Analysis Method for Customer Value of Aviation Big Data Based on LRFMC Model

Yang Tao(✉)

Nanchang Institute of Science and Technology, Nanchang, China
taoyangxp@163.com

Abstract. In the era of Big Data, enterprise marketing focuses on customers instead of product center, and customer relationship management has become the core issue. In the aviation field, how to tap the high-quality customer base is more important. How to classify customers according to the characteristics of air passengers, and then make personalized marketing strategies for them, is the key problem to be solved. Aiming at optimizing resource allocation, with the help of aviation big data, a customer value analysis method is proposed based on the LRFMC model. First, Python was applied to clean, reduce and transform the data on the big data platform, and then to classify them. Moreover, characteristics of different customer categories were analyzed, and the customer value was evaluated. Finally, optimization methods based on K-Means algorithm were proposed, and the data were visualized, so that personalized services can be developed for different customers.

Keywords: LRFMC · Data analysis · Big data · Data visualization

1 Introduction

In modern society, China's civil aviation industry develops rapidly. With the continuous improvement of informatization, focusing on the civil aviation transportation industry, a large number of aviation customer data has been generated. With the continuous growth of airline network and air traffic volume, the competition among airlines is increasingly fierce, and the level of competition is also upgrading and expanding. In order to enhance competitiveness and attract aviation customers, airlines have put forward their own marketing plans, such as "Phoenix bosom friend" of Air China, "Oriental WanLiXing" of China Eastern Airlines, and "Pearl member" of China Southern Airlines. However, these alone cannot meet the needs of the explosive growth of the number of airline customers. The general membership level system has been difficult to assess the value and loyalty of airline passengers [1], nor to dig out the real value of massive customer data.

With the coming of the information age, the marketing focus of enterprises has changed from product center to customer center, and customer relationship management has become the core issue of enterprises. The key problem of customer relationship management is customer classification. Through customer classification, we can distinguish between valueless customers and high-value customers. Enterprises specify and optimize personalized services for different value customers, adopt

J. Zeng et al. (Eds.): ICPCSEE 2020, CCIS 1257, pp. 89–100, 2020.
https://doi.org/10.1007/978-981-15-7981-3_7

different marketing strategies, concentrate the priority marketing resources with high-value customers, and achieve the goal of maximizing profits. Accurate customer classification result is an important basis for the distribution of marketing resources. The more and more customers, the more and more become one of the key problems to be solved in the customer relationship management system [2–4].

In the face of accumulated market competition, various airlines have launched more preferential marketing methods to attract more customers. Domestic capable airlines are faced with business crisis such as passenger loss, competitiveness decline and insufficient utilization of aviation resources. Just like the big data application system[5–7] in all walks of life, it is necessary and effective to provide personalized customer service for different customer groups by establishing a reasonable customer value evaluation model, grouping customers, analyzing and comparing customer values with different customer groups, and formulating corresponding marketing strategies. At present, the airline has accumulated a large number of member file information and flight records. Therefore, airlines usually need to adopt some models to effectively classify customers. However, the traditional processing method has low performance, long time-consuming, poor accuracy and poor interaction experience effect.

In order to solve the above problems and more effectively classify customers, this paper proposes a method to realize customer value analysis, which is based on LRFMC customer value evaluation model [8] and machine learning K-Means [9] clustering analysis algorithm. Through the use of Python data analysis module pandas as the main body, the whole process of data exploration, data preprocessing, modeling analysis and visualization is fully interpreted. Through data preprocessing, standardized feature vector and K-Means clustering analysis algorithm, rapid and efficient clustering analysis [10] is carried out to achieve the purpose of obtaining accurate customer value information.

2 Basic Theory and Models to Identify Customer Value

Dataology and Data Science [11] are the science of data, which are defined as the theory, method and technology to study and explore the mysteries of data in cyberspace. There are two main connotations: one is to study the data itself; the other is to provide a new method for natural science and social science research, called the data method of scientific research. It is widely used in multidimensional matrix and vector computation.

Our goal is customer value identification, which is to identify customers with different values through airline customer data. The most widely used model to identify customer value is to evaluate and analyze through three indicators, as follows:

- Recent Consumption Interval (Recency)
- Consumption frequency (Frequency)
- Consumption (Monetary)

This is our common model called the RFM model. By analyzing the three indicators of individual consumers, the consumer group customers can be segmented to identify high-value customers. In the RFM model, the consumption amount represents the total amount of the customer's purchase of the company's products over a period of time. As the air fare is affected by various factors such as transportation distance and class of cabin, different passengers with the same consumption amount have different values to the airline. Therefore, this single indicator is not applicable to the specific application scenarios of airline customer value analysis.

