative return amount, these kinds of customers are referred to have low loyalty to the stores. According to the small number of products in a single purchase, customers of type 1 are referred to only purchase the target products and are not inclined to store goods. According to their short payment delay time, impulsive customers are easy to buy products impulsively, but they are also easier to regret after the purchase, so the probability of return behavior is higher than the other two types.

Customer Type 2: Economical Customer

Type 2 customers are defined as an economical customer because they have the minimum average price of purchased products among the three types. They account for 30% of the total samples. Its typical features are the medium number of consumption orders, fewer number of return orders, a higher amount of cumulative consumption, and the medium cumulative return amount compared with the other two customer types. The proportion of return orders is low, the average price of purchased goods is low, the quantity of products in a single purchase is large, the payment delay time is medium, and the payment enthusiasm is medium.

The customer of type 2 is defined as the economical customer. Considering their fewer number of return orders, the lower average price of purchased products, and a large number of products in a single purchase, economic customers, are referred to think a lot about whether the price is affordable or not when they choose products. They are more sensitive to the price of products and more likely to buy products at one time to get preferential discounts. Besides, taking their medium number of consumption orders into consideration, this kind of customer has a certain degree of loyalty after finding a shop with appropriate prices.

Customer Type 3：Cautious Customer

Customers of type 3 can be defined as cautious customers. According to their high payment delay time, cautious customers are more cautious when buying goods. The typical features of customer type 3 are the relatively large number of consumption orders and return orders, high cumulative return amount, the proportion of return orders, and the payment delay time compared with other customer types.

They don't like to make a hasty decision to buy products. They tend to consider carefully before the payment. Customers of type 3 are referred to require products of high quality, so the proportion of return orders is high. The loyalty of cautious customers is high, and they tend to choose familiar stores for repeated purchase, so the number of consumption orders is large, and the cumulative return amount is high.

Customer Cluster Compare with VIP Groups by ANOVA

There are many statistical tests that are normally used to perform the clustering process. In this paper, ANOVA is employed for various analyses and data processing, including clustering and data mining. The ANOVA is invariably used in comparing more than one means or centroids. In its simplest form, ANOVA provides a statistical test of whether or not the means or centroids of several groups are all equal or not.

We first compare the variable means of traditional clusters, i.e., NON VIP, VIP1, and VIP2, and then we show the comparison among the means of new clusters, i.e., type1, type2, and type3. As shown in Table.3, 4 chosen variables are Consumption order, Accumulated consumption amount, Proportion of return orders, and Payment delay time(s). The reason is that these variables have similar character as RFM. The Payment delay time is a time character instead of the recency (how recently a customer has made a purchase and take the place of R. Accumulated consumption amount in one year is the frequency (how often a customer makes a purchase), and Accumulated consumption amount shows how much a customer spends (monetary). The proportion of return orders is an additional variable to represent how much loss they bring to retailers.

Table 3: One-Way ANOVA（Analysis of Variance）of Existing VIPS.

	(I) Existing customer groups	(J) Existing customer groups	Mean Difference（I-J）	Std Error	Sig
Consumption order	non VIP（mean=2.99）	VIP1（mean=3.02）	-0.032	0.069	0.890
		VIP2（mean=3.51）	-0.514*	0.123	0.000
Accumulated consumption amount	NOn VIP(mean=210.64340)	VIP1（mean=220.852）	-10.209	8.263	0.433
		VIP2（mean=280.311）	-69.667*	16.805	0.000
Proportion of return orders	NOn VIP(mean=0.454846)	VIP1(mean=0.449)	0.005	0.006	0.728
		VIP2(mean=0.422)	0.032*	0.110	0.011
Payment delay time(s)	NOn VIP(mean=61.9223)	VIP1(mean=53.253	8.668	12.066	0.753
		VIP2(mean=63.271)	-1.349	15.872	0.996

*The mean difference is significant at the 0.05 level

Source: This study.

