Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 4680959, 11 pages
https://doi.org/10.1155/2020/4680959

Research Article

Capturing the Characteristics of Car-Sharing Users: Data-Driven Analysis and Prediction Based on Classification

Jun Bi,[1,2] **Ru Zhi**,[1] **Dong-Fan Xie**,[1] **Xiao-Mei Zhao**,[1] **and Jun Zhang**[3]

[1]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China*
[2]*Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing, China*
[3]*Yunan Travelsky Airport Technology Co. Ltd., Yunnan, China*

Correspondence should be addressed to Dong-Fan Xie; dfxie@bjtu.edu.cn

Received 24 August 2019; Accepted 16 January 2020; Published 9 March 2020

Academic Editor: Zhi-Chun Li

This work explores the characteristics of the usage behaviour of station-based car-sharing users based on the actual operation data from a car-sharing company in Gansu, China. We analyse the characteristics of the users' demands, such as usage frequency and order quantity, for a day with 24 1 h time intervals. Results show that most car-sharing users are young and middle-aged men with a low reuse rate. The distribution of users' usage during weekdays shows noticeable morning and evening peaks. We define two attributes, namely, the latent ratio and persistence ratio, as classification indicators to understand the user diversity and heterogeneity thoroughly. We apply the k-means clustering algorithm to group the users into four categories, namely, lost, early loyal, late loyal, and motivated users. The usage characteristics of lost users, including maximum rental time and travel distance, minimum percentage of same pickup and return station, and low percentage of locals, have noticeable differences from those of the other users. Late loyal users have lower rental time and travel distance than those of the other users. This manifestation is in line with the short-term lease of shared cars to complete short- and medium-distance travel design concepts. We also propose a model that predicts the driver cluster based on the decision tree. Numerical tests indicate that the accuracy is 91.61% when the user category is predicted four months in advance using the observation-to-judgment period ratio of 3 : 1. The results in this study can support enterprises in user management.

1. Introduction

The increasing number of vehicles has caused congestion and deterioration of the environment in many large cities worldwide. A new and convenient energy-saving transportation mode called car-sharing with full-electric vehicular fleets has emerged with the development of mobile internet technology to alleviate these traffic issues [1]. Car-sharing is an autonomous car rental mode in which users can use cars for a short period of time by borrowing them from their operators instead of acquiring ownership [2]. Car-sharing users only need to use smartphones to complete a series of self-registration, unlocking, payment, and other car rental programmes on a fixed site. Accordingly, car-sharing can meet users' temporary personal demands satisfactorily.

The car-sharing mode has remarkable social benefits. For instance, it can improve the utilisation rate of vehicles and decrease the number of vacant resources effectively; moreover, this mode is conducive to the reduction of car ownership to a certain extent [3]. Martin and Shaheen [4] found that each shared car could replace six to seven private cars on average. Nijland and Meerkerk [5] found that car ownership among car-sharing users had decreased by more than 30%. The car-sharing travel mode can reduce users' demand for parking; thus, it can alleviate the shortage of public resources, cut down user travel cost, and reduce unnecessary vehicle travel distance [6]. Wang et al. [3] found that more than 70% (50% on average) of car-sharing users cut their journeys with a shorter travel distance. Nijland and Meerkerk [5] found that the vehicle travel distance of private

car owners was decreased by 15%–20% after the introduction of the car-sharing system. Loose [7] showed that the car-sharing service could reduce 28%–45% of vehicle mileage. The energy cost and emissions will be greatly reduced because of the common launch of clean-energy vehicles by car-sharing operators. Loose [7] reported that each car-sharing user reduces carbon dioxide emissions by 39%–54% on average. Nijland and Meerkerk [5] found that user's carbon dioxide emission decreased by 13%–18%.

The site-based car-sharing model is a popular model in many areas and is convenient for operation and management; many studies on car-sharing are also based on the site-based model, where the user's age is mostly distributed in the range of 20 to 30 years [8]. Some scholars found that incorporating electric and nonelectric vehicles into a fleet increases users' interest and participation because most users prefer hybrid electric vehicles under the same conditions [9]. A study found that car-sharing users prefer environmentally friendly vehicles [10] when selecting their own car. This notion indicates that developing car-sharing promotes the purchase of new energy vehicles to a large extent. Long-distance travellers do not opt for electric vehicles or plug-in hybrid ones because of the mileage limitation of new energy vehicles [11], and 83% of travellers use car-sharing for short-distance travel [9].

Some scholars have analysed the factors that affect the demand and willingness of car-sharing on the basis of the user model; numerous studies have been carried out to develop various logit models, for example, the multiple logistic regression model [12], ordered logit model [13], and binomial logit model with sequence correlation (Greeks. [14]). These studies, which focused on the impact of demographic characteristics and travel attributes on the willingness to use shared car, reported that environmentally conscious low- and middle-income people are willing to use car-sharing. The usage characteristics and demand of car-sharing users vary across different regions [10]. Car-sharing demand is concentrated in specific time periods and regions [8]. Accordingly, parking demand variations are related to geographical areas and parking types [15]. Regional attributes greatly affect the user's demand. Urban users have lower car ownership, and suburban ones have more trips than nonusers.

Describing users' usage behavioural patterns is a key issue in the car-sharing field; scholars can divide all users into active and sporadic ones based on their usage activity [16]. Users can also be divided into five clusters by combining user attributes [17], wherein most long-term users go on temporary short- or middle-distance travel. The usage of sporadic users is more likely to occur on weekends with high travel and time lengths. Hui et al. [18] realised the cluster of the travel chain by using the travel distance, home-based travel chain, and parking time in a certain place. The authors found that users have varied travel purposes under different travel chain modes. Predicting the dependent variables effectively and analyzing the influence of various attribute factors also have important practical significance. Habib et al. [19] proposed a user behaviour econometric model. Such model can predict the duration of users' continuous usage, determine the month in which users become active members, and estimate the quantity of active members' usages each month. Some researchers have analysed users' vehicle selection behaviour and influence of various attribute factors and proposed vehicle selection models, such as multiple discrete continuous extremum model [20], accelerated failure model, and space hazard-based model [21].