Instead of the amount of consumption, we choose two indicators: the customer's accumulated mileage M within a certain period of time and the average value C of the discount coefficient corresponding to the passenger's cabin class within a certain period of time. In addition, considering the length of the membership time of airline members can affect the customer value to a certain extent, the length of customer u relationship L is added to the model as another indicator to distinguish customers. The five indicators of customer relationship length L, consumption interval R, consumption frequency F, flight mileage M, and discount coefficient C are used as the indicator of airline customer identification (Table 1). Based on the simulation, the model adjusted to meet the special needs of the local scene is called the LRFMC model.

Table 1. LRFMC model (airline customer value model)

Model item	Meaning
L	The number of months since the member's joining time from the end of the observation time
R	Number of months since the member's last flight from the end of observation time
F	The total number of times the member has flown during the observation period
M	Miles accumulated during member observation time
C	The average value of the discount factor used by the member during the observation period

3 Customer Value Analysis

For the LRFMC model of the airline, if the attribute binning method analyzed by the traditional RFM model is used, as shown in Fig. 1 (it is divided according to the average value of the attribute, where the value greater than the average is represented as ↑, and the value less than the average is represented as ↓), although the customers with the best value can also be identified, but there are too many subdivided customer groups, which increases the cost of targeted marketing.

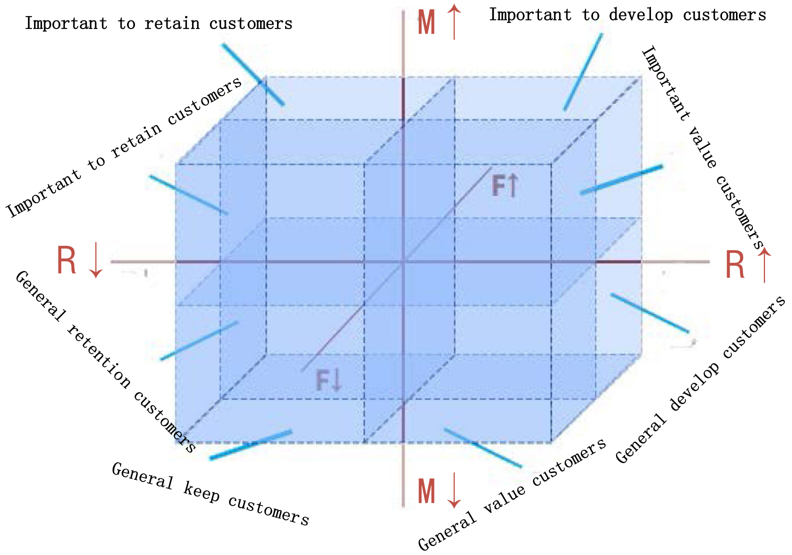
Important to retain customers
M ↑
Important to develop customers
Important to retain customers
Important value customers
R ↓
R ↑
F↑
F↓
General retention customers
General develop customers
General keep customers
M ↓
General value customers

Fig. 1. Attribute binning method.

Therefore, the clustering method was used to identify customer value. By performing K-Means clustering on five indicators of the LRFMC model of airline customer value, the most valuable customers were identified.

3.1 Data Analysis Overall Process

After we have a preliminary understanding of the LRFMC model, we need to make a complete plan for the overall data mining and data analysis. According to the standard process of data mining analysis, we divide the whole process into the following five parts, as shown in Fig. 2.

Part 1: sorting and viewing data sources in the business system;
Part 2: data extraction according to business analysis requirements;
Part 3: data exploration and preprocessing;
Part 4: data modeling and application;
Part 5: result feedback and visual presentation.

