A Prediction Model of Online Car-Hailing Demand Based on K-means and SVR

Bang Chen¹, Shenghan Zhou¹, Houxiang Liu¹, Xinpeng Ji¹, Yue Zhang¹, Wenbing Chang¹, Yiyong Xiao¹, and Xing Pan¹,*

¹School of Reliability and Systems Engineering, Beihang University, Beijing, China

*Corresponding author email: panxing@buaa.edu.cn

Abstract. The paper proposed a prediction model of online car-hailing demand based on K-means and support vector regression (SVR) methods. In the past few years, online car-hailing market demand has grown rapidly, and prediction of rapid demand growth has become a hot topic. This study takes the initial longitude and latitude of online car-hailing orders as the eigenvalues for K-means clustering. The clustering results are taken as the result of area division. The number and size of potential demand areas could be determined automatically. This method of area division solves the shortcomings of traditional artificial meshing division and existing administrative division methods. The model takes a small-sample data set as the application object and uses the SVR method for data regression. Finally, we conduct an empirical study on Didi's real data set in the core area of Chengdu City, China. The final experimental results suggest that the area division method based on K-means is reasonable and that the demand prediction model based on K-means and SVR is effective.

1. Introduction

With the development of Internet technology, the online car-hailing industry, as a new sharing economic model, has been developing rapidly, and a large number of online car-hailing platforms such as Uber and Didi have been established. Online car-hailing platforms have attracted a large number of users because of their convenience, quickness, flexibility and lower price [¹]. People's demand for online car-hailing has also increased dramatically. For example, Didi, China's largest online car-hailing platform, handled approximately 14.1 billion orders in 2018 [²]. Therefore, the accurate prediction of online car-hailing demand has become an important research direction, which is of great significance for drivers to receive more orders and for online car-hailing platforms to run more efficiently.

The prediction of online car-hailing demand mainly includes the following three parts: area division, characteristic variable selection and data regression. Area division divides orders into areas to predict the demand of different areas. At present, scholars generally use the methods of artificial meshing division [³-⁴] and existing administrative division [⁵-⁶] to divide the prediction areas. However, the method of artificial meshing division cannot automatically determine the number and size of grids, and the method of existing administrative division lacks certain rationality. Selecting characteristic variables includes selecting the key factors that affect the demand of online car-hailing, which may be longitude and latitude, timestamp, weather and air quality [⁷-⁹]. However, for different data sets, the characteristic variables are often different. For example, for data sets from cities with little change in weather conditions, weather may not be a characteristic variable for demand. Therefore, further correlation analysis is needed when selecting the characteristic variables. Data regression generates the
prediction model by regressing the value of historical feature variables and the value of historical target variables. To date, most studies generally adopt an artificial neural network (ANN) method to forecast the online car-hailing demand. For example, Ke et al. proposed a deep learning approach, named fusion convolutional LSTM (Long Short-Term Memory) net, for short-term online car-hailing demand prediction and verified the validity of the model on the real data set of DiDi in Zhejiang, China. Wang et al. proposed an end-to-end framework called Deep Supply-Demand (DeepSD) to predict the gap between the online car-hailing supply and demand in a certain area in the next few minutes, based on a novel deep neural network, and performed validation experiments on DiDi's real data set. Nejadettehad et al. used three kinds of recurrent neural networks, including simple RNN units, GRU (Gate Recurrent Unit) and an LSTM neural network, to predict the short-term demand of online car-hailing, and their experimental results proved the effectiveness of the model. However, these methods based on ANNs often do not perform well on small-sample data sets because the learning effect of ANNs often depends on the size of the data.

To solve the above problems, this paper proposes an online car-hailing demand prediction model based on K-means and SVR. A method based on K-means clustering is adopted to perform area division, which can automatically determine the number and size of each area, to solve the problem of blindness in traditional area division methods. The SVR method, which performs well on the small-sample data set, is used to perform data regression to improve the performance of the proposed prediction model on the small-sample data set. At the same time, to verify the effectiveness of the prediction model proposed in this paper, we conducted an empirical study on the real DiDi data set in the core area of Chengdu, China.