In the free-floating car-sharing mode, the distance between the user and the vehicle influences the possibility of selecting a vehicle [22]. The reasonable layout of the charging station is also the research focus. Schussler and Bogenberger [23] investigated the charging behaviour of different user groups and provided a strategy to determine the locations of public charging stations. Space attributes are emphasised in the cluster analysis of usage patterns because the spatial location is an important issue for users to consider the use of shared cars [24]. Although most car-sharing organisations have used hybrid and low-emission vehicles, several users are unaffected by the mileage limitations of battery electricity vehicles (BEVs) because they can meet most of the travel demands of users (when 80% of the travel distance is less than 20 km) [22]. Accurate prediction of order quantity can provide a practical significance for operation. Müller et al. [25] developed a negative binomial statistical model to predict the reservation quantity. The related influencing factors must be determined to predict the car-sharing demand accurately. Wang et al. [26] recently discovered that three factors, namely, selection behaviour of the car-sharing mode, maximum acceptable price for car-sharing, and willingness to give up buying a car, were influential on Chinese individual user acceptance.

The car-sharing mode has been developed for more than 20 years in some countries. Nevertheless, car-sharing is only emergent in recent years in China. To date, this mode is only officially operated in more than 10 large- and medium-sized cities in China. Car-sharing is a newly developed transportation mode that is beneficial in solving traffic-related problems. This mode also brings huge opportunities for car-sharing companies [27, 28]. The car-sharing mode in China is still in its initial developing stage; thus, many issues exist and should be addressed. In particular, most car-sharing companies cannot capture people's usage demands accurately. Consequently, these companies cannot identify potential users and retain high-value ones. The widespread geographical distribution of residents in China, which is characterised by differences in urban development and cultures, results in distinct characteristics of the user's travel behaviour. Consequently, the characteristics of car-sharing users in different regions of China must be deeply understood. However, only several studies have addressed this issue. To this end, this study analyses the rule of users' demand on the time axis based on the actual operational data of a car-sharing company in Gansu. The demand of different user categories is examined thoroughly. A category prediction model is developed to realise accurate advanced prediction of user categories. This work proposes to find a balance between predictability (longer time span between the current and the forecast points) and accuracy by dividing the observation and judgment periods for the first time

effectively. This model can provide data support for operators' dynamic resource management.

The rest of this work is organised as follows. In Section 2, we clean the acquired data and explore the rule of car rental and car return on a 24-h time axis. In Section 3, we use k-means clustering to divide all users and comparatively analyse the usage attributes of different user categories for determining various usage behaviour. In Section 4, we use C 5.0 decision tree to develop the user classification prediction model. Such model can predict the user category in advance based on the user's partial usage attributes during the observation period. Finally, in Section 5, we present the conclusions.

2. Car-Sharing Data Analysis

2.1. Dataset. The data are from a car-sharing company in Gansu, China, which was established in 2017 and provided a one-way station-based car-sharing service. The user can return a rented car to any car-sharing station, which may not be the origin station.

Users only need to register their personal information and pay a deposit once during the company's application. Thereafter, the users can complete a series of loan–return operations at a fixed site without managers. In September 2018, the company has a total of 1272 car-sharing stations and 655 shared cars, including 5 types of pure electric vehicles. The vehicle types E200, ZHIDOU2, EC200, and Lease Edition are economical, and E5 is comfortable.

The data contain the car-sharing rental order information of the company from May 2017 to September 2018 with a total number of 290,266 transactions. Table 1 shows the attribute contents of the acquired order data.

We initially clean the data by deleting duplicate orders and preoperation test data with an actual travel distance less than 1 km, which was created before the actual operation, to ensure their authenticity. In summary, 18,501 records are deleted. The remaining data contain 271,765 records. A total of 10,345 car-sharing users are available from May 2017 to September 2018.

A user may have several orders. Therefore, we conduct the analysis from the perspectives of the order and user. First, the analysis is performed in terms of each of the order data. We can analyse the usage pattern in different time periods of one day and observe whether a peak period of renting and returning of shared cars exists. Second, we can deeply analyse the individual characteristics from the user's perspective. Specifically, we analyse the multidimensional usage characteristics presented by the same user through multiple orders. The usage attributes of each category of users are analysed comparatively in detail after the users are classified.

2.2. Car-Sharing Order Data Analysis

2.2.1. Analysis of User Information. We use 10,345 users' registration information of car-sharing to analyse their attribute characteristics. Only the age and gender are selected in the following analysis because of data limitations.

Figure 1 shows the distribution of the user's age. The users' age spans are large. Most users are distributed in the age range of 22–38 years, with a proportion of 78.90%. Furthermore, only 1% of users are older than 57 years. Young and middle-aged people have a high tendency to use shared cars.

Figure 2 reveals the gender ratio of the user. The proportion of male users is 81.25%, whereas that of female users is 18.75%. The number of male users is almost 4.3 times that of female users. Therefore, a man has greater possibility of using shared cars than a woman.

2.2.2. Order Quantity. We draw a trend chart of the cumulative quantity of daily orders to understand the usage of car-sharing and provide an accurate grasp of the market prospects (Figure 3). The company's cumulative daily order quantity remains basically stable before February 2018. A trough is observed during the Spring Festival in February. Starting from March 2018, the order quantity rapidly increased until June 2018 and then remains basically stable again. The results of data observation and analysis indicated that the order quantity displays periodic fluctuations, and the two peak points correspond to Saturday and Sunday. To this end, the subsequent analyses are based on a weekly cycle.

We focus on the variation of demand within a day. At present, we have 16 months of order data, which include 74 weeks. Considering the periodic change rule of order demand in a week, we superimpose the data from Monday to Sunday separately by calculating the correlation coefficient between car rental and return from the aforementioned days (Table 2). The result showed that the similarity coefficient between weekdays, between weekends, and between weekdays and weekends is as high as 0.99, 0.9, and approximately 0.85, respectively. Evidently, the distribution from Monday to Thursday is similar; hence, we regard the average of weekdays as a research object and the average of weekends as another object. Therefore, we take the mean value for subsequent analysis, and a day is divided into 24 1 h time periods in detail.

The number behind the diagonal line corresponds to the similarity coefficient of the number of car returns.