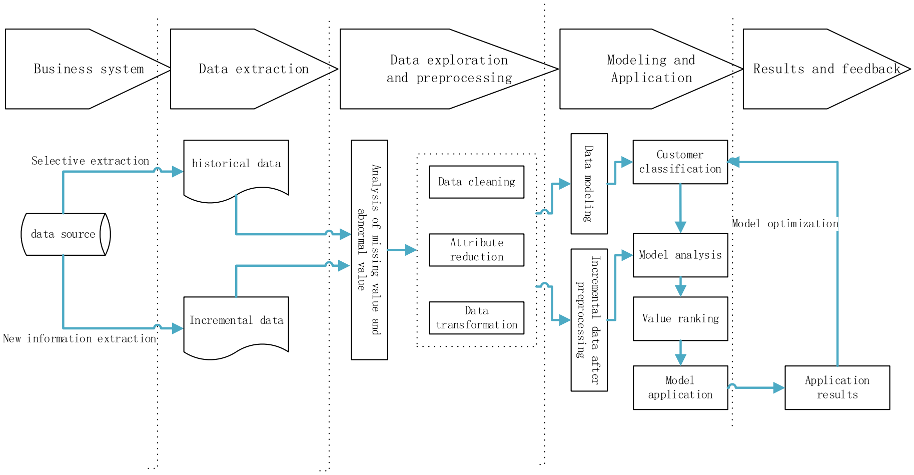

Fig. 2. Data analysis overall process.

Through the above five parts, we can complete the objective of data mining modeling and costumer value analysis of our project.

3.2 Data Mining Steps

The data mining of aviation customer value information mainly includes the following steps:

Step 1: selectively extract and add data from the data source of the airline to form historical data and incremental data respectively.
Step 2: carry out data exploration analysis and preprocessing for the two data sets formed in step 1, including data missing value and abnormal value exploration analysis, data attribute specification, cleaning and transformation.
Step 3: Based on the LRFMC model of customer value, the customer groups are divided and the characteristics of each customer group are analyzed.
Step 4: according to the customers who get different value from the model results, adopt different marketing methods to provide customized services.

After we have a preliminary understanding of the LRFMC model, we need to make a complete plan for the overall data mining and data analysis.

4 Application Case Analysis

We select a period of customer sample data of an airline as an example to analyze the application of the method. Take 2014-03-31 as the end time, select the two-year period as the analysis observation window, and extract the detailed data of all customers with boarding records in the observation window to form historical data. For the subsequent new customer details, the latest time point in the subsequent new data is taken as the end time, and the same method is used for extraction to form incremental data.

According to the last flight date, the detailed data of all passengers from April 1, 2012 to March 31, 2014 are extracted from the detailed data of basic customer information, flight information and integral information in the airline system, with a total of 62988 records. It includes 44 attributes, such as membership card number, joining time, gender, age, meeting card level, city of work, province of work, country of work, end time of observation window, integral points of observation window, kilometers of flight, times of flight, time of flight, interval between flights and average discount rate. Some of the aviation data are shown in Table 2.

Table 2. Aeronautical information data sheet (partial column slice data)

MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	FFP_TIER	WORK_COUNTRY
47229	2005/4/10	2005/4/10	6	CN
28474	2010/4/13	2010/4/13	6	US
58472	2010/2/14	2010/3/1	5	FR
13942	2010/10/14	2010/11/1	6	FR
45075	2007/2/1	2007/3/23	6	CN
47114	2005/1/15	2005/3/17	6	CN
54619	2006/1/7	2006/1/8	6	CN
12349	2008/6/16	2008/6/27	6	CN
35883	2006/4/11	2007/4/18	6	CN
56091	2004/11/25	2005/2/10	6	CN
2137	2005/4/11	2005/5/3	6	CN
27708	2006/3/20	2006/3/25	6	CN
28014	2006/12/1	2011/1/7	6	FR

In the data exploration and analysis, the missing value (null value) analysis and abnormal value analysis are carried out to analyze the data rules and abnormal values. Through the observation of the data, it is found that there are records in the original data that the ticket price is null, the minimum ticket price is zero, the minimum discount rate is zero, and the total flight kilometers are greater than zero. The data with null fare may be caused by the fact that the customer does not have a flight record, and other data may be generated by the customer taking a 0% discount ticket or exchanging points. Through reading and writing Python CSV file, data description and matrix transposition in Pandas module library, the number of empty values, maximum value and minimum value of each column of attribute observation value can be found. The results of data exploration and analysis are shown in Table 3.