2. Methodology

2.1. Model Framework

The framework of the proposed model is shown in figure 1. First, for given data sets, order data, weather data and data preprocessing are needed. Second, K-means clustering is used to perform area division to add the area labels to the data sets. Third, the data sets are analyzed as a whole, such as the changes in the number of orders with the weather and times of day, to preliminarily filter the possible influencing feature variables of demand. Fourth, further correlation analyses are performed to select the final feature variables. Finally, SVR is used to perform data regression to generate the prediction model.

![Figure 1. Prediction model framework.](image)

2.2. Area Division Method based on K-means

In online car-hailing demand prediction research, scholars generally use the methods of artificial meshing division and existing administrative division to divide the prediction areas. However, the two methods are generally blind and cannot automatically determine the optimal number and size of the areas. Therefore, based on K-means clustering, this paper proposes an area self-division method, which can automatically determine the optimal number and size of the areas. The core idea of this method is that the generation of online car-hailing orders has regional agglomeration, that is, the orders of online car-hailing tend to spread around core buildings. For example, core buildings of orders around the schools are schools, while core buildings of orders around shopping malls are...
shopping malls. Based on this idea, the initial longitude and latitude of online car-hailing orders are taken as the eigenvalues for clustering to automatically determine the number and size of areas. For selection of the clustering method, we adopt the K-means clustering method, which is based on the distance proposed by Macqueen [11].

The process of K-means clustering is [12-13] as follows: 1) randomly select \(k\) initial cluster centers from the data sets; 2) calculate the Euclidean distance between the remaining data objects and the cluster centers \(C_i (1 \leq i \leq k)\); 3) find the closest cluster centers \(C_i\) to the target data object and assign the target data object to the cluster class \(i\); 4) calculate the average value of the data objects in each cluster class as the new cluster center and carry out the next iteration until the cluster center does not change, which indicates that the objective function has converged.

The formula of Euclidean distance between data objects and cluster centers \(C_i\) in space is as follows:

\[
d(x, C_i) = \sqrt{\sum_{j=1}^{m}(x_j - C_{ij})^2}
\]

where \(x\) is the data object, \(C_i\) is the cluster center, and \(m\) is the dimension of the data object.

K-means adopts the minimum error sum of squares as the objective function \(S\), and its specific formula is as follows:

\[
f = \min \sum_{i=1}^{k} \sum_{x \in X} |d(x, C_i)|
\]

where \(k\) is the number of cluster classes, and \(X_i\) is the set of all sample points with \(C_i\) as the cluster center.

Since K-means cannot automatically determine the optimal number of clusters, the sum of the squared errors (SSE) and the silhouette coefficient (SC) are usually used to select the optimal number of clusters. The calculation of SSE is small, but the accuracy is low. When \(k\) is less than the true number of clusters, the increase in \(k\) will greatly increase the degree of aggregation of each cluster, so SSE will decrease greatly. When \(k\) reaches the true number of clusters, the decrease range of SSE will decrease sharply and finally will tend to be stable [14]. Therefore, SSE can be used to judge the approximate range of the optimal number of clusters. While the calculation of SC is large, the accuracy is high. The calculation formula of SC is as follows:

\[
SC = \frac{\max(a(i), b(i))}{b(i) - a(i)}
\]

where \(a(i)\) is the intracluster similarity, and \(b(i)\) is the intercluster dissimilarity. Therefore, the smaller \(a(i)\) and the larger \(b(i)\), the larger \(SC\) and the better the clustering effect.

### 2.3. SVR Regression

To solve the problem that the traditional online car-hailing demand prediction models based on ANN are not suitable for small-sample data sets, this paper uses the SVR method to carry out data regression. SVR is a machine learning method based on the Vapnik-Chervonenkis dimension [15] and structure risk minimization. The goal of SVR is to construct an optimal hyperplane to minimize the error between training samples and the optimal hyperplane to achieve better fitting performance and generalization ability for unknown samples [16].

