International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2022): 7.741

Airline Customer Value Analysis Based on Entropy Weight Method and WKmeans Clustering Algorithm

Saixin Wu[1]

[1]North China Electric Power University, School of Control and Computer Engineering
No. 2 Beinong Road, Changping District, Beijing, China
18010466972[at]163.com

Abstract: ***As the industry becomes more competitive, airlines are placing more emphasis on customer experience and personalized services for different customers in their marketing strategies. This requires us to make accurate customer segmentation so that we can target our limited resources to different types of customer groups to maximize the benefits. Since customer groups are not marked in advance, this problem is a typical unsupervised problem. In this paper, based on the traditional LRFMC customer analysis model, we propose a clustering analysis method combining entropy weight method and WKmeans algorithm to achieve customer classification, and finally give corresponding marketing strategies for different types of customers.***

Keywords: LRFMC model, entropy weight method, WKmeans algorithm, data mining

1. Introduction

In today's era, the Internet is becoming more and more connected to all walks of life, and with it comes massive amounts of data and information. To a large extent, the value that Big Data can bring to us depends on our ability to obtain information from the data, and the magnitude of the value is reflected in the level of our data mining techniques. In order to discover and exploit the potential value of this data, research on various data mining tools has received increasing attention [1].

Clustering analysis is an important subfield of data mining, which can divide unlabeled objects into different groupings based on similarity, which corresponds to unsupervised learning in machine learning [2]. As a popular research direction in recent years, it has been applied to numerous cross-cutting areas. In the field of biology, a large amount of gene expression data exists in public databases, and cluster analysis has been used as an automated analysis tool to explore unknown gene functions, perform automatic classification of pathological features, and so on, in order to obtain useful information. In agriculture, the genetic diversity of germplasm resources is of extreme importance, and the use of cluster analysis and other agricultural tools can be used to analyze different germplasm resources for traits and avoid optional blindness. This paper is to apply the clustering algorithm to the marketing strategy of airline companies, and discover the different values of customers with different characteristics through cluster analysis to provide a decision basis for airline companies' operation.

2. Model and Algorithm

2.1 LRFMC Model

In the field of customer relationship management, the RFM model [3] is arguably the most widely used. It is an important way to rate the value of customers, where the three elements of the most recent consumption (R), the frequency of consumption (F), and the total amount of consumption (M) constitute the best indicators for data analysis.

The subject of this paper is airline customers, so it is necessary to combine the well-established RFM theories with the airline realities. This paper adopts the LRFMC model, which uses five indicators - length of customer's membership (L), consumption interval (R), consumption frequency (F), total flight miles (M) and mean value of cabin discount coefficient (C) - as indicators for airlines to identify customer value. The meaning of each indicator is shown in Table 1.

Table 1: Meaning of LRFMC model indicators

	Indicator Meaning
L	Interval between joining and current time
R	The interval between the latest consumption time and the current time
F	Number of rides in the observation time window
M	Number of miles flown within the observation time window
C	Average value of discount factor corresponding to class of cabin

2.2 Entropy Method

The value of data is influenced by various factors, and the key to weighing these factors for value analysis lies in the assignment of index weights. The commonly used assignment methods are divided into two categories: subjective assignment and objective assignment. Delphi method and hierarchical analysis are commonly used subjective assignment methods, while principal component analysis, coefficient of variation method and entropy method are commonly used objective assignment methods [4].

The results obtained by the subjective assignment method are easily influenced by subjective factors and are not convincing [5]. Although the two methods, principal

component analysis and coefficient of variation method, improve the objectivity of the assessment process to a certain extent, they both have their own shortcomings. In view of the shortcomings of the above methods, the weight calculation method used in this paper is the entropy weight method. Based on the theoretical basis of information entropy, the entropy weight method uses the intensity and unevenness of the data itself to reflect the importance of the indexes, and can ensure that the assigned index weights can reflect the vast majority of the original information [6].