Table 3. Data exploration and analysis data sheet (partial column slice data)

Attribute name	Null value number	Max	Min
MEMBER_NO	0	62988	1
FFP_DATE	0		
FIRST_FLIGHT_DATE	0		
GENDER	3		
FFP_TIER	0	6	4
WORK_CITY	2269		
WORK_PROVINCE	3248		
WORK_COUNTRY	26		
AGE	420	110	6
LOAD_TIME	0		
FLIGHT_COUNT	0	213	2
BP_SUM	0	505308	0
EP_SUM_YR_1	0	0	0

4.1 Data Cleaning

Through data exploration and analysis, it is found that there are missing values in the data, the minimum value of ticket price is zero, the minimum value of discount rate is zero, and the total flight kilometers are greater than zero. Due to the large amount of original data, the proportion of such data is small, which has little impact on the problem, so it is discarded. The specific rules and standards are as follows:

Discard the record with empty ticket price.
Discard the record that the fare is zero, the average discount rate is not zero, and the total flying kilometers are greater than zero.

The steps of data cleaning are as follows:

Step1: read the original sample CSV data.
Step2: set data filtering conditions according to standard requirements, and realize data clearing.
Step3: write the final result to the data file.

4.2 Attribute Reduction

There are too many attributes in the original data. According to the LRFMC model of airline customer value, six attributes related to the LRFMC model indexes are selected: FFP_DATE, LOAD_TIM, FLIGHT_COUNT, AVG_DISCOUNT, SEG_KM_SUM, and LAST_TO_END.

Delete attributes that are not related, weakly related or redundant, such as membership card number, gender, city of work, province of work, country of work and age. Through the data slicing function of pandas module, attribute reduction is realized. The data set after attribute selection is shown in Table 4.

Table 4. Data set after attribute selection (partial column slice data)

FFP_DATE	LOAD_TIME	FLIGHT_COUNT	AVG_DISCOUNT	SEG_KM_SUM
2006/11/2	2014/3/31	210	0.961639043	580717
2007/2/19	2014/3/31	140	1.25231444	293678
2007/2/1	2014/3/31	135	1.254675516	283712
2008/8/22	2014/3/31	23	1.090869565	281336
2009/4/10	2014/3/31	152	0.970657895	309928
2008/2/10	2014/3/31	92	0.967692483	294585
2006/3/22	2014/3/31	101	0.965346535	287042
2010/4/9	2014/3/31	73	0.962070222	287230
2011/6/7	2014/3/31	56	0.828478237	321489
2010/7/5	2014/3/31	64	0.708010153	375074
2010/11/18	2014/3/31	43	0.988658044	262013
2004/11/13	2014/3/31	145	0.95253487	271438
2006/11/23	2014/3/31	29	0.799126984	321529

4.3 Data Transformation

Data transformation is to transform data into "appropriate" format to meet the needs of mining tasks and algorithms. In this project, the main data transformation method is attribute construction. Because the original data does not directly give the five indicators of LRFMC, the five indicators need to be extracted from the original data. The specific calculation method is as follows:

- L = LOAD_TIME - FFP_DATE

The number of months between the time of membership and the end of observation window = the end time of observation window - the time of membership [unit: month].

- R = LAST_TO_END

The number of months from the last time the customer took the company's aircraft to the end of the observation window = the time from the last flight to the end of the observation window [unit: month].

- F = FLIGHT_COUNT

Number of times the customer takes the company's aircraft in the observation window = number of flights in the observation window [unit: Times].

- M = SEG_KM_SUM

Accumulated flight history of the customer in observation time = total flight kilometers of observation window [unit: km].

- C = AVG_DISCOUNT

Average value of the discount coefficient corresponding to the passenger space during the observation time = average discount rate [unit: none].

4.4 Data Normalization

At the same time, the standardized data is more conducive to the accuracy of model analysis. We use Z-score method to normalize the LRFMC index data. Z-score normalization is also known as standard deviation normalization. The normalized data is normally distributed, i.e. the mean value is zero, and the standard deviation is a formula as follows:

$$\chi^{n} = \frac{\chi - \mu}{\sigma} \tag{1}$$

Where μ is the mean value of all sample data and σ is the standard deviation of all sample data. Implemented in Python, the code is as follows:

$$\text{data} = (\text{data} - \text{data.mean}\ (\text{axis} = 0))/\text{data.std}\ (\text{axis} = 0)$$

The difference between the standard deviation and the standard deviation is that the standard deviation only reduces the variance and mean deviation of the original data by multiple, while the standard deviation standard deviation makes the standardized data variance one. This is more advantageous to many algorithms, but its disadvantage is that if the original data is not Gaussian distribution, the standardized data distribution effect is not good.