Figure 4 shows the variations of rental and return demand during weekdays and weekends. The demand of car rental on weekdays has two evident peak time points. The time point of the early peak is 7 : 00 am, whereas that of the evening one is 5 : 00 pm. The car rental demand during the morning peak on weekends is approximately 1.8 times than that during weekdays. The high trend of car rental during weekends continues until 5 : 00 pm and then presents a downward trend. No rapid growth occurs during the evening peak. Thus, users have greater and more dispersed demand for car-sharing during weekends than that during weekdays. The car return demand during weekdays also has two evident peak time points and lags 1 h behind the car rental demand. Such demand increases on weekends from the morning until its peak at 6 : 00 pm and then begins to decline. The cumulative quantity of returning cars after 4 pm on weekends is higher than the car rental demand. This time period may be the end time after users' short trips on weekends.

Table 1: Attributions of the order data.

Fields	Description
Order ID	Set of numbers generated by the operating platform that is a unique identifier of the order
Total cost	Actual price paid by the user
Car rental station	Station where the user rents the car
Car return station	Station where the user returns the car
Start using time	Time when the user begins using the car
Actual rental time	Time period when the user actually uses the car from renting to returning
Time fee (RMB)	Cost due to time in the order (monetary unit of measurement in China)
Actual travel distance	Traveling distance of the order, which is the total distance of all discontinuous usages
Distance fee (RMB)	Cost due to mileage in the order (monetary unit of measurement in China)
User gender	Sex attribute of user
User ID number	User's ID number contains the user's age
User origin	Place of the user's birth or origin (obtain this information by identifying the user's ID number)
User mobile number	The user's mobile number is its unique identifier

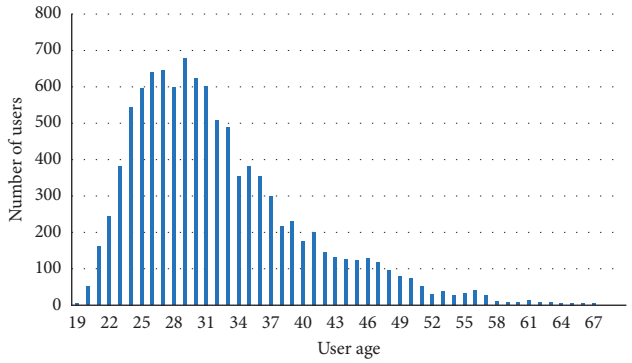

Figure 1: User age distribution histogram.

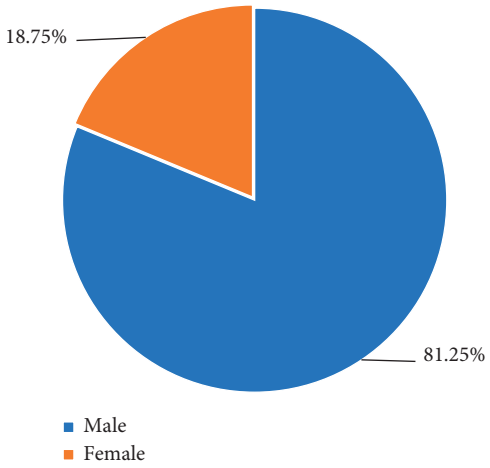

Figure 2: User gender-scale sector diagram.

On the basis of the abovementioned analysis, we can clearly recognise the peak period of car rental and return demand in one day. The operators can reasonably dispatch the vehicles by using the peak periods of renting and returning of the shared cars at each station and the number of parking spaces and other information. For the users, it is possible to understand the usage demand of the other users in advance and is beneficial to plan the rental time in advance.

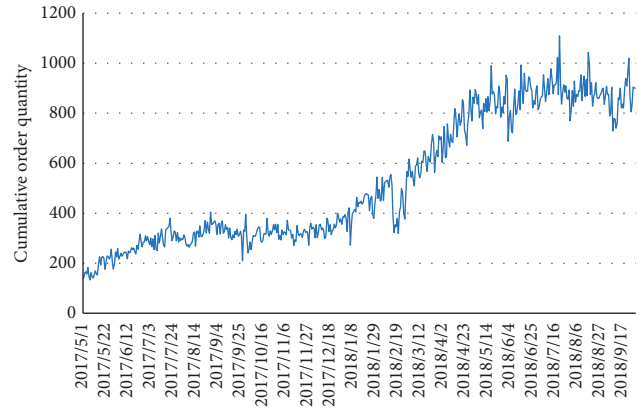

Figure 3: Daily order quantity trend.

2.2.3. Orders of Each Car-Sharing User. Considering that car-sharing is a new type of transportation mode, numerous issues still should be carefully addressed, particularly for the characteristics of users. To this end, this section attempts to analyse the characteristics of user demand in terms of the transactions for each car-sharing user. Figure 5 shows the frequency distribution of the number of orders for each user. Users who have only used shared cars once have a large proportion (nearly 2000 people). The number of users gradually decreases as the number of orders increases. In summary, 20% of the users only use shared cars once, 50% of the users have more than seven orders, 30% of the users have more than 20 orders, and only 10% of the users have more than 72 orders. A small proportion of the users have high-usage frequency. Nearly one-third of the users only have one to two orders, and these individuals are regarded as small-value users. The targeted activities for such users should be considered to stimulate user cost and promote the popularisation of car-sharing.

3. Clustering of Users

Given that diversity is a basic characteristic of users, this section classifies users based on the attribute data of users. We use the *k*-means clustering method to classify all users into different categories according to the two proposed user

Table 2: Similarity coefficient of days in a week.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	1	0.998	0.997	0.997	0.993	0.872/0.914	0.83/0.884
Tuesday	0.998	1	0.999	0.998/0.999	0.994	0.868/0.913	0.823/0.884
Wednesday	0.997	0.999	1	0.998	0.993	0.865/0.91	0.821/0.978
Thursday	0.997	0.998/0.999	0.998	1	0.995	0.878/0.921	0.834/0.893
Friday	0.993	0.994	0.993	0.995	1	0.885/0.934	0.845/0.908
Saturday	0.872/0.914	0.868/0.913	0.865/0.91	0.878/0.921	0.885/0.934	1	0.991/0.994
Sunday	0.83/0.884	0.823/0.884	0.821/0.978	0.834/0.893	0.845/0.908	0.991/0.994	1