For a given online car-hailing demand data set, \(D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}\) is the training sample data set, \(x_i \in \mathbb{R}^m\) is the input vector and \(y_i \in \mathbb{R}\) is historical online car-hailing demand. Therefore, the corresponding optimization problem of SVR is shown in formulas (4) and (5):

\[
\min_{w, b, \xi, \hat{\xi}} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)
\]

\[
s.t. \begin{cases} 
    f(x_i) + b - y_i \leq \epsilon + \xi_i, \ 0 \leq \xi_i \ (i = 1, 2, ..., m) \\
    y_i - f(x_i) - b \leq \epsilon + \hat{\xi}_i, \ 0 \leq \hat{\xi}_i \ (i = 1, 2, ..., m)
\end{cases}
\]

where \(w\) is the weight, \(b\) is the offset, \(C > 0\) is the penalty factor, \(\xi_i, \hat{\xi}_i\) and \(\epsilon\) are the insensitive loss function parameters and s.t. is the constraint condition. According to KTT (Karush-Kuhn-
Tucker) conditions and the Lagrangian function, the regression expression formula (6) is obtained as follows:

\[ f(x_i) = \sum_{i=1}^{m} (\hat{a}_i - a_i) K(x_i, x_j) + b \] (6)

where \( K(x_i, x_j) \) is the kernel function and \( \hat{a}_i \) and \( a_i \) are Lagrangian multipliers. A variety of kernel functions can be used in the SVR model, and the prediction accuracy of approximate models constructed by different kernel functions is quite different. The typical kernel functions include the linear kernel function, polynomial kernel function, radial basis function (RBF) kernel function, etc. Due to the advantages of fewer parameters and low computational complexity, the RBF function is selected as the kernel function of the SVR regression model, and its specific expression is shown in formula (7) as follows:

\[ K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right) \] (7)

where \( \delta \) is the width parameter of the RBF kernel function, \( x_i \) is an input sample vector and \( x_j \) is the center of the RBF kernel function.

3. Experiments and Results Analysis

Table 1. Dididrivers’ driving track data chart.

| DRIVER ID     | ORDER ID           | TIME     | LNG    | LAT     |
|---------------|--------------------|----------|--------|---------|
| 3a7013bfbdcb48f7f7203ed5d3 | 464b015cf9532223c07df5abb9 | 1475299381 | 104.05892 | 30.65445 |
| 3a7013bfbdcb48f7f7203ed5d3 | 464b015cf9532223c07df5abb9 | 1475299399 | 104.0593  | 30.65445 |
| 3a7013bfbdcb48f7f7203ed5d3 | 464b015cf9532223c07df5abb9 | 1475299421 | 104.06025 | 30.65443 |
| ...          | ...                | ...      | ...    | ...     |
| 1ffbc7dc647c28e3f1bc6b6e  | a89a6e16b5adaea128aa4afdb7 | 1475310924 | 104.07356 | 30.65847 |
| 1ffbc7dc647c28e3f1bc6b6e  | a89a6e16b5adaea128aa4afdb7 | 1475310927 | 104.07356 | 30.65847 |

Table 2. Weather data chart.

| Date       | Maximum Temperature (°C) | Minimum Temperature (°C) | Weather Type | AQI |
|------------|--------------------------|--------------------------|--------------|-----|
| 2016-10-01 | 30                       | 19                       | Cloudy       | 64  |
| 2016-10-02 | 32                       | 19                       | Sunny        | 89  |
| 2016-10-03 | 29                       | 20                       | Cloudy       | 97  |
| ...        | ...                      | ...                      | ...          | ... |
| 2016-10-30 | 17                       | 11                       | Cloudy       | 67  |
| 2016-10-31 | 19                       | 11                       | Sunny        | 73  |