4.5 Modeling Analysis

The construction of customer value analysis model mainly consists of two parts. The first part: according to the data of five indicators of airline customers, cluster and group customers. The second part: combined with the business to analyze the characteristics of each customer group, analyze its customer value, and rank each customer group. K-Means clustering algorithm is used to classify customers into 5 categories (the number of customer categories needs to be determined by combining the understanding and analysis of business). In the specific implementation, the K-Means clustering algorithm we used is located in the clustering word library (sklearn.cluster) under the Scikit-Learn module library. We use SK-Learn module to create K-Means model object instance, use fit() function of K-Means object instance to complete model analysis, and finally get the model analysis result, as shown in Table 5.

Table 5. Cluster analysis results

Cluster No	ZL	ZR	ZF	ZM
4183	0.052674226	−0.002018209	−0.22675176	−0.231377928
24649	−0.700316003	−0.415473188	−0.160700141	−0.160404636
15746	1.160429765	−0.377379525	−0.087102357	−0.095201216
12139	−0.314170608	1.684986848	−0.573955288	−0.536720331
5334	0.483815811	−0.799389015	2.483818101	2.425261912

4.6 Data Visualization

According to the analysis results of the previous model, we want to visualize the distribution of customer clustering population by using histogram. First of all, we use the Pandas module to intercept the relevant data and generate the parameter data structure and type required for histogram. Then we build the basic structure of histogram based on python. Finally, we use Matplotlib module to generate histogram through data binding. Through the above steps, the distribution of customer clustering population is shown in Fig. 3.

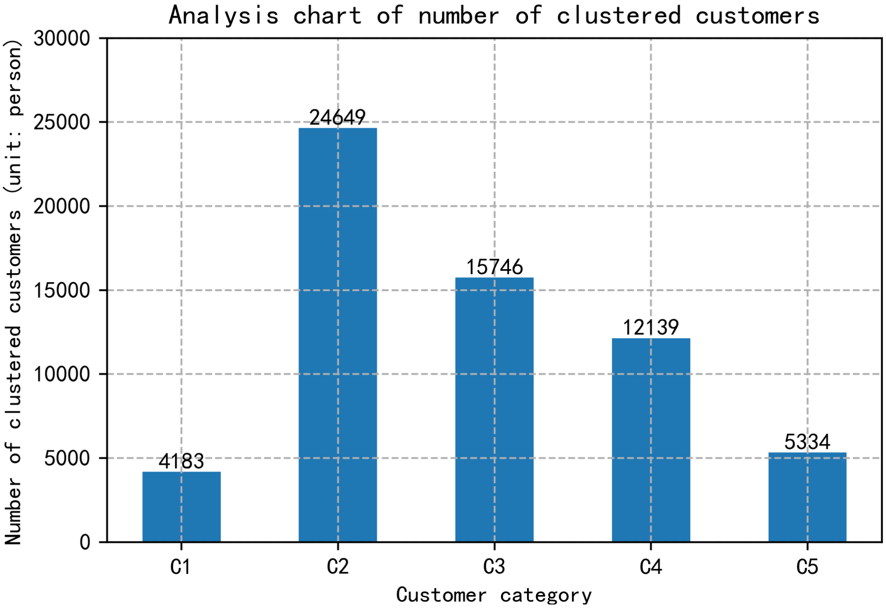

Fig. 3. Distribution of customer clustering population is shown.

Based on the above visualization of clustering population distribution, radar map can be constructed to visualize customer clustering eigenvalues, and the results are shown in Fig. 4.

