Research on Application of Data Mining Algorithm Based on RFM in Customer Segmentation of Logistics Enterprises

Xueyuan Wang, Wenyao Qu*, Jiajun Dang
Institute of Management Engineering, Zhengzhou University, No. 100 Science Street, Zhengzhou, China
wangxueyuan@zzu.edu.cn
**Corresponding author: quwenyaohust@163.com*

Keywords: logistics enterprises, K-means algorithm, customer segmentation

Abstract: Market competition is becoming increasingly fierce, bringing serious challenges to the survival and development of logistics enterprises. Today, enterprises are unable to win only by the products in the market competition. Moreover, the enterprise focuses on the customers. To give differentiated service strategies based on classification of customers is an important part of implementing customer relationship management. Traditionally, to classify customers in accordance with the amount of consumption ignores two important factors between customers and business, including trading time and trading frequency, thus the accuracy rate is lower. In this paper, the RFM model is introduced into the process of classifying customers, and each attribute is given a certain weight. After the data are normalized by statistical knowledge, the RFM value of each customer is calculated. K-means method is used for classification. Then according to the clustering results, some different services and marketing strategies are adopted to different types of customers. Thereby the core competitiveness of enterprises is developed.

1. INTRODUCTION

In twenty-first Century, the development of all enterprises depends on two aspects, which are enterprise brand and customer satisfaction (Xinchen Wang, 2016). More and more organizations are implementing customer relationship management system to support employee service activities, and require their employees to use these systems (HSIEH J. J. Po-An, 2012). For the logistics industry, mostly there are 7 ways of logistics and transportation, namely: railway transportation, marine transportation, inland transport, air transport, pipeline transportation, road transport and multimodal transport. There are tens of thousands of logistics enterprises running in each transportation mode whatever large and small, thereby the competition between the industry is extremely fierce. If the logistics enterprises want to be invincible, they must improve customer satisfaction. In the field of logistics, there are only a few large logistics enterprises implementing customer relationship management, management and application of most enterprise customer relationship has been not mature. And the customer segmentation as the key to the successful implementation of customer relationship management step, its value has not been well reflected. The rise of online shopping makes the logistics industry market demand and competition intensified at the same time, the extend of customer choice space determines the customer requirements of the enterprise is enhanced, thus it is vital to take effectively customer classification and personalized marketing strategy .

Nowadays, both entrepreneurs and scholars at home and abroad have recognized the value of customer segmentation, and have laid the foundation for effective implementation of customer relationship management with the help of data mining technology. In our nation, subject domain is divided and data model was built by Dong Chen according to FS-LDM model, and then realized the prediction model of customer churn and the dynamic model of customer segmentation by using clustering technology and logistic regression techniques (Dong Chen, 2013). Firstly, square statistics was used by Chuanyu Ji to analyze and select attributes and K-means algorithm was used for quantization of each attribute value. Then, customers will be divided into three categories according to the DBSCAN algorithm based on the density. Finally, he used FP-tree based on two-way clustering algorithm to cluster the three types of customers to obtain more detailed information. (Chuanyu Ji, 2017). Study abroad not only confined to the traditional customer classification according to customer value, but the introduction of some other indicators on the customer segmentation. Followings are some representative research: Lazer classified customers based on the way of their life (LAZER, WILLIAM, 1963). Malaika sums up two elements which are customer attitudes and customer psychological quality into standards of customer classification (MALAIKA B, 2005). Husghes establishes the R (recentcy) F (Frequency) M (Monetary) model to describe the customer classification result (HUGHES A, 1994). Bradley improves the initial clustering point of K-means algorithm, which makes the classification structure more clear (BRADLEYP, 2010).

Data mining is an important tool to effectively implement customer relationship management, accompanied by the whole process of customer relationship management. Only by applying data mining technology, can enterprises find the rules of customer behavior from the complex data in the database, so that the business can operate well under the guidance of data mining results.