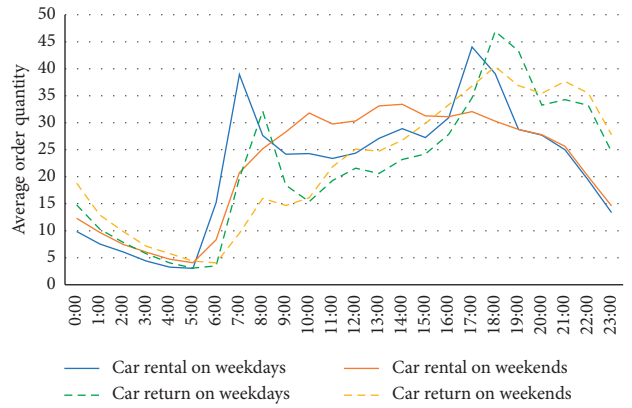

Figure 4: Average rental and return distribution on the time axis of one day for weekdays and weekends.

usage indicators. On this basis, we further analyse the representative usage attributes for the different user categories. This analysis is expected to be helpful for operators in designing the corresponding service strategies to fit the usage habits of different user categories.

3.1. Cluster Algorithm. Clustering is an important method in data mining. This method is a process of grouping physical or abstract objects into clusters. The objects within a cluster have high similarity, and those between clusters have low similarity. *K*-means clustering is an excellent and simple method for data mining. Therefore, this study uses the *k*-means clustering method to classify car-sharing users.

The *k*-means algorithm, also known as fast clustering method, has good scalability and efficiency and is, thus, appropriate for processing large datasets. The specific steps of the *k*-means algorithm are explained in detail in the Appendix section.

3.2. User Clustering. This study performs clustering to classify users with different loyalty effectively. Operators can determine the usage rules, thereby helping them improve the loyalty of users, prolong their use span, and create high value. The current user attributes cannot measure users' loyalty. Therefore, we use the *k*-means clustering method with two proposed indicators that measure the loyalty of users to cluster all users. We summarise the user attributes used in later definition to describe the classification indicators clearly (Table 3).

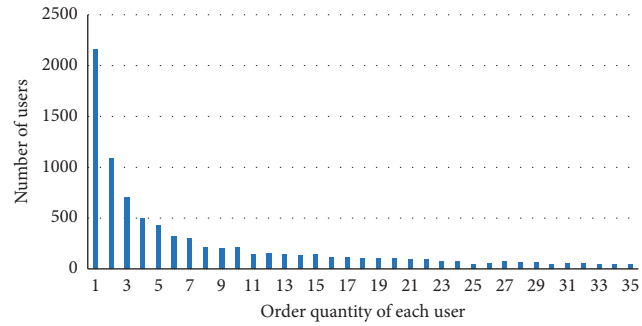

Figure 5: Frequency distribution of users based on the number of orders of each user.

The first indicator refers to the latent rate (LR). This indicator measures the time node when a user enters the car-sharing market. A user will use the shared car late when the value is large. LR can be calculated as follows:

$$\mathrm{LR} = \frac{\mathrm{LP}}{\mathrm{LC}}, \tag{1}$$

where LP is the latent period, which refers to the time span from the statistical start time to the first use, amongst which the user has yet to utilise a shared car; and LC is a life cycle and is a fix value that measures the time span of the statistic. These two indicators can be calculated as follows:

$$\begin{aligned} \mathrm{LP} &= (\text{first time} - \text{start time}) + 1, \\ \mathrm{LC} &= \text{end time} - \text{start time}. \end{aligned} \tag{2}$$

The second indicator refers to the persistence rate (PR). This indicator represents the time proportion of a user that exists in the car-sharing market. Such indicator can be used to measure the loyalty of users and the persistence of their demands as follows:

$$\mathrm{PR} = \frac{\mathrm{DA}}{\mathrm{SP}}, \tag{3}$$

where DA is the duration, which represents the time span of a user's first use of shared car to the last use; and SP is the sustainability period, which is the time span from the first use until statistical end time. These two indicators can be calculated as follows:

$$\begin{aligned} \mathrm{DA} &= (\text{last time} - \text{first time}) + 1, \\ \mathrm{SP} &= \text{end time} - \text{first time}. \end{aligned} \tag{4}$$

Table 3: Description of user time attributes.

User time attribute	Field description
First time	Time when a user first used a shared car
Last time	Last time the user used a shared car
Start time	Start time of statistics, that is, the day when the first order appeared (1 May 2017 in the dataset)
End time	Deadline of statistics, that is, the day after the last order appeared (30 September 2017 in the dataset)

The value of the two proposed indicators is within the range [0, 1], which avoids the problem caused by the difference in magnitude between the two indicators. These two indicators are directly used as input attributes of the k-means cluster algorithm. The final number of classification categories is determined by combining the DBI indicators and is used to measure the classification effect. The final aim is to achieve high similarity within classes and low similarity between classes. Specific calculation methods of indicators are shown in the Appendix. Table 4 shows the indicator results under different user classifications.

The results in Table 4 manifest that the users can be divided into four categories (Table 5) with low DBI values. For statement convenience, the four clusters are named as lost users, early loyal users, late loyal users, and motivated users.

Table 5 exhibits that the characteristics of different clusters are substantial. These characteristics are discussed as follows:

(i) The users of Cluster 1 have a relatively small latent and persistence rate values. This notion indicates that the users enter the market early but did not continue to use shared cars in the later period of statistics. Specifically, the users will no longer utilise shared cars after a short-term concentrated usage. Therefore, we define such individuals as lost users with short-term demand or just-for-trial usage. The proportion of the number of lost users is only 19.79%. The total cost of lost users is the second lowest, accounting for 11.07%, who belong to low-value users.

(ii) Cluster 2 is named as early loyal users. The users of Cluster 2 have a maximum persistence rate of 0.8796 and a small latent rate of 0.1605. Thus, the users have begun to use shared cars early and have a long usage duration. The early loyal users account for 19.00% of the total users. The created cost proportion is 49.02%, who belong to high-value users.

(iii) The users of Cluster 3 enter the market late with a high latent rate of 0.7474. However, the persistence rate is up to 0.8457, only second to the early loyal users, which is a user category that maintains high demand. We define these individuals as late loyal users based on the abovementioned analysis. The quantity of late loyal users accounts for 26.93%. The cost accounts for 32.57%, who belong to high-value users.