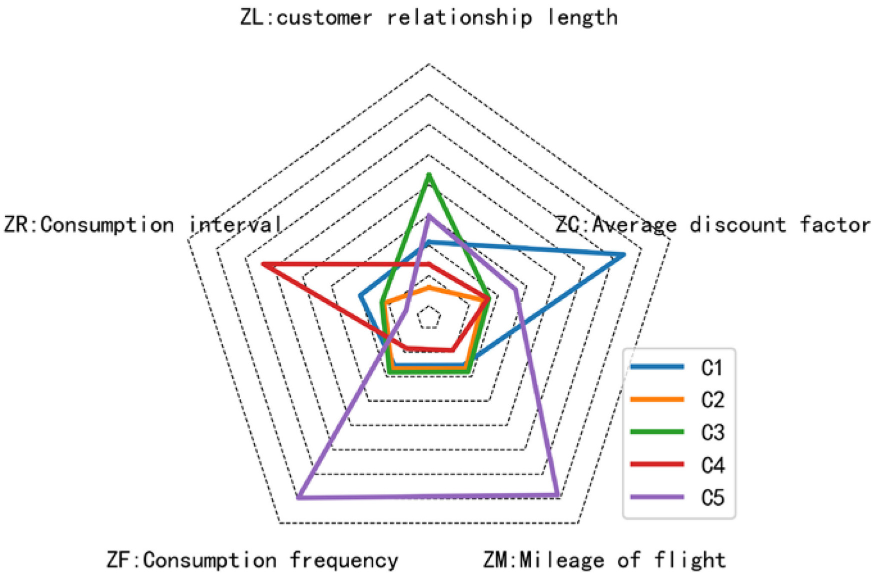

Fig. 4. Visualization of customer clustering eigenvalues is shown.

Based on the data and visual chart, the clustering results are analyzed. Through the chart observation, we can see that:

- customer group 1 is the largest in C attribute and relatively average in L, R, F and M;
- customer group 2 is relatively average in the five dimensions of LRFMC;
- customer group 3 is the largest in L attribute and relatively average in R, F, M and C, with no significant difference;
- customer group 4 is the largest in R attribute and relatively average in L, F, M and C;
- customer group 5 is the largest in F and M attributes, and the average in L, R and C attributes.

Combined with the business analysis, the paper evaluates and analyzes the characteristics of a group by comparing the size of each index among groups. There are maximum, minimum, sub maximum and sub minimum values in each indicator attribute, so we need to analyze and count them. For example, customer group 5 has the largest attribute of F and m, and the smallest attribute of R, so it can be said that f and m are the advantages of customer group 5. By analogy, F and m are inferior characteristics in customer group 4.

5 Conclusion

Starting from customer classification, based on the customer value analysis model of LRFMC, this paper analyzes customer resources and customer value mining to maximize customer value, and proposes a general standard process for big data analysis. Based on Python, according to the algorithm model of scientific computing and data analysis and machine learning, this paper implemented the application analysis based on aviation big data, in order to achieve the best customer classification results, and verified the effectiveness of the proposed method.

Acknowledgment. The project is provided with technical training guidance by China Soft International Education Technology Group. This paper is supported by the Ph.D. Research Initiation Fund of Nanchang Institute of Science and Technology with the Project (No. NGRCZX-18-01).

References

1. Ban, H.J., Joung, H.-W.: The text mining approach to understand seat comfort experience of airline passengers through online review. Culinary Sci. Hospitality Res. **25**(9), 38–46 (2019)
2. Kun, Q., Li, C.: Construction and research of airline customer relationship management system based on Python. J. Baotou Vocat. Tech. Coll. **20**(04), 9–14 (2019)
3. Xie, W.: Diagnostic Research on Customer Relationship Management System of China Eastern Airlines. China University of Geosciences, Beijing (2015)
4. Chiang, W.-Y.: Establishing high value markets for data-driven customer relationship management systems. Kybernetes **48**(3), 650–652 (2019)
5. Zhiliang, M., Mingkun, T., Yuan, R.: Building energy consumption information model for big data analysis. J. S. China Univ. Technol. (Natural Science Edition) **47**(12), 72–77 (2019)
6. Luo, X., Wen, X.: Student behavior analysis model based on artificial intelligence and big data. Educ. Obs. **8**(17), 11–13 (2019)
7. Sun, D., Wu, J., Wen, H., Xue, M.: Study on big data analysis model and application of mountain slope disaster resilience–taking Chengkou County as an example. J. Chongqing Norm. Univ. (Nat. Sci. Ed.) **36**(03), 64–71 (2019)
8. Tahanisaz, S., Sajjad, S.: Evaluation of passenger satisfaction with service quality: a consecutive method applied to the airline industry. J. Air Transp. Manage. **83**, 101764 (2020)
9. Li, J., Cao, Y., Wang, Z., Wang, G.: An algorithm for cloud simplification based on K-means clustering and Hausdorff distance. J. Wuhan Univ. (Inf. Sci. Ed.) **45**(02), 250–257 (2020)
10. Song, Y., Peng, G., Sun, D., Xie, X.: Active contours driven by Gaussian function and adaptive-scale local correntropy-based K-means clustering for fast image segmentation. Sig. Proc. **174**, 107625 (2020)
11. Data Science and Data Science. Tianjin Econ. (07), 63–64 (2017)