To verify the validity of the prediction model proposed in this paper, we conducted an empirical study on DiDi’s real data set in the core area of Chengdu, China. This data set records the driving track data of DiDi drivers from the upper half of the second ring road in Chengdu, China, in October 2016, and the details are shown in table 1. This data set records the drivers’ id, orders’ id, timestamp, latitude and longitude every three seconds. There are 28201506 data records in the whole data set, and the data set length is 2.91 GB. At the same time, considering the impact of weather conditions on the online car-hailing demand, we searched the daily weather conditions in October 2016 and obtained the weather data set, as shown in table 2. This data set records the date, maximum temperature, minimum temperature, weather types, and air quality index (AQI). Obviously, the original data sets cannot be used directly, so a series of preprocessing steps was carried out. We first clean up the data sets, including duplicate items, missing items and irrelevant items. For irrelevant items, it needs to be explained here. In DiDi drivers’ driving track data, only the first data record of each order that contains the time and location of the order is useful. Therefore, we treat the rest of the data records as irrelevant items. Then, we transform the data set to transform the timestamp into standard time and code the weather type. The whole empirical study is divided into the following three parts: the result
analysis of K-means area division, the overall analysis of data sets, the selection of characteristic variables and the comparative analysis of regression prediction models.

3.1. Result Analysis of K-Means Area Division
To solve the limitations of the traditional methods of area division, this paper proposes a new area division method based on K-means. First, we extract the longitude and latitude from DiDi drivers’ driving track data and then take the results as the eigenvalues for clustering.

To determine the optimal number of clusters, the paper first calculates the $SSE$ value when $k \in [2,15]$ to determine the approximate range of the optimal number of clusters $k$. The results are shown in figure 2. It shows that $SSE$ begins to decrease significantly after $k = 2$, which indicates that $k$ starts to approach the optimal number of clusters. Then, it decreases sharply after $k = 5$, indicating that $k$ has reached the optimal number of clusters. Therefore, it is judged that the number of optimal clusters $k \in [3,5]$. Finally, the $S_C$ value of $k \in [3,5]$ is calculated to determine the exact value of the optimal number of clusters $k$. The results are shown in figure 3. It can be seen from figure 3 that $S_C$ obtains the maximum value when $k = 4$, so the optimal number of clusters is 4.

![Figure 2. SSE varying with $k$ value.](image1)

![Figure 3. $S_C$ varying with $k$ value.](image2)

### Table 3. Geographical location of each cluster center.

| Class | LNG of Cluster Center | LAT of Cluster Center | Nearby landmark |
|-------|-----------------------|-----------------------|-----------------|
| 0     | 104.1055              | 30.7067               | Chengdu Zoo     |
| 1     | 104.0638              | 30.6637               | Tianfu Square   |
| 2     | 104.1056              | 30.6640               | Chengdu Bilingual Experimental School |
| 3     | 104.0521              | 30.7049               | Southwest Jiaotong University |

![Figure 4. K-means clustering results.](image3)

![Figure 5. The distribution of clustering centers on the map of Chengdu.](image4)
On the basis of determining the optimal number of clusters, the longitude and latitude coordinates are used as the eigenvalues to divide the areas. The clustering results are shown in figure 4, which shows that K-means divides the orders in different locations into four areas, and the orders in each area are evenly distributed around the cluster center. To further judge the rationality of the division results, this paper queries the corresponding positions of the clustering centers on the map and marks them on the map, as shown in table 3 and figure 5. Table 3 shows that all cluster centers are located near the landmark buildings. For example, the cluster centers of area 1 are located near Tianfu Square in the center of the city, and the cluster centers of area 3 are located near Southwest Jiaotong University. From figure 5, it can be further found that each cluster center is located in the east, south, west and north of the upper part of the second ring road of Chengdu. Therefore, based on table 3 and figure 5, the rationality of the areas divided by K-means clustering is further proven.

