BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Management Science & Engineering Research
https://ojs.bilpublishing.com/index.php/jmser

ARTICLE

Clustering Analysis of User Loyalty Based on K-means

Qiushui Fang* **Zhiming Li** **Mengtian Leng** **Jincheng Wu** **Zhen Wang**

Guangdong lingnan Pass co., LTD., Guangzhou, 510000, China

ARTICLE INFO

Article history
Received: 15 May 2020
Accepted: 26 May 2020
Published Online: 30 June 2020

Keywords:
Machine learning
Public transportation
K_means
KDE

ABSTRACT

In recent years, the rise of machine learning has made it possible to further explore large data in various fields. In order to explore the attributes of loyalty of public transport travelers and divide these people into different clustering clusters, this paper uses K-means clustering algorithm (K-means) to cluster the holding time, recharge amount and swiping frequency of bus travelers. Then we use Kernel Density Estimation Algorithms (KDE) to analyze the density distribution of the data of holding time, recharge amount and swipe frequency, and display the results of the two algorithms in the way of data visualization. Finally, according to the results of data visualization, the loyalty of users is classified, which provides theoretical and data support for public transport companies to determine the development potential of users.

1. Introduction

With the continuous improvement of domestic economic level, the process of China's urbanization is speeding up, the city scale is expanding, and the personal vehicle ownership and road traffic flow are also increasing rapidly, which leads to the increasingly prominent contradiction between traffic supply and traffic demand. In order to solve this problem, public transport has been greatly developed and become one of the important bases to maintain the operation of the city.

With the development of public transport, public transport companies such as public transport companies, subway companies and all in one card companies have a large number of user transaction data. In order to improve the competitiveness and viability of the company, expand the revenue channels of the company, tap the potential value of data, and realize the dimension expansion of data has become an urgent means to be considered.

In terms of potential value mining of public transport data, there are many researches. Tao Zhou et al [1] put forward the data analysis and processing system framework of public transportation IC card based on GIS, and established the system analysis process to realize the calculation of user travel OD; Xiangyun Li et al. [2] used Gaussian process regression to realize the prediction of bus arrival time; Li Goldman Sachs et al. [3] used the neural network model of LSTM to learn the long-term sequence data of boarding and alighting passenger flow at multiple stations, and realized the short-term prediction of public transport passenger flow;In addition, there are a lot of researches such as station matching [4], bus route optimization [5], travel feature analysis [6], and bus operation optimization [7], which greatly expand the dimension of public transport data and tap the potential value of public transport data. However, there is no research on the loyalty attribute of public transport users.

**Corresponding Author:*
Qiushui Fang,
Guangdong lingnan Pass co., LTD., Guangzhou, 510000, China;
Email: xinsuile1991@163.com

DOI: https://doi.org/10.30564/jmser.v2i2.1851

Public transport user loyalty refers to the degree of public transport users' support and use of public transport. Research on customer loyalty can understand the current situation of customers in enterprises as a whole and evaluate the level of customer loyalty. It can not only make enterprises clearly understand the current customer quality, but also clarify the position and weight of enterprises in the market [8].For public transport, users who use public transport travel, namely customers, study the loyalty attribute of users, and classify the loyalty level of users, which can make the public transport related enterprises clear the value and potential of various users.User loyalty data can serve for product design, marketing activities and other businesses, provide theoretical and data support for public transport enterprises to determine the development potential and development value of users, and provide reference for public transport enterprises to select target customers in industry expansion and business application.

In order to analyze the loyalty attribute of public transport users, this paper uses traffic IC data (including public transport data and subway data) and K-means algorithm to cluster the user's card duration, card swiping frequency and recharge amount, and then uses kernel density estimation algorithm to analyze the clustering results and realize data visualization. Finally, the traffic IC card users are divided into four loyalty levels, which are public The joint transportation enterprises provide data support for the determination of users' potential and value.

2. K-means Clustering Algorithm and Kernel Density Estimation

In this paper, K-means algorithm is used to cluster user loyalty, and then kernel density estimation algorithm is used to further analyze the clustering results.

K-means clustering algorithm is a classical distance based clustering algorithm. The so-called clustering, according to the principle of similarity, divides the data objects with higher similarity into the same cluster, and divides the data objects with higher dissimilarity into different clusters, so as to achieve the division of data categories. K-means algorithm first randomly selects k points (called clustering centers) in the data set, and then associates each data point with the nearest center point according to the distance from k center points. All points associated with the same center point are classified into one group, and then calculates the mean value of each group of data, taking the mean value as a new clustering center and following it repeatedly Finally, when the cluster center does not change, the cluster is completed.