The entropy method is calculated as follows.

Step 1: Data normalization process.

The entropy weight method is calculated on the premise that there is a normalized evaluation matrix. Assuming that there are n objects and m evaluation indicators in this evaluation system, the matrix takes the following form.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \end{bmatrix} \quad (1)$$

Step 2: Find the information entropy of each indicator.

The information entropy of the j-th indicator is calculated as

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij} \ln p_{ij} (j = 1,2,...,m) \quad (2)$$

where

$$p_{ij} = \frac{s_{ij}}{\sum_{i=1}^{n} s_{ij}} (j = 1,2,...,m) \quad (3)$$

Step 3: Calculation of variability coefficients and weights.

The coefficient of variability for the j-th indicator is calculated as

$$\alpha_j = 1 - E_j (j = 1,2,...,m) \quad (4)$$

The entropy weight of the j-th indicator is calculated as

$$W_j = \frac{\alpha_j}{\sum_{j=1}^{m} \alpha_j} (j = 1,2,...,m) \quad (5)$$

2.3 WKmeans Algorithm

Clustering is the process of dividing a set of objects into clusters, where objects in the same cluster are more similar than objects in different clusters according to defined criteria [7]. The most widely used of all clustering algorithms is the Kmeans algorithm which is a shape-center based division method [8]. The algorithm uses the distance between samples as a measure of similarity between them, and the closer the distance between two samples, the higher the similarity, and the more likely they are to belong to the same subset.

The Kmeans algorithm tables the clusters by calculating the mean of all sample points within the same cluster to obtain the shaped centroid of the cluster. The number of clusters is represented by the constant K, which is specified before training the dataset. According to the known number of sample centers, K cluster centers are randomly selected within the data set, and the most similar group is found by calculating the distance to each center point to divide the sample points into them, so that a subset is obtained according to the initial cluster center. The shape centers of the sample points within the subset change, so it is necessary to recalculate the shape centers, divide the sample points again, and repeat the process until the shape centers are stable.

The whole process of Kmeans algorithm is actually an optimization process with the following objective function.

$$J = \min \sum_{j=1}^{K}\sum_{i=1}^{n} dist(x_i^j, c_j) \quad (6)$$

where n is the total number of sample points in the dataset, K is the number of cluster centers, c_j is the position of the j-th center, and dist () is the distance from the i-th sample point belonging to the j-th cluster to its cluster center.

The importance of different indicators in the actual situation is different, and the distinction of the importance of indicators is not reflected in the above Kmeans algorithm idea. Thus, Joshua Zhexue Huang et al. proposed the WKmeans algorithm [9] in 2005. The core idea of this paper is to initialize a weight value for each feature dimension, and by the time the objective function converges, the weight corresponding to the noise dimension will converge to 0, thus making it possible to ignore the effect of the noise dimension as much as possible when calculating the distance between samples.

The objective function of WKmeans algorithm is as follows.

$$J = \min \sum_{j=1}^{K}\sum_{i=1}^{n} \omega_j^{\beta} dist(x_i^j, c_j) \quad (7)$$

subject to

$$\sum_{j=1}^{m} \omega_j = 1 \quad (8)$$

Compared with the traditional Kmeans algorithm, the WKmeans algorithm simply adds a weight parameter to the objective function, which works by calculating the weighted

distance sum of each dimension when minimizing the whole intra-cluster distance, i.e., the influence of each dimension on the clustering results is adjusted by different weight values. And, when β = 0, the objective function (7) also degenerates to the objective function (6) of the Kmeans clustering algorithm.

3. Experiment and Analysis

3.1 Data Processing

The dataset selected for this experiment is the customer information of an airline, which contains a total of 62, 988 detailed records, and each customer record includes 44 attribute tags related to information such as basic customer information, flight information, and point's information.