Table 4: Indicator results under different categories.

	2	3	4	5	6
DBI	1.0922	0.8541	0.8230	0.9090	0.9890

(iv) The users of Cluster 4 have the highest latent rate and the smallest persistence rate. Considering that Cluster 4 is the last users to enter the market, its usage characteristics are only partially displayed. These users also tend to keep high loyalty by operators and take incentives for stimulating users in utilising shared cars. Accordingly, these individuals are defined as motivated users, and their number is the largest amongst the four user categories. A total of 34.28% of the total users make 7.34% of the total cost. The value created by motivated users is low because of their short duration in the car-sharing market.

3.3. Characteristic Analysis for Users of Various Clusters. This section analyzes and compares various attributes to understand the characteristics of different user categories deeply. These attributes include average order quantity, average rental time, average travel distance, percentage of same pickup and return station, percentage of locals, and working day ratio.

We need to eliminate the problem caused by the difference in magnitude between these attributes to draw all of them on a graph conveniently and realise the comparison of attributes amongst different users. Thus, we normalise these indicators by adopting the maximum–minimisation method:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{5}$$

where x, $x_{\max}$, and $x_{\min}$ are the attribute, maximum, and minimum values, respectively. Consequently, the normalised data are within the range [0, 1].

Figure 6 shows the distribution of multiple usage attributes for different users. The classification of users in the average order quantity is the same as that in the percentage of locals. The loyal users generate additional orders with the increase in the percentage of locals. The percentage of locals of lost users is relatively low, that is, outsiders are likely to lose. The rental time and travel distance of the lost users are higher than those of the other users. In combination with the aforementioned analysis, the order quantity of the lost users

Table 5: Clustering results based on k-means algorithm.

Cluster centre	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster name	Lost users	Early loyal users	Late loyal users	Motivated users
Latent rate	0.2070	0.1605	0.7474	0.7755
Persistence rate	0.1075	0.8796	0.8457	0.1011
Quantity ratio (%)	19.79	19.00	26.93	34.28
Cost ratio (%)	11.07	49.02	32.57	7.34

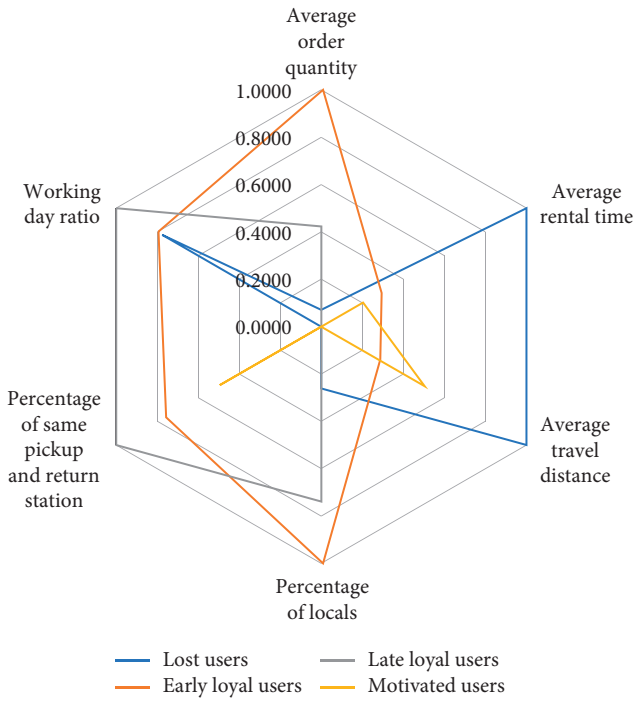

Figure 6: Radar charts of multiple user attributes for different users.

is less, but the demand for the time and distance is high. Late loyal customers tend to travel short time and distance. The short-term and short-distance rental characteristics of car-sharing are gradually accepted by people. The analytical result of users' rental and return vehicle stations indicates that loyal users tend to rent and return vehicles at the same station. By contrast, lost users tend to rent and return vehicles at different stations. The working day ratio of motivated users is lower than that of the three other user categories, indicating that such users are likely to use shared cars during weekend.

4. Prediction Model for User Classification

Car-sharing, as an emerging transportation mode, and operators should pay extensive attention on the early prediction of user categories and targeted measures. To this end, this section develops a prediction model that captures the characteristics of a user in advance based on the decision tree.

4.1. Decision Tree Prediction Model. This study develops a prediction model for user classification based on the decision tree. The decision tree is a classical nonparametric classification model that can predict a new sample set by summarising and refining the existing data inclusion rules in terms of the existing training set. This model is characterised by the good anti-interference of the extremum. The decision tree has excellent data analytical capabilities and intuitive visual graphical display. The amount of data used in this work is large. The partial variables are nonlinear, and some user attributes have outliers. Therefore, we use the decision tree model with good tolerance and interpretation ability. We use the C 5.0 algorithm to develop a decision tree. Formula (6) presents the expression of the developed decision tree model.

$$\text{class}(i) = f_{\text{DT}}\left(x_1^i, x_2^i, \ldots, x_n^i\right), \tag{6}$$

where i represents user i, x^i represents the usage attribute of user i, f_{DT} represent the algorithm of the decision tree, and $\text{class}(i)$ represents the user's final prediction category.

The C 5.0 algorithm introduces the self-adaptive enhanced boosting technique. This technique can iteratively generate a series of decision trees by increasing the sample probability of the misjudgment of the previous decision tree. Finally, all the decision tree models are combined together for classification prediction. This algorithm can improve the accuracy and enhance the robustness of the model. The growth algorithm of C 5.0 adopts the branch rule based on the information gain rate to find the optimal grouping variables and segmentation points.

4.2. Performance Indicators.

$$\begin{aligned} \text{accuracy}_{\text{model}} &= \sum_{m=1}^{k} \frac{N(m,m)}{N(m)}, \\ \text{accuracy}_{\text{class}(m)} &= \frac{N(m,m)}{N(m)}, \end{aligned} \tag{7}$$

where $\text{accuracy}_{\text{model}}$ is the prediction accuracy of the model, $\text{Accuracy}_{\text{class}(m)}$ is the prediction accuracy of the users whose user category is m, k represents the k categories of users, $N(m)$ represents the number of users whose actual user category is m, and $N(m,m)$ represents the number of users whose actual user category is m who also predicted to be m.