3.2. Overall Analysis of Data Sets and Selection of Characteristic Variables

According to previous studies [3,6,10], online car-hailing demand may be affected by holidays, regions, time periods, lowest temperature, highest temperature, weather types and air quality. However, for different data sets, the characteristic variables are often different. Among the factors preliminarily selected in this paper, holidays, areas, time periods and weather types are discrete variables, while lowest temperature, maximum temperature and air quality are continuous variables. Therefore, this paper first analyzes the change of the online car-hailing demand as a whole to directly determine some easily recognized discrete characteristic variables. For some continuous characteristic variables that are not easy to identify, the Pearson correlation coefficient $\gamma$ is used to determine the results.

![Figure 6. Thermal chart of the order distribution in each hour per day.](image)

In this paper, the daily order volume of DiDi in the upper half of the second ring road in Chengdu, China, in October 2016 was counted, and the results are shown in figure 6. Figure 6 shows that there are more orders from 9:00 to 21:00 every day and fewer orders from 0:00 to 6:00. Because of the obvious difference of orders in each hour per day, time periods are a characteristic variable. After that, we calculated the average order volume percentage by area type, weather type and date type, and the results are shown in figure 7. Figure 7 shows that the order volume of different areas shows obvious differences, among which the order volume of area 1 located in the center of the city accounts for the highest proportion, while the average order volume proportion of area 0 outside the second ring road is the lowest, which further explains the rationality of area division. However, there is no obvious difference in the order volume proportion of different weather types date types. Therefore, for this data set, weather types and date types are not characteristic variables. For the three continuous variables of the lowest temperature, the highest temperature and the AQI, the Pearson correlation coefficient $\gamma$ is
used to determine the correlation between them and the demand. The specific results are shown in table 4. From table 4, we can see that the minimum temperature, maximum temperature and AQI have weak correlations with demand for this data set, so they are not characteristic variables. Therefore, the time periods and area types are finally determined as the characteristic variables of this data set. It is necessary to explain why the variables such as date types, weather types, minimum temperature, maximum temperature and AQI are not characteristic variables of this data set. The root cause of this situation is the limited amount of data set because this data set only records the data of DiDi for one month. However, our basic methods and ideas are correct, and the model can be trained more accurately with enough data in the future.

![Proportion of Orders Under Different Areas](image1)

![Proportion of Orders Under Different Weather](image2)

![Proportion of Orders Under Different Date](image3)

**Figure 7.** Average order volume percentage by area type, weather type and date type.

**Table 4.** Pearson correlation $\gamma$ between the continuous variables and the demand.

| Variable   | Pearson Correlation Coefficient $\gamma$ |
|------------|------------------------------------------|
| Maximum    | -0.026                                   |
| Temperature| -0.016                                   |
| Minimum    | -0.088                                   |

3.3. **Comparative Analysis of Regression Prediction Models**

**Table 5.** Performance index of three regression models.

| Group ID | Method                | $R^2$ | RMSE   | MAE    |
|----------|-----------------------|-------|--------|--------|
| A        | SVR + Administrative Division | 0.64  | 675.68 | 897.87 |
| B        | SVR + K-means Division  | 0.93  | 385.67 | 275.06 |
| C        | BP + K-means Division   | 0.71  | 435.51 | 589.63 |

On the basis of determining the time periods and area types as the characteristic variables of this data set, this paper makes statistics on the order quantity of different areas in different time periods in October and obtains 2976 valid samples, which means it is a typical small-sample data set. To verify the effectiveness of the model proposed in this paper based on K-means and SVR, three groups of comparative experiments are set up in the regression prediction part: SVR + administrative division, SVR + K-means division and back propagation (BP) neural network + K-means division. We used goodness of fit $R^2$, root mean square error $RMSE$ and mean absolute error $MAE$ to compare the performance of the three regression models. The specific results are shown in table 5. It can be seen from table 5 that group B has the largest $R^2$, which indicates the best fitting degree, while the minimum $RMSE$ and $MAE$ indicate the highest accuracy. Therefore, in general, the regression model trained by group B has the best performance, which also proves that the K-means division method proposed in this paper can effectively improve the performance of the prediction model and that SVR regression is more effective than BP neural network regression in small-sample data sets.
4. Conclusion and Discussion