2. ANALYSIS OF THE NECESSITY OF CUSTOMER SEGMENTATION

2.1 The main content of logistics enterprise customer relationship management

The customers with a wide range and a large number of logistics enterprises bring customer segmentation a certain degree of difficulty. The standards Customer segmentation are not unique, one customer may belong to the different types according to different standards. Reasonable customer classification plays an important role in daily operation of enterprises. According to statistics, in modern enterprises, 80% of the profits are provided by 20% of the big value customers (Gang Ma, 2012), while most of the remaining 80% of the customers contribute little to the enterprises; even consume the resources of the enterprise. The key to the successful implementation of customer relationship management is to identify the big value customers who provide the 80% profit, and maintain the relationship between them. Customer segmentation is the beginning of customer relationship management of enterprises, and also is the foundation of personalized marketing and customer retention, so the premise of the successful implementation of customer relationship management of logistics enterprises is the effective and comprehensive classification of customers. For enterprises, the implementation of customer relationship management system mainly has the following 4 aspects (Xin Wang, 2015):

(1) Customer identification: the process of finding potential customers based on some experimental observation.

(2) Customer segmentation: customers will be divided into different groups according to customers' different characteristics or behaviors. Customers in the same group have same or similar characteristics and behaviors.

(3) Personalized marketing: to implement different marketing strategies based on the preference and purchase behavior of customers.

(4) Customer retention: to take some measures to stimulate customers to consume before customers lose, thus prevent them from loss.

2.2 Customer identification and segmentation

Customer identification includes identifying potential customers and identifying customer value. Customer segmentation standards are not unique, so customer classification results are not same according to the different indicators. For example, customers can be divided into general customers, key customers, VIP customers in accordance with customers' value; customers can be divided into potential customers, new customers and old customers according to the customer's trading amount. Therefore, customer identification is a prerequisite for the segmentation of customers, but also a part of the segmentation of customers.

2.3 Customer classification and personalized marketing

However, in the actual business activities, the needs of customers naturally vary from different customers' identity. It not only costs too much for us to statistical each customer's needs, but also is very difficult to achieve, so firstly customers with the similar features should be summarized in a group, to form a set. Then develop different marketing plans according to different sets of customers. Analysis of association rules can be applied for the orders data of customers to find purchase association rules of different products. Clustering techniques help us summarize what products customers have preferences, neural network and regression method can be used to establish the prediction model of customer to purchase products.

2.4 Customer classification and customer retention

Customer retention is to remain old customers. If customer identification is the beginning of customer relationship management, customer retention runs through the whole process of customer relationship management. In order to reduce the operating costs of enterprises, there are parts of customers need not reserve. The old customers with retention can create value for the enterprise, but it is unknown whether the new customers can contribute to the enterprise. Therefore, we should firstly classify customers to find these old customers who need to be retained. Due to the increase of market competition, customers loss phenomenon exists in each enterprise, the application of data mining technology can not completely solve the problem of the loss of customers, mainly establish a customer churn prediction model according to the existing data in the database, and analyze the factors which may lead to the loss of customers, so as to alert raise for the potential loss of customers, timely take some certain measures of improvement and maintenance.

3. METHODS RELATED TO CUSTOMER SEGMENTATION

3.1 Clustering technique

The application of data mining technology in customer classification has been gradually mature in recent years. Using data mining technology to classify customers, we can use statistical methods, or clustering or classification technology. Commonly, clustering technology is used.

What is the difference between classification technology and clustering technology? When we choose to use classification technology, we know how many categories the customers will be divided into or what standards to be used to customer classification, and then choose customer data according to the attributes affecting the classification results; finally we get the correct classification results. Before we use clustering technology to classify customers, we don't know what kind of customers we can divide into, but we know how many categories customers can be divided into. When the customers' data are gathered naturally, the data in

each cluster is analyzed, and the similarity of customers in the same type is summed up. The common clustering algorithms are k-means algorithm, k-medoids algorithm, and birch algorithm and so on. To use clustering technology, we need to understand the following three basic concepts:

(1)Dissimilarity, which is used to represent the difference between two objects, and the calculation method is related to the attribute values of the two objects. In particular, the dissimilarity of numerical data is usually described by distance, and then is expressed in the form of dissimilarity matrix:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix} \quad (1)$$

Explanation: d (i, j) represents the distance between the object i and j, that is, the dissimilarity. In general, d (i, j) >0, and d (i, j) =d (j, i), d (i, i) =0.