4.3. Period Definition and Division. The data acquisition period of user attributes is long, and the final prediction accuracy is high when the starting point and user prediction point of car-sharing life cycle are fixed. However, in reality,

we aim to judge the user categories early. In this case, we cannot obtain long-term usage data, that is, the high predictability corresponds to few data. Therefore, we must determine the number of months in advance to predict whether users can determine the optimal balance between predictability (short data cycle) and accuracy. This problem is investigated in this section.

Before the user classification prediction model is developed, the datasets should be preprocessed, and the observation and judgment periods should be divided reasonably. The users have already used shared cars during the observation period. Meanwhile, the judgment period is assumed to be the future time relative to the observation period. We predict the users' categories after the end of the judgment period. Specifically, the final proposed model predicts the users' category in the judgment period by using the multiple usage attributes of the user during the observation period as input. In this section, we need to find a balance between the observation and the judgment periods to provide practical support for future user prediction model construction.

4.4. Input Attributes. This study develops the prediction model with the judgment period from 1 month to 8 months by inputting user attributes. The users' data are divided into 70% training and 30% testing sets to obtain the optimal time division rule according to prediction accuracy. A total of 10 decision trees are constructed and combined with adaptive boosting technology.

In the initial analysis, the input attributes of the model only include two indicators, namely, the latent ratio and persistence ratio. These two indicators are used in the clustering method. In the next analysis, we add the users' usage attributes to obtain the optimal model.

4.5. Output Category. The final model output is the user prediction category. This output consists of four categories, namely, lost, early loyal, late loyal, and motivated users.

4.6. Framing and Testing. Pruning branches and leaves is a method used to overcome noise effectively. This method simplifies the decision tree and makes its structure easy to understand. Such method can also improve the classification efficiency and accuracy. Undercutting and overcutting reduce prediction accuracy. Therefore, the key to constructing a reasonable and efficient decision tree model is to select a suitable degree of pruning according to the selection of the pruning severity. In this study, we select 75 as the centre point and set different degrees of pruning with a standard deviation of 10. The optimal pruning degree is determined when the prediction accuracy of the model is high.

The boosting technology can improve the prediction accuracy of the model effectively. However, such technology can also cause the model to over fit. We use the optimal time division as the model basis for determining the optimal number of iterations of the boosting technology and set different iteration times to construct the model. The optimal number of iterations is achieved when the classification prediction accuracy is high.

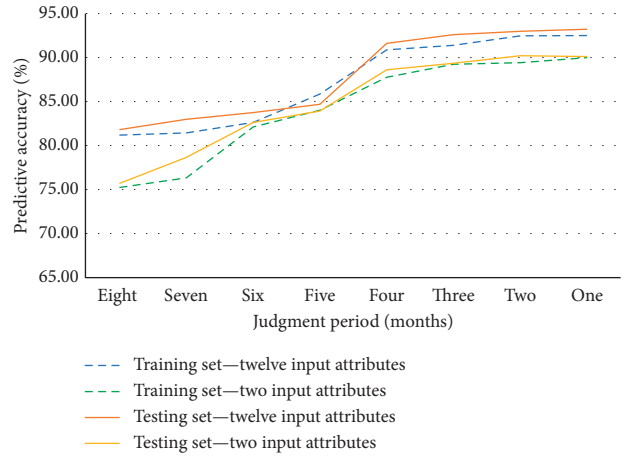

Figure 7: Prediction accuracy under different input properties and judgment periods.

Figure 7 shows the trend of the model accuracy with different user attributes under various judgment periods. When numerous usage attributes are used as input, the model accuracy is high. The attributes include average order quantity, total cost, average rental time, average travel distance, percentage of same pickup and return station, percentage of locals, working day ratio, average time-to-cost ratio, maximum rental time, and maximum travel distance. These attributes are analysed in detail in the third section. If the length of statistical period changes, then the values of these attributes will change in magnitude. These input attributes are standardised by adopting the maximum–minimisation method to make the model universal and usable.

Each prediction point in the graph is the result of the optimal model under this type of condition. The observation result manifests that the prediction accuracy has been greatly improved by predicting user categories four months in advance, and then it gradually increases. This time point can enhance forecast of users four months in advance and can achieve an optimal balance between the data volume and prediction accuracy. At this time, the observation-to-judgment period ratio is 3 : 1.

After the model is run according to the classification principle of the decision tree algorithm, the importance degree of each attribute of the optimal decision tree on the classification prediction result can be obtained and sorted on the basis of size. The mean and variance of each user's attributes, which rank the top six in the importance of user attributes, are calculated and separately expressed by M and S (Table 6). We use unprocessed data for statistical calculation to display the actual value of input attributes fully. Thus, the most important impact attributes in the model construction include the persistence ratio and latent ratio, followed by the average order quantity, the maximum travel distance, the maximum rental time, and the average travel distance.

In the final optimal model, the training set has a prediction accuracy rate of 90.87%, and the test set has an

Table 6: Ranking of importance indexes of various users under the optimal model.

Usage attribute	Importance	Lost users		Early loyal users		Late loyal users		Motivated users	
		M	*S*	*M*	*S*	*M*	*S*	*M*	*S*
Persistence ratio	0.20	0.15	0.19	0.83	0.26	0.75	0.32	0.25	0.31
Latent ratio	0.19	0.27	0.17	0.21	0.17	0.84	0.11	0.86	0.09
Average order quantity	0.07	8.69	15.46	55.50	69.02	20.70	27.31	4.48	7.49
Maximum travel distance	0.07	86.47	122.19	135.94	191.96	68.82	66.50	53.53	73.20
Maximum rental time	0.07	1154.11	2586.66	1666.97	2712.90	511.49	930.14	429.49	1700.16
Average travel distance	0.07	44.13	46.52	33.19	28.60	28.07	20.42	35.68	36.06

Table 7: Prediction accuracy of various users under the optimal model.