This paper proposed an online car-hailing demand prediction model based on K-means and SVR, which is suitable for small-sample data sets. We use the initial longitude and latitude of online car-hailing orders to perform K-means clustering to divide areas, which can automatically determine the number and size of areas. The final experimental results also verify the rationality and effectiveness of this method of area division. To solve the problem that the traditional prediction model based on ANN is not suitable for small-sample data sets, SVR regression is used in the demand prediction model proposed in this paper. Through the comparative experiment, it is verified that SVR regression is more effective than BP regression on a small-sample data set. In conclusion, the K-means areas and SVR regression can effectively improve the performance of the online car-hailing demand prediction model on a small-sample data set. However, the model does not consider the impact of road congestion on the demand for online car-hailing in the regression process. At the same time, due to the limitation of data volume, the impact of weather conditions on online car-hailing demand cannot be fully reflected. Therefore, in the future, we will actively seek cooperation with DiDi to obtain more data and add road congestion as a characteristic variable into the model to further improve its accuracy.

Acknowledgement

Thanks for the support of data source from Didi Chuxing GAIA Initiative. The study is supported by the National Natural Science Foundation of China (Grant No.71971013 & 71871003), the Fundamental Research Funds for the Central Universities (YWF-20-BJ-J-943), the key project of the production and education integration collaborative education series of the research branch of the Chapter of Industry-Education Integration Research, China Association of Higher Education (CJRH1901). The study is also sponsored by the Graduate Student Education & Development Foundation of Beihang University.

References

[1] Gilibert M, Ribas I, Rosen C and Siebeneich A 2019 Proc. Int. Conf. on Transportation Research Procedia (Barcelona) vol 47 (Amsterdam: North-Holland/American Elsevier) p 323-330
[2] Yi W and Yan J 2020 Appl. Energy vol. 277 p 115549
[3] Ke J, Zheng H, Yang H and Chen X 2017 Transp. Res. Pt. C-Emerg. Technol. vol 85 p 591-608
[4] Chiang M, Hoang T and Lim E 2015 Proc. Int. Conf. on Advances in Geographic Information Systems (Seattle Washington) vol 1 (New York: Association for Computing Machinery) p 1-10
[5] Nie M 2017 Transp. Res. Pt. C-Emerg. Technol. vol. 79 p 242-256
[6] Nejadettehad A, Mahini H and Bahrak B 2020 Appl. Artif. Intell. vol 34 no 9 p 674-689
[7] Moreira-Matias L, Gama J, Ferreira M, Mendes-Moreira J and Damas L 2013 IEEE Trans. Intell. Transp. Syst. vol 14 no 3 p 1393-1402
[8] Chen M, Zahiri M and Zhang S 2017 Transp. Res. Pt. C-Emerg. Technol. vol 76 p 51-70
[9] Farber S 2015 Q. J. Econ. vol 130 no 4 p 1975-2026
[10] Wang D, Cao W, Li J and Ye J 2017 Proc. Int. Conf. on Data Engineering (San Diego) vol 1 (Washington, DC: Institute of Electrical and Electronics Engineers) p 243-254
[11] MacQueen J 1967 Proc. Int. Conf. on Mathematical Statistics and Probability (California) vol 1 (California: University of California Press) no 2 p 281-297
[12] Bhattacharya A, Jaiswal R and Kumar A 2018 Theor. Comput. Syst. vol 62 no 1 p 93-115
[13] Guo W, Huang Z, Hou Y and et al. 2020 Proc. Int. Conf. on Control and Decision (Hefei) vol 1 p 5449-5453
[14] Vapnik N and Chervonenkis Y 1971 Theory Probab. Appl. vol 16 no 2 p 264-280
[15] Drucker H, Burges C, Kaufman L, Chris C, Kaufman L, Smola A and Vapnik N 1996 Proc. Int. Conf. on Neural Information Processing Systems (Cambridge, MA, USA) vol 1 (Cambridge: MIT Press) p 155-161