(2) Distance, the commonly used expression method is Euclidean distance. It is defined as:

$$\sqrt{\left|X_{i1} - X_{j1}\right|^2 + \left|X_{i2} - X_{j2}\right|^2 + \dots + \left|X_{in} - X_{jn}\right|^2} \quad (2)$$

Where i = (x_{i1}, x_{i2},…, x_{in}) and j= (x_{j1}, x_{j2},…, x_{jn}) are objects of two n dimensions.

(3) The average distance between two clusters is the average length between two objects in two clusters.

3.2 RFM model

Traditionally, the amount of consumption is regarded as the value of customers and customers are classified by their value. Facts have proved that there are many shortcomings in this method of discrimination. The Arthur Hughes, a database marketing research institute in the U.S.A, proposed to increase two properties that are the amounts of purchase and frequency of purchase as evaluating indictors of customer value. Then we called it RFM model, which R represents Recency, the last trading time until now, and F is on behalf of Frequency, the frequency of customer transactions in the most recent period, and M is on behalf of M Monetary, the transaction amount of customers in the most recent period. The traditional customer relationship management focuses on the distinction between the total amounts of customer consumption, while the RFM model pays more attention to the study of customer purchase behaviors. In the process of applying the RFM model, three attributes are assigned different weight values or grouped according to the corresponding rules, which can divide the customers into different levels. The roles of the RFM model are (Lan Xiong, 2017):

(1) Divide customers in different levels, such as VIP customers, important customers and general customers and so on. Then make different marketing plans for different levels of customers, and make different promotions.

(2)Find out lost or dormant members, and make solutions to activate these customers.

(3) Take a variety of marketing methods, such as SMS, telephone, mail, etc. to identify VIP customers in the marketing process to establish a model.

(4) Pay attention to the old customers; maintain the relationship with the old customers, so as to improve customer satisfaction and loyalty.

4. APPLICATION OF K- MEANS METHOD BASED ON RFM MODEL IN CUSTOMER SEGMENTATION

4.1 K- means method

K-means method often writes K-means algorithm, which belongs to a clustering method based on partition. The central idea of K-means algorithm is(Bin Zhang, 2017): firstly, select first few cluster centers in accordance with the relevant methods; secondly, let the other points gather to its nearest cluster center, to complete the preliminary classification; then adjusted classification does not meet the requirements according to the recently distance principle, until to meet the conditions finally. K-means algorithm is based on the average value of the objects in the cluster, and the output is K clusters. The advantage of K-means algorithm is that it is easy to implement and easy to understand; and the disadvantage is that if there are outliers in the data, it will affect the quality of clustering. The iterative steps of the K-means algorithm are as follows:

(1) Determine the number of clusters K, and then the database or data set that needs to be classified is introduced.

(2) Randomly select K objects as the initial cluster centers.

(3) Let other objects cluster to cluster centers so as to complete the initial classification.

(4)Calculate the average distance value of the objects in the cluster, and reclassify the objects that do not meet the requirements.

(5) Calculate the average distance value of the clusters again, that is, the average of the objects in each cluster.

(6) When the above average values are no longer changed, the clustering is finished and the classification results are obtained.

4.2 Application steps of data mining

4.2.1 Data extraction and processing

Extracting data is the first step in modelling, and only actual data can be used for analysis. The data sources of customer information of logistics enterprises mainly include customer database and expressed information of orders. In this case, take Anneng logistics company as an example, and take the June 1, 2015 as the statistical time node, the order data before the time node in the statistical database is removed, and the irrelevant attributes are deleted. Part of the order data is shown in the table1 below.