K-means clustering algorithm is a classical distance based clustering algorithm. The so-called clustering, according to the principle of similarity, divides the data objects with higher similarity into the same cluster, and divides the data objects with higher dissimilarity into different clusters, so as to achieve the division of data categories. K-means algorithm first randomly selects k points (called clustering centers) in the data set, and then associates each data point with the nearest center point according to the distance from k center points. All points associated with the same center point are classified into one group, and then calculates the mean value of each group of data, taking the mean value as a new clustering center and following it repeatedly Finally, when the cluster center does not change, the cluster is completed.

Kernel density estimation is used to estimate the unknown density function in probability theory, which belongs to one of the nonparametric test methods.Generally, there is a big gap between the basic assumption of the parameter model and the actual physical model. These methods do not always achieve satisfactory results.The kernel density estimation makes full use of the information of the data itself to avoid the prior knowledge brought in by human subjective, so it can approximate the sample data to the greatest extent (relative to the parameter estimation). Because kernel density estimation method does not use the prior knowledge of data distribution and does not attach any assumptions to data distribution, it is a method to study the characteristics of data distribution from the data sample itself. Therefore, it is highly valued in the field of statistics theory and application.

3. Data Preprocessing

This paper analyzes the data of Guangzhou IC card in May 2019 (including the data of public transportation and subway), the data of related users' card issuing and the data of users' recharge in that month, and obtains the data of users' card duration, frequency of card swiping in that month and recharge amount in that month.

Due to the different types of people represented by each card type (ordinary card, senior card, student card, etc.) and the preferential policies for riding, it needs to be analyzed separately. This paper takes the data of ordinary card as an example.

Before K-means clustering, we need to preprocess the data.There may be incomplete (missing values), inconsistent and abnormal data in the massive original data, which will seriously affect the clustering effect of K-means algorithm, and even cause the clustering result is unreasonable, so data cleaning is particularly important. The transaction data obtained from various channels can not be directly used for data analysis, and the missing and in-

 DOI: https://doi.org/10.30564/jmser.v2i2.1851

complete data can be processed.Because there are enough samples in the data set, the number of data in the above cases is generally relatively small, which can be discarded.Statistics and observation of the data are carried out.

The average value, maximum value and minimum value of the data are shown in Table 1:

Table 1. data description

	Card duration (month)	recharge amount in the month (yuan)	swipe frequency in the month (Times)
Average	47.93	49.18	23.70
Max	204	9750	1040
Min	0	0	1

From the above table, it can be found that the maximum and average values of the recharge amount and the frequency of card swiping are extremely unbalanced. Through the data statistics, it is found that the capacity of sample data set is more than 8 million pieces, and the data with more than 130 times of swiping card accounts for 0.5%, only 5 pieces of data with more than 500 times; 0.2% of the data with a recharge amount of more than 300 yuan, and only 1254 pieces of data with a recharge amount of more than 500 yuan, only 1 / 10000 of the data set. These oversized data only account for a small part and belong to outliers. Although the value of outliers is beyond the normal range, some of the data are real, so it can not be simply discarded. At the same time, it can be seen from the above table that the numerical distribution range of the selected features in this analysis is quite different. In order to reduce the incongruity of numerical values between features, it is necessary to normalize or standardize the data first.

The Z-score standardized formula model is as follows:

$$x^* = \frac{x - x_{avg}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x)^2}}$$

Where x is a piece of data of the sample, which represents the average value of the sample, n is the sample capacity, which represents the piece of data of the sample, which is the result after Z-score standardization.

When the difference between X and the average value of the sample is larger, the standard deviation represented by the denominator in the formula is larger, which greatly reduces the impact of extreme maximum value on clustering results.

4. User Loyalty Clustering Based on k-means

4.1 Selection of Cluster Number

Because K-means clustering algorithm can not determine the number of clustering categories K, so we need to choose the appropriate K value.When the number of clusters can not be defined according to the actual situation of clustering objects, the commonly used methods are: empirical method, elbow method, contour coefficient method and interval statistics method.For user loyalty clustering, the number of clustering categories can be self-defined, that is, empirical method. In this paper, the user loyalty clustering categories are set to four categories, which correspond to high loyalty group, lower loyalty group, general loyalty group and low loyalty group respectively.

4.2 User Loyalty Clustering

In this paper, K-means algorithm is used to cluster the three characteristics of traffic IC card users, i.e. card duration, recharge amount and card frequency.After clustering, the output results will divide the users into four categories of loyalty groups. Although k-means algorithm divides the groups, the corresponding loyalty levels of the four groups cannot be directly confirmed. It is necessary to visually analyze the three characteristics of the analysis group, i.e. cardholder duration, recharge amount and swiping frequency, in order to classify the loyalty levels of each category.