Most of the comparisons are not significant. For example, when considering the variable Consumption order, the results of Comparison 1 (Non-Vip vs. VIP1) is not significant (p = 0.890>0.05), and for Comparison 2 (non VIP vs. VIP2), the difference is significant (p = 0.00<0.05), so the Univariate ANOVA results revealed no differences in Consumption order, Accumulated consumption amount, Proportion of return orders and Payment delay time between the existing shopping types（Non-VIP and VIP1）. This suggests a homogeneous online shopping sample with respect to demographics. For Non-VIP and VIP2, the results show a significant difference in Consumption order, Accumulated consumption amount, and Proportion of return orders. Their mean differences are -0.514, -69.667, and 0.032. However, these differences are less than the differences between the new types as 2.077, 104.933, and -0.075 (in the latter table). This demonstrates that the degree of differentiation is still not high enough in existing clusters. The one-way ANOVA statistics applied to new groups are listed in Table 4. It was very clear from that statistic test that four chosen variables were significantly different across the three final clusters obtained by k-means clustering, as shown.

Table 4: One-Way ANOVA（Analysis of Variance）of new Groups.

	(I)proposed used groups	(J)proposed used groups	Mean Difference（I-J）	Std Error	Sig
Consumption order	TYPE1（mean=2.010）	TYPe2（mean=3.920））	-1.906*	0.018	0.000
		Type3（mean=4.090）	-2.077*	0.023	0.000
Accumulated consumption amount	TYPE1（mean=124.878356）	TYPE2(mean=344.543）	-219.664*	2.261	0.000
		TYPE3(mean=229.781）	-104.903*	2.635	0.000
Proportion of return orders	TYPE1(mean=0.499)	TYPE2(mean=0.297)	-0.202*	0.001	0.000
		TYPE3(mean=0.574)	-0.075*	0.002	0.000
Payment delay	TYPE1(mean=51.198)	TYPE2(mean=68.589)	-17.390*	6.802	0.029

time(s)		TYPE3(mean=77.648)	-26.449*	5.863	0.000

*The mean difference is significant at the 0.05 level

Source: This study.

CONCLUTIONS AND IMPLICATIONS

Based on the customer data related to return features in e-commerce platform stores, this paper constructs a customer segmentation model and a customer prediction model between customer features and customer types. In the experiment, the customer segmentation classifies customer samples into three types, including impulsive customer, economical customer, and cautious customer. These two models can also be promoted to other products selling in the online stores to realize customer type segmentation quickly and well, so as to analyze the purchase habits and preferences of store customers. They also provide some basis for e-commerce platform stores to make differentiated sales service and reduce the high return rate.

However, there are still some limitations in the customer segmentation models designed in this paper. Follow-up work can be further studied from the following aspects: first, the algorithm used in the models still has its own defects. FCM algorithm needs to determine the number of clustering groups in advance, so it is difficult to apply to dynamic data. New algorithms can be used to improve the clustering result in the future. Secondly, the accuracy of the PNN neural network algorithm in the customer prediction model may not be stable in different data environments. Therefore, it is necessary to carry out more experiments to test the accuracy of different neural network algorithms in different environments. Finally, it is very important to combine the theoretical models with e-commerce practice. Only in this way can the theoretical method really help the actual development of the e-commerce platform.

REFERENCES

Abdulhafedh, A.(2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, 3(1),12-30. doi:10.12691/jcd-3-1-3

Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, 7(1), 1-23.

Amine, A., Bouikhalene, B., & Lbibb, R. (2015). Customer segmentation model in e-commerce using clustering techniques and LRFM model: The case of online stores in Morocco. *International Journal of Computer and Information Engineering*, 9(8), 1993-2003.

Cheng, C., & Chen, Y. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, 36(3), 4176-4184.

Cullinane, S., & Cullinane, K. (2021). The logistics of online clothing returns in Sweden and how to reduce its environmental impact. *Journal of service science and management*, 14(1),72-95.