Table 1 One of the order data sheet

Company name	time	number of orders	amount of trade
Huitianli trading limited company	2014.12.29	2724	2220000

The second step is to convert the data in the database or data warehouse into intermediate data by mathematical algorithm, and then convert it into the modelling data which can be used for analysis through factor analysis. The list data is the recorded customer initial data. These data are large and complex, even scattered, and cannot be used to model, so the list data must be converted to intermediate data. Summarizing the list data is the intermediate data. The conversion rules are shown in table2:

Table 2 Standardization rules based on the RFM model

Attributes	Weights	Standardization rule
R'	0.3	*R'*=(max-x)/(max-min)
F'	0.2	*F'*=(x-min)/(max-min)
M'	0.5	*M'*=(x-min)/(max-min)

Where x represents the original data value, max represents the maximum value in the original data, min represents the minimum value in the original data, and x 'represents the normalized data. According to the above conversion rules, the order data is quantified, which is shown in table 3:

Table 3 Standardization rules based on the RFM model

Company name	R'	F'	M'
Huitianli trading limited company	0.9787	1	0.0746

4.2.2 Operation and result analysis of data

In this case, the software SPSS which is very good in the field of data mining is used to implement clustering algorithm and the running results are sorted out as follows:

Table 4 Traditional clustering algorithm SPSS classification results

	Clustering				
	1	2	3	4	5
Value	0.80	0.98	0.50	0.04	0.18
Percentage (%)	13.98	4.92	46.97	4.52	11.56

Regarding the amount of consumption as the customer's value to the enterprise, we divided the customer into 5 categories in accordance with the amount of consumption of customers, and from left to right are the important customers, VIP customers, major customers, low value customers, and general customers. Customer classification of RFM model based on the results is shown in table5.

We call the second kind of customers VIP customers, although their amount occupies a small proportion of the total customer, but this kind of customer interact with enterprise frequently, and their purchasing ability is strong, the value for enterprise is the biggest. Enterprises should develop more personalized service strategies for these customers, and highlight their value, so that they have priority and privileges in services and products, such as priority delivery. The discount rate is greater, so that they feel out of the ordinary. The fifth kind of customer consumption amount is highest, their amount proportion is bigger, but their recent interaction with the enterprise is very few, therefore enterprise should take the corresponding measure to stimulate these kinds of customers, and prevent the customer outflow.

The first type of customers interact with enterprises frequently in recent time, they have a largest number but the amount of their consumption is relatively small, enterprises also need to maintain good relationship with them to stimulate them consume, making them into an VIP client. The recent transaction amount and transaction frequency of the fourth kind of customers, which are called general customers are few, we conclude they are possibly the new customers. The third type of customer interact with enterprise seldom, it is likely that they are some scattered customers or have been lost, for such customers, there is no need to waste too much energy to maintain them, just some routine maintenance is enough.

Table 5 Clustering results based on RFM model

	Clustering				
	1	2	3	4	5
R'	0.9441	0.8910	0.1809	0.9391	0.0000
F'	0.8419	0.1744	0.0282	0.1200	0.1961
M'	0.0203	0.7832	0.1050	0.1412	0.9550
Value	0.4617	0.69378	0.11241	0.3763	0.51672
P (%)	54.4587	5.2626	12.4114	4.6824	15.2782

For the cluster classification result based on RFM model, customers are divided into 5 categories according to the total value of RFM, and from left to right are the main customers, VIP customers, low value customers, general customers and key customers, comparison of two classification results is shown in Figure 1:

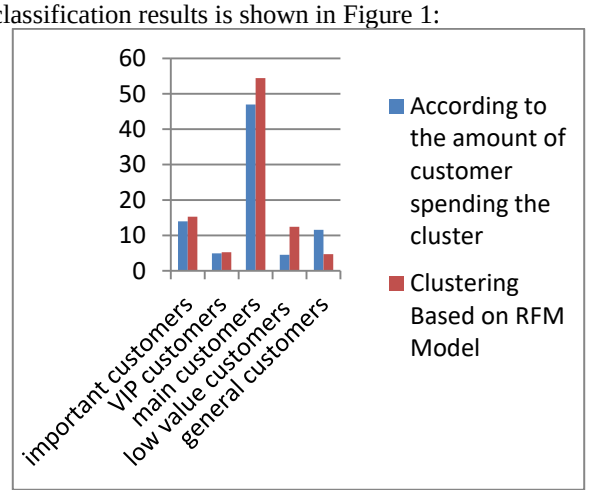

Fig. 1 Comparison of two clustering results

From the figure above, we can see that the segmentation result with introduction of RFM model in clustering analysis is more accurate for identifying VIP customers, important customers and major customers. The amount of VIP customers changed from 4.92% to