	Optimal model (%)	Lost users (%)	Early loyal users (%)	Late loyal users (%)	Motivated users (%)
Training set	90.87	98.31	90.40	85.70	86.86
Test set	91.61	97.93	91.82	87.01	87.94

accuracy rate of 91.61%. Table 7 shows the prediction accuracy of the four user categories on the training and test sets.

On the basis of aforementioned research, the multidimensional usage attributes can be used as the input to construct the classification prediction model four months in advance. This model can achieve high accuracy prediction of user categories. Setting the ratio of observation to judgment periods to 3 : 1 can not only achieve the early prediction of user categories but also ensure high accuracy. In this work, the observation and judgment periods are defined for the first time, and the mechanism of dividing them is analysed. The final conclusion can provide support for determining the prediction time point and defining the observation period to obtain sufficient effective data for completing the prediction.

The model construction can provide quantitative decision reference basis for the operations and managements of car-sharing. This basis is beneficial for operators in conducting scientific and reasonable user management to retain users and encourage continuous usage effectively.

5. Conclusions

The data analysis can confirm that the quantity of car-sharing orders maintains a high growth rate. Therefore, the car-sharing market has a good developmental prospect. However, the proportion of users who repeatedly utilize car-sharing is small, and almost 50% of these users use car-sharing less than 6 times. Morning and evening peaks can be observed in the distribution of car rental and return on weekdays. On weekends, orders are primarily concentrated from 10 : 00 am to 7 : 00 pm without demand peak.

We use the *k*-means clustering method to divide all users into four categories, namely, lost, early loyal, late loyal, and motivated users, by combining the two indicators, namely, the latent ratio and persistence ratio. Lost users' rental time and travel distance tend to be higher than those of the other users. The lost users account for 19.79% of the total users and create 11.07% of the total cost; therefore, they are considered low-value ones. Early loyal users account for 19% of the total users and create 49.02% of the total cost; thus, they are considered high-value ones. Motivated users have a great possibility to become loyal users. Therefore, operators are required to take measures for extending these users' duration to promote substantial economic benefits. Late loyal users have the lowest rental time and travel distance. This observation is the same as conclusion drawn by some Chinese scholars. The use of car-sharing for a short-time travel will also become a major trend.

On the basis of the user classification with the *k*-means clustering method, the C 5.0 decision tree classification prediction model takes the user's multidimensional usage attributes as the input. The aforementioned model also predicts the user category four months in advance with an accuracy of 91.61%. Accordingly, the optimal balance of predictability and accuracy of prediction is achieved. The prediction model can predict the user category in advance according to the attributes of the person by using the car-sharing service over a period of time. The prediction effect has reached a relatively good level when the ratio of observation to judgment periods is 3 : 1. The prediction is beneficial for operation managers in executing measures for different user categories in a targeted manner and rationally arranging resource delivery. This approach can also provide basic research for the operation scheduling and site layout of upcoming car-sharing operation.

In this research, we use data from a company in the early stage of car-sharing development. Therefore, further studies need to be carried out. We can consider the influence of other external factors, such as weather and incentives, on car-sharing usage. We can also analyse the data of different regions and identify the developmental rules and usage characteristics. In this manner, we can perform the early estimation of user characteristics and rationally arrange vehicle resources in diverse regions [29].

Appendix

A. Basic Principle and Algorithm Steps of *K*-Means Clustering Algorithm

K-means, an unsupervised clustering method, is commonly used to partition samples into *k* clusters automatically. This clustering method aims to assign all N samples into k

clusters by minimising the sum of point-to-centre distances as follows:

$$J = \arg\min_C \sum_{i=1}^{k} \sum_{x \in C_i} \left\| \vec{x} - \vec{\mu_i} \right\|, \quad (A.1)$$

where $C = \{C_1, C_2, \ldots, C_k\}$ indicates k clusters; $\vec{x}$ is an $N \times R$ feature matrix; R represents the dimensions of the matrix, each row is a single observation or sample; and $\vec{\mu_i}$ indicates the cluster centre of the ith cluster. The detailed process of this clustering method is presented as follows:

Step 1: initialisation

Randomly select k cluster centre for all feature samples.

Step 2: assignation

Assign each sample to the nearest cluster centre by measuring the distance between the sample and each centre as follows:

$$C_i^t = \left\{ x_p : \left\| x_p - \mu_i^t \right\| \le \left\| x_p - \mu_j^t \right\|^2 \forall j, 1 \le j \le k \right\}. \quad (A.2)$$

Step 3: update

Find all samples in each cluster and determine the new cluster centre using

$$\mu_i^{t+1} = \frac{1}{N_i^t} \sum_{x_j \in C_i^t} x_j, \quad (A.3)$$

where N_i^t is the sample number of the ith cluster at the tth iteration.

Step 4: repeat Steps 2 and 3 until the cluster centre remains unchanged or the function converges.

B. Evaluation Index of K-means Clustering Algorithm

Clusters of clustering results are divided into $C = \{C_1, C_2, \ldots, C_k\}$, and the following attributes are defined as follows:

$$\operatorname{avg}(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \le i \le j \le |C|} \operatorname{dist}\left(x_i, x_j\right), \quad (B.1)$$

$$d_{\text{cen}}\left(C_i, C_j\right) = \operatorname{dist}\left(\mu_i, \mu_j\right), \quad (B.2)$$

where $\operatorname{dist}(,)$ is used to calculate the distance between two samples; μ represents the central point of cluster C, $\mu = (1/|C|)\sum_{1 \le i \le |C|} x_i$; $\operatorname{avg}(C)$ corresponds to the average distance between samples in cluster C; and $d_{\text{cen}}(C_i, C_j)$ corresponds to the distance between centre point of clusters C_i and C_j.

The two indexes for measuring clustering performance can be deduced on the basis of formulas (B.1) and (B.2). The specific calculation is presented as follows:

Davies–Bouldin index (DBI),

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \ne i} \left(\frac{\operatorname{avg}(c_i) + \operatorname{avg}\left(c_j\right)}{d_{\text{cen}}\left(C_i, C_j\right)} \right). \quad (B.3)$$

The clustering effect is good when the DBI value is small.

Data Availability

The data used in this paper were provided by a Gansu travel-sharing company.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (91846202 and 71961137008).