After exploratory analysis of the data table, some apparently illogical outliers were found, so the first step in data processing was to eliminate the records containing outliers from the table to improve the data quality. After data cleaning, 62, 043 records remained in the table.

Then, on the basis of the original data table, data extraction and calculation were performed to construct the five indicators needed for the LRFMC model. In this way, an original evaluation matrix with dimension 62043*5 is formed.

Finally, the data are standardized. This step is to avoid the influence of the variability of each assessment index in terms of magnitude and other aspects on the data value assessment results, and the specific standardization formula is shown in (9) and (10).

$$b_{ij} = \frac{a_{ij} - (a_{ij})_{\min}}{(a_j)_{\max} - (a_j)_{\min}} \quad (9)$$

$$b_{ij} = \frac{(a_{ij})_{\max} - a_{ij}}{(a_j)_{\max} - (a_j)_{\min}} \quad (10)$$

Where, b_{ij} is the normalized value of the ith data over the j-th indicator, and the value of b_{ij} is between [0, 1]; $(a_{ij})_{max}$ and $(a_{ij})_{min}$ are the maximum and minimum values of the j-th indicator in all samples, respectively. Since there are two kinds of indicators, positive and negative, the larger the value of the positive indicator the better, and the smaller the value of the negative indicator the better. Therefore, the two indicators need to be processed separately: equation (9) is chosen to standardize the positive indicators; equation (10) is chosen for the negative indicators. Table 2 shows an example of the first five rows of the matrix after the normalization process.

Table 2: Example of standardized data

L	R	F	M	C
0.7620	1.0	0.9857	1.0	0.6053
0.7265	0.9917	0.6540	0.5054	0.8184
0.7324	0.9863	0.6303	0.4882	0.8201
0.5475	0.8684	0.0995	0.4841	0.7000
0.4723	0.9945	0.7109	0.5334	0.6119

3.2 Entropy Method Analysis

The data have been standardized in subsection 3.1, and here it is only necessary to calculate the entropy, coefficient of variation and weight of each indicator in the model according to the formula in subsection 2.2. The calculation results are shown in Table 3.

Table 3: Calculation results of entropy weight method

	Entropy	Coefficient of variability	Weights
L	0.9724	0.0275	0.1851
R	0.9939	0.0060	0.0407
F	0.9359	0.0640	0.4313
M	0.9534	0.0465	0.3132
C	0.9956	0.0043	0.0294

This shows that the indicators with the greatest weight are F and M, followed by L, and the smallest are R and C. This indicates that the number of rides and mileage are the most influential indicators, the time of joining the membership is an indicator worthy of reference, and the time interval between recent purchases and the discount factor has the lowest reference value.

3.3 Cluster Analysis

The clustering analysis was performed using the WKmeans algorithm, the number of cluster centers was determined as 5 using the empirical method, the weights of the objective function (7) were the weights of the indicators obtained in subsection 3.2, and the distance was calculated using the Euclidean distance, and the final clustering results are shown in Table 4.

The table shows the clustering centers, number of clients, percentage and importance of the five categories. The cluster center of each category is the mean value of all sample coordinates in that category, so the center is representative of the category it belongs to. By analyzing the values of L, R, F, M and C indicators of each cluster center, we can know the value of the customers they represent, and the process of analysis is as follows.

Category 1 customers, accounting for 14.99%. The main characteristics of this category of customers are that they have the highest number of rides and mileage and have been enrolled for the longest period of time, making them high value customers. They contribute the most to the company, but account for a smaller percentage. According to the two-eight principle, the company should prioritize resources to the head customers, provide them with personalized services, and keep the loyalty and satisfaction of such customers.

Category 2 customers, accounting for 28.87%. The main characteristics of this category of customers are that they have been enrolled for a short period of time, have recently taken a flight for a short period of time, and have an intermediate number of trips and mileage, which are important development customers. These customers have greater potential for development, so we can target them with more promotional activities to develop customer

stickiness.