4.3 Visual Analysis of Clustering Results

Because the clustering results are tens of thousands of data, it can't be directly displayed by simple drawing such as scatter diagram, so further analysis of the characteristics of each clustering result is needed.In this paper, KDE (kernel density estimation) algorithm is used to analyze the clustering results visually. KDE algorithm makes full use of the information of data itself to avoid the prior knowledge brought in by human subjective, so as to achieve the maximum approximation of sample data. KDE algorithm can objectively and directly show the probability distribution of data in the numerical range, which is suitable for this paper to estimate the distribution of three characteristics of all kinds of people.

The formula model of nuclear density estimation is as follows:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{h}\sum_{i=1}^{n} K_h\left(\frac{x - x_i}{h}\right)$$

H is the bandwidth of kernel density estimation, which needs to be set according to the actual situation. N is the sample size, which is the kernel function of kernel density estimation. It is similar to the kernel function in support vector machine (SVM), mean IFT and other algorithms.

DOI: https://doi.org/10.30564/jmser.v2i2.1851

The common kernel functions in kernel density estimation include uniform function, triangle function, Biweight function, triweight function, epanechnikovnormal function, etc.In this paper, the kernel function used for kernel density estimation of three features is Gaussian function, and its formula model is as follows:

$$K_h(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$$

In the kernel density estimation algorithm model, different bandwidth will lead to great difference in the final fitting results. If h is too small, there will be too few points involved in fitting in the field, and if h is too large, waveform fusion may occur. The selection of H depends on the specific situation. If the fitted probability distribution curve is considered to be too flat, the h parameter can be reduced appropriately. If the fitted probability distribution curve is considered to be too steep, the bandwidth h can be increased appropriately. In this paper, the bandwidth selected for the core density estimation of the card duration is 1.0, the bandwidth selected for the recharge amount is 5.0, and the bandwidth selected for the card frequency is 1.0.

The curve chart of card duration, recharge amount and card frequency of various other groups is as follows:

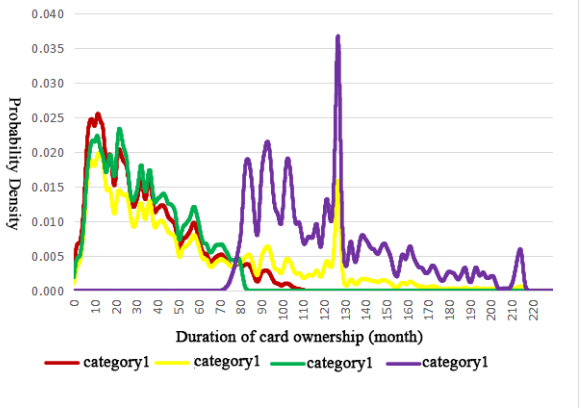

Figure 1. KDE curve of card ownership time

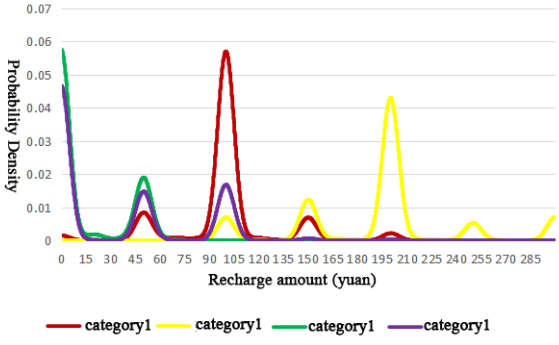

Figure 2. KDE curve of recharge amount

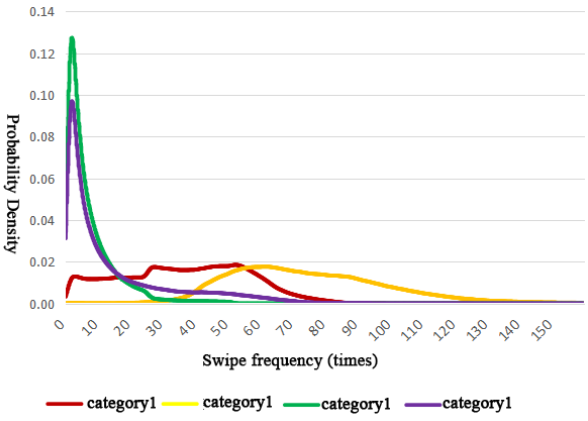

Figure 3. KDE curve of card swiping frequency

From the above three figures, it can be seen that the length of the cardholder of category 1 is relatively low, generally less than 100 months, and the recharge amount is relatively moderate, mainly concentrated in 100 yuan, and the frequency of card swiping is relatively moderate, generally less than 80 times, so the group of category 1 belongs to the higher loyalty group;The duration of the card of category 2 group is moderate, generally less than 170 months, which is mainly concentrated in the area with low value. The recharge amount is relatively high, mainly 200 yuan, and the frequency of card swiping is relatively high, mainly between 40 and 125 times. Therefore, category 2 group belongs to the group with high loyalty;Category 3 people have relatively low card duration, generally less than 80 months, relatively low recharge amount, generally no more than 60 yuan, relatively low card frequency, generally less than 30 times, so Category 3 people belong to low loyalty group;The length of cardholder in Category 4 is generally within 70-220 months, mainly between 70-130 months, relatively high, and the amount of recharge is relatively low, generally lower than 120 yuan, mainly concentrated at 0 yuan, that is, this group generally does not recharge or recharge less, and the frequency of card swiping is relatively low, generally lower than 70 times, and mainly concentrated between 0-20 times and more inclined to 0 times, so category 4 group Belong to the general loyalty group.

5. Increase Loyalty

In the field of public transport, improving user loyalty needs to consider improving user experience. For example, if the cost allows, the distance between stations can be shortened by adding stations to make it more convenient for users to ride; more shifts can be added to reduce the waiting time of users and reduce the congestion in the car; WiFi or shared tissue can be provided in the car to

 DOI: https://doi.org/10.30564/jmser.v2i2.1851

facilitate users, etc. There are many ways to improve the loyalty of users. Public transport enterprises can improve the loyalty of users with the permission of cost, and have more and better customer resources for their industry expansion and business application.

6. Summary

In the era of fierce competition, in order to improve the company's competitiveness and viability, expand the company's revenue channels, tap the potential value of data, and realize the dimension expansion of data has become an urgent means to be considered. In this paper, traffic IC data (including bus data and subway data) is used to correlate user card issuing data and user recharge data of the current month, so as to obtain the user's card duration, card swiping frequency and recharge amount data of the current month. After data cleaning and Z-score standardization, according to different card types, K-means algorithm is used to cluster the three user characteristics of user's card duration, card swiping frequency and recharge amount, and then kernel density estimation algorithm is used to analyze the clustering results, and the analysis results of kernel density estimation are displayed by data visualization. After the core density map is Through observation, traffic IC card users are finally divided into four loyalty levels, which provide data support for public transport enterprises to determine the potential and value of users, and also provide reference for public transport enterprises to select target customers in industry development and business application.

Reference

[1] Tao Zhou, Changxu Zhai, Zhigang Gao. Research on OD calculation technology based on bus IC card data[J]. Urban transportation, 2007 (03): 48-52.

[2] Xiangyun Li, Shuai Ren, Weigang Zhang, Juanjuan Wu, Jing Wu. Prediction method of bus arrival based on Gaussian process regression[J/OL]. Computer technology and development, 2019 (09): 1-7.

[3] Li Goldman Sachs, Ling Peng, Xiang Li, Tong Wu. Study on short-term passenger flow prediction of urban bus stations based on LSTM[J]. Highway transportation technology, 2019, 36(02): 128-135.

[4] Siyuan Zhou, Jiayu Liu, Jiayi Chen, Yue Ren, Wanfeng Dou. Station matching method based on bus IC card passenger flow data[J]. Electronic technology and software engineering, 2017(12): 173-174.

[5] Zeda Xu, Minfeng Yao. Optimization method of conventional public transport under the common line relationship between rail transit and conventional public transport[J]. Journal of Huaqiao University (Natural Science Edition), 2018, 39(04): 562-568.

[6] Tieyan Zhang. Population division and travel characteristics analysis based on public transportation IC card data - Taking Qingdao as an example[A] Academic Committee of urban transport planning of China Urban Planning Society. Innovation driven and intelligent development: Proceedings of 2018 China Annual Conference of urban transport planning[C]. Academic Committee of urban transport planning of China Urban Planning Society: Urban Transport Research Institute of China urban planning and Design Institute, 2018: 9.

[7] Zhengwu Wang, Anqi Liu, Kangkang Tan. Optimization of DRC bus operation cycle under uneven passenger distribution[J]. Transportation science and engineering, 2016, 32(02): 85-88.

[8] Hui Li. Analysis and Research on potential customer development and customer loyalty[D]. Yanshan University, 2016.

DOI: https://doi.org/10.30564/jmser.v2i2.1851