5.26%, and the amount of important customers changed from 13.98% to 15.27%, that is to say, the traditional classification method omitted 0.34% of the VIP customers and 1.29% of the important customers. Therefore, the introduction of RFM model in customer segmentation has important significance to identify the VIP customers and important customers. The interpretation of customer clustering segmentation results based on RFM model is following:

5. CONCLUSIONS

For the logistics enterprises, the customer is the fundamental survival and profitability. However 80% of the enterprises profit is from 20% of the customers with the big value. Practice shows that the traditional way of customer classification has some shortcomings, for that there are omissions for VIP customers and key customers, and unable to meet the actual needs of the development of enterprises. The key of implementation of customer relationship management is to improve the existing technology to adapt to the pace of development, and then accumulate customer knowledge from seemingly out of order and massive customer information, so that potential value can be find by using data mining technology. By combining the RFM model and the K-means algorithm, the customer classification error can be reduced, the work efficiency can be improved, the operating cost can be reduced, and the core competitiveness of the enterprise can be improved.

ACKNOWLEDGEMENTS

This research is supported by National Natural Science Foundation of China (Grant No. U1604150), Humanities & Social Sciences Research Foundation of Ministry of Education of China (Grant No. 15YJC630148), Distinguished Young Teacher Development Foundation of Zhengzhou University (1421326092), and Key Research Foundation of University Education in Henan province (17A520058). The authors would like to thank the editors and anonymous referees for their careful and fruitful comments to improve the quality of this paper.

REFERENCES

[1] Xinchen Wang, 2016. The application of web data mining in enterprise database marketing and customer relationship management, Information and Computer (theoretical), 4(22): 155-156.

[2] HSIEH J. J. Po-An, ARUN RAI, STACIE PETTER., 2012. Impact of user satisfaction with mandated CRM use on employee service quality, MIS Quarterly, 12(4): 1065.

[3] Dong Chen, 2013. Research on application of data warehouse and data mining technology in CRM system, Tigital Technology and Application. (08): 66.

[4] Chuanyu Ji, Lili He, 2017. Research on customer segmentation based on two-way clustering methods, Industrial Control Computer. 30(09): 107.

[5] LAZER, WILLIAM, 1963. Life Style Concept and Marketing toward Scientific Marketing, American Marketing Assn. Chicago.

[6] MALAIKA B, MAGGIE G, BERT W, et al, 2005. Segmenting Internet Shoppers Based on their web-usage-related lifestyle: Across-cultural Validation, Journal of Business Research, 6(58): 79-88.

[7] HUGHES A, 1994. Strategic Database Marketing: the Master Plan for Starting and Managing A Profitable Customer Based Marketing Program, Irwin Professional Pub, Beijing.

[8] BRADLEY P, STEVEN T, et al, 2010. Minimal incision surgery as a risk factor for early failure of total hip arthroplasty, Clinical Orthopaedics and Research, 7(89): 421-423.

[9] Gang Ma, 2012. Customer Relationship Management, Northeast University of Finance and Economics Press. Liaoning, 2nd edition.

[10] Xin Wang, Wen Xue, 2015. Application of Data Mining in Customer Relationship Management System, Journal of Northeast Dianli University, 35 (04): 73-78.

[11] Lan Xiong, Bing Gao, 2017. Study on customer segmentation based on RFM multi-level customer value model, Commercial Economics Research, (05): 55-57.

[12] Bin Zhang, Qiyuan Peng, 2017. Study on segmentation method of Chinese railway freight customers based on KFAV, Journal of Transportation Systems Engineering and Information Technology, 17 (03): 235-242.