References

[1] H. Ohta, S. Fujii, Y. Nishimura, and M. Kozuka, "Psychological analysis of acceptance of pro-environmental use of the automobile: cases for car-sharing and eco-car," *Group Decision & Negotiation*, vol. 18, no. 6, pp. 537–566, 2009.

[2] M. Wang, E. Martin, and S. Shaheen, "Car-sharing in Shanghai, China: analysis of behavioral response to a local survey and potential competition," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2319, no. 1, pp. 86–95, 2011.

[3] R. G. Meijkamp, "Changing consumer behavior through eco-efficient services: an empirical study on car sharing in the Netherlands," *Business Strategy & the Environment*, vol. 7, no. 4, pp. 234–244, 2001.

[4] E. W. Martin and S. A. Shaheen, "Greenhouse gas emission impacts of carsharing in North America," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1074–1086, 2011.

[5] H. Nijland and J. V. Meerkerk, "Mobility and environmental impacts of car sharing in the Netherlands," *Environmental Innovation & Societal Transitions*, vol. 23, pp. 84–91, 2017.

[6] S. Tai, P. L. Mokhtarian, and S. Shaheen, "Car-sharing and the built environment: a GIS-based study of one U.S. operator," *Institute of Transportation Studies Working Paper*, vol. 11, 2008.

[7] W. Loose, *The State of European Car-Sharing*, Bundesverband Car Sharing e. V Brussels, Brussels, Belgium, 2010.

[8] J. Kang, K. Hwang, and S. Park, "Finding factors that influence carsharing usage: case study in Seoul," *Sustainability*, vol. 8, no. 8, p. 709, 2016.

[9] S. Shaheen, L. Cano, and M. Camel, "Exploring electric vehicle carsharing as a mobility option for older adults: a case study of a senior adult community in the San Francisco Bay Area," *International Journal of Sustainable Transportation*, vol. 10, no. 5, pp. 406–417, 2016.

[10] R. Clewlow, "Carsharing and sustainable travel behavior: results from the San Francisco Bay Area," *Transport Policy*, vol. 51, pp. 158–164, 2016.

[11] S. M. Zoepf and D. R. Keith, "User decision-making and technology choices in the U.S. carsharing market," *Transport Policy*, vol. 51, pp. 150–157, 2016.

[12] N. Wang and R. Yan, "Research on consumers' use willingness and opinions of electric vehicle sharing: an empirical study in Shanghai," *Sustainability*, vol. 8, no. 1, p. 7, 2015.

[13] D. Efthymiou and C. Antoniou, "Modeling the propensity to join carsharing using hybrid choice models and mixed survey data," *Transport Policy*, vol. 51, pp. 143–149, 2016.

[14] A. Cartenì, E. Cascetta, and S. D. Luca, "A random utility model for park & carsharing services and the pure preference for electric vehicles," *Transport Policy*, vol. 46, pp. 49–59, 2016.
[15] T. H. Stasko, A. B. Buck, and H. Oliver Gao, "Carsharing in a university setting: impacts on vehicle ownership, parking demand, and mobility in Ithaca, NY," *Transport Policy*, vol. 30, no. 4, pp. 262–268, 2013.
[16] Y. Hui, W. Wang, and Q. L. Sun, "A study on the use behavior characteristics of car-sharing users: a case study of Hangzhou Che Fen Xiang," *Communication Shipping*, vol. 3, no. 5, pp. 24–29, 2016.
[17] C. Qian, W. Li, M. Ding, Y. Hui, Q. Xu, and D. Yang, "Mining carsharing use patterns from rental data: a case study of Chefenxiang in Hangzhou, China," *Transportation Research Procedia*, vol. 25, pp. 2583–2606, 2017.
[18] Y. Hui, M. Ding, K. Zheng, and D. Lou, "Observing trip chain characteristics of round-trip carsharing users in China: a case study based on GPS data in Hangzhou city," *Sustainability*, vol. 9, no. 6, p. 949, 2017.
[19] K. M. N. Habib, C. Morency, M. T. Islam, and V. Grasset, "Modelling users' behaviour of a carsharing program: application of a joint hazard and zero inflated dynamic ordered probability model," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 2, pp. 241–254, 2012.
[20] S. Jian, T. H. Rashidi, and V. Dixit, "An analysis of carsharing vehicle choice and utilization patterns using multiple discrete-continuous extreme value (MDCEV) models," *Transportation Research Part A Policy & Practice*, vol. 103, pp. 362–376, 2017.
[21] S. Jian, T. H. Rashidi, K. P. Wijayaratna, and V. V. Dixit, "A spatial hazard-based analysis for modelling vehicle selection in station-based carsharing systems," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 130–142, 2016.
[22] T. Niels and K. Bogenberger, "Booking behavior of free-floating carsharing users," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2650, no. 1, pp. 123–132, 2017.
[23] M. Schussler and K. Bogenberger, "Fusion of carsharing and charging station data to analyze behavior of free-floating carsharing BEVs," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 541–546, Las Palmas, Spain, September 2015.
[24] S. Schmöller, S. Weikl, J. Müller, and K. Bogenberger, "Empirical analysis of free-floating carsharing usage: the Munich and Berlin case," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 34–51, 2015.
[25] J. Müller, G. Correia, and K. Bogenberger, "An explanatory model approach for the spatial distribution of free-floating carsharing bookings: a case-study of German cities," *Sustainability*, vol. 9, no. 7, p. 1290, 2017.
[26] Y. Wang, X. Yan, Y. Zhou, Q. Xue, and L. Sun, "Individuals' acceptance to free-floating electric carsharing mode: a web-based survey in China," *International Journal of Environmental Research and Public Health*, vol. 14, no. 5, p. 476, 2017.
[27] X. U. Qing, D. Y. Yang, Y. Hui, X. J. Lai, and Y. A. Liu, *The Car Sharing in China is on the Way*, Traffic & Transportation, Beijing, China, 2014.
[28] B. Roland, *Sharing the Future—Perspectives on the Chinese Car Sharing Market*, Roland Berger Strategy Consultants, Munich, Germany, 2014.
[29] Y. Hui, W. Wang, M. Ding, and Y. Liu, "Behavior patterns of long-term car-sharing users in China," *Transportation Research Procedia*, vol. 25, pp. 4662–4678, 2017.