Category 3 customers, accounting for 19.95%. This category of customers has been enrolled for a longer period of time, and the number of rides and mileage are lower, so they are general customers and do not need special treatment.

Category 4 customers, accounting for 21.13%. This type of customer has a high number of flights and mileage, but the recent flight time is far away, which is an important retention customer. The possibility of losing these customers is high, so we can hold some old customer return activities to retain them.

Table 4: Clustering results

Category	Clustering Center					Number of customers	Percentage	Customer Value
	L	R	F	M	C			
1	0.8248	0.8653	0.0911	0.0510	0.4517	9300	14.99%	High Value Customers
2	0.0958	0.8643	0.0394	0.0255	0.4162	17913	28.87%	Important Development Customers
3	0.5969	0.7344	0.0304	0.0209	0.4324	12378	19.95%	General Customers
4	0.3326	0.8829	0.0697	0.0400	0.4343	13110	21.13%	Important Retention Customer
5	0.1666	0.3497	0.0091	0.0102	0.4237	9342	15.06%	Low Value Customers

Category 5 customers, accounting for 15.06%. The main characteristic of this category of customers is that they have lower indicators in all aspects, especially the number of rides and mileage is significantly less than other customers, and they are low-value customers. They provide the least revenue for the company, have a shorter membership time, are more mobile and account for less overall, and therefore do not require additional investment in human and material resources.

4. Conclusion

In the information age where massive data is abundant, data mining is widely used in different fields as a science and technology that can extract potential and effective information from unknown data. In this paper, a weight-based Kmeans algorithm is used to cluster the airline customers. The weights in the objective function are calculated by the objective entropy weight method, which fully considers the importance of different indicators to the customer value. The clustering algorithm classifies customers into 5 categories, with obvious distinction between different categories, and after analysis they can be classified as high value customers, important development customers, important retention customers, general customers and low value customers. The paper also gives personalized marketing strategies for the characteristics of each of these five categories of customers, which has certain reference value.

In other traditional fields involving big data, we can also take full advantage of data mining to discover the potential connections of data and uncover valuable information.

References

[1] Jiawei H, Micheline K. Data mining: concepts and techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2006, 5 (4): 1 - 18.
[2] Everitt B. Cluster analysis [J]. Quality & Quantity, 1980, 14 (1): 75-100.
[3] Ying Sun, Baolong Ma, Jinlin Li. Research on customer value identification of loyalty program members based on RFM model approach [J]. The Practice and Understanding of Mathematics, 2011, 41 (15): 75-79.
[4] Maolin Li. Research on hybrid storage data migration method under load balancing [D]. Xi'an University of Architecture and Technology, 2020.
[5] Zhanfeng Dong, Chunxu Hao, Qianqian Liu, Xiaodong Yan, Chazhong Ge. A study on provincial environmental performance index in China based on entropy weight method [J]. Environmental Pollution and Prevention, 2016, 38 (08): 93-99.
[6] Qiushuang Huang, Sen Zeng, Jiefeng Wang, Tiankai Wang. Entropy-COPARS-based sustainability evaluation of electric power companies Research [J]. Science and Technology Management Research, 2019, 39 (11): 101-106.
[7] Jain A K, Dubes R C. Algorithms for clustering data [M]. Prentice Hall, 1988.
[8] Juanying Xie, Shuai Jiang, Chunxia Wang, Yan Zhang, Weixin Xie. An improved global K-means clustering algorithm [J]. Journal of Shaanxi Normal University (Natural Science Edition), 2010, 38 (02): 18-22.
[9] Huang J Z, Ng M K, Rong H, et al. Automated variable weighting in k-means type clustering [J]. 2005, 27 (5): 657-668.

Author Profile

Saixin Wu received her bachelor's degree in software engineering from North China Electric Power University in 2019. She is currently a graduate student in Computer Science and Technology at the school. Her research interests include big data, machine learning, and clustering algorithms.