Gu, Y., Zhang, Z., Zhang, D., Zhu, Y., Bao, Z., & Zhang, D. (2020). Complex lithology prediction using mean impact value, particle swarm optimization, and probabilistic neural network techniques. *Acta Geophysica*, 68(6), 1727-1752.

Hassan, A. A. H., Shah, W. M., Othman, M. F. I., & Hassan, H. A. H. (2020). Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks. *International Journal of Electrical & Computer Engineering* (IJECE), 10(2), 1515-1523.

Jayarathna, S., Patra, A., & Shipman, F. (2015, June). Unified relevance feedback for multi-application user interest modeling. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 129-138).

Jouzdani, J., Shirouyehzad, H., Maaroufi, N., & Javaheri, A. (2020). Identification, ranking and clustering of electronic banking services based on customer satisfaction: case study in bank industry. *International Journal of Services, Economics and Management*, 11(2), 167-190.

Kanavos, A., Iakovou, S. A., Sioutas, S., & Tampakas, V. (2018). Large scale product recommendation of supermarket ware based on customer behaviour analysis. *Big Data and Cognitive Computing*, 2(2), 11.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57-63..

Lee, S., & Lee, K. C. (2020). "Comparative study of service quality on VIP customer satisfaction in internet banking: South Korea case." *Sustainability*,12(16): 6365.

Li, H., Yang, X., Xia, Y., Zheng, L., Yang, G., & Lv, P. (2018, July). K-LRFMD: method of customer value segmentation in shared transportation filed based on improved K-means algorithm. *In Journal of Physics: Conference Series* (Vol. 1060, No. 1, p. 012012). IOP Publishing.

Li, K., Deolalikar, V., & Pradhan, N. (2015, October). Mining lifestyle personas at scale in e-commerce. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1254-1261). IEEE.

Lysenko-Ryba, K., & Zimon, D. (2021). Customer behavioral reactions to negative experiences during the product return. *Sustainability*,13(2), 448.

Magatef, S. G., & Tomalieh, E. F. (2015). The impact of customer loyalty programs on customer retention. *International Journal of Business and Social Science*, 6(8), 78-93.

Marisa, F., Ahmad, S. S. S., Yusof, Z. I. M., Hunaini, F., & Aziz, T. M. A. (2019). Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using K-means clustering and LRFM model. *International Journal of Integrated Engineering*, 11(3).

Mohanty, M. N., & Palo, H. K. (2020). Child emotion recognition using probabilistic neural network with effective features. *Measurement*, 152, 107369.

Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of computer and System Sciences*, 80(1), 57-71.

Rohm, A. J., & Swaminathan, V. (2004). A typology of online shoppers based on shopping motivations. *Journal of business research*, 57(7), 748-757.

Satheesh, M. K., & Nagaraj, S. (2021). Applications of artificial intelligence on customer experience and service quality of the banking sector. *International Management Review*, 17(1), 9-86.

Sivaguru, M., & Punniyamoorthy, M. (2020). Modified dynamic fuzzy c-means clustering algorithm-Application in dynamic customer segmentation. *Applied Intelligence*, 50(6), 1922-1942.

Song, X., Liu, M. T., Liu, Q., & Niu, B. (2021). Hydrological cycling optimization-based multi objective feature-selection method for customer segmentation. *International Journal of Intelligent Systems*, 36(5), 2347-2366.

Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T. & Hussmann, H. (2020). Predicting personality from patterns of behavior collected with smartphones. In *Proceedings of the National Academy of Sciences*, 117(30), 17680-17687.

Tekin, A. T., Kaya, T., & Cebi, F. Customer lifetime value prediction for gaming industry: fuzzy clustering based approach. J*ournal of Intelligent & Fuzzy Systems*, (Preprint), 1-10.

Tripathi, S., Bhardwaj, A., & Poovammal, E. (2018). Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology*, 7(3.12), 802-807. doi: 10.14419/ijet.v7i3.12.16505

Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, 61, 102588.

Zhou, K., & Yang, S. (2020). "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering." .*Pattern Analysis and Applications* 23(1): 455-466.