Research on Taxi Operation Characteristics by Improved DBSCAN Density Clustering Algorithm and K-means Clustering Algorithm

Saisai Jian¹, *, Dongyi Li², Yaqi Yu³
¹College of Electronic and Information Engineering, Liaoning Technical University, Huludao 125000, China
²School of Resources and Environmental Sciences, Wuhan University, Hubei 430070, China
³Entrepreneur College, Xi'an Jiaotong-liverpool University, Suzhou 215028, China
*Corresponding author: saisaijian2020@lntu.edu.cn

Abstract. With the development of urbanization, the problem of urban traffic congestion is becoming more and more serious. An improved k-means clustering algorithm was proposed to solve the problem that the traditional k-means clustering center could easily be affected by the clustering center and fall into the local optimal solution. Based on the big data of New York City taxis, the operational characteristics are analyzed. The experimental results show that the improved K-means clustering algorithm has a better clustering analysis effect in terms of hot demand for taxis.

1. Introduction
In recent years, with the rapid development of cities, the daily travel of urban residents and the city's public transportation are closely linked and mutually influenced. Therefore, it is urgent to build a good public transportation system. Taxi is one of the critical transportation tools for urban residents to travel daily. In recent years, due to the low operating efficiency of urban taxis, the number of taxis in China and the total size of passenger traffic fluctuate and decrease. With the rapid development of our economy, how to timely and accurately get a taxi in the city is the distribution features and the operating characteristics of urban taxis has become a problem to be solved in urban road traffic design and planning.

This paper mainly makes the following contributions:
(1) Conduct statistical analysis and descriptive analysis of taxi data of a certain period of time in a large city;
(2) Based on the traditional K-means clustering algorithm, optimization is carried out for the characteristic that the number of clusters obtained by the clustering algorithm needs to be predetermined;
(3) Based on the improved K-means clustering algorithm, the taxi data is clustered, and the time and place distribution rule of taxi traffic demand is mined.
2. Related work
In recent years, the clustering analysis based on the operation trajectory data of urban taxis has become a research hotspot for domestic and foreign scholars. A large number of K-means improved algorithms have been proposed successively. In 2017, Pan Wei et al. [1] proposed a clustering algorithm based on semi-supervised PCA attribute weighting K-means, effectively removing the influence of irrelevant attributes and noise attributes. In 2019, Wei Kangyuan et al. [2] overcame the local optimum by introducing the adaptive concept and immune clonal selection mechanism, and proposed the A2-GSA-Kmeans algorithm. In 2016, Cheng Jing et al. [3] proposed a clustering method combining distance measurement of time series and self-correlation of time series, which is of great significance for exploring urban functional areas and structural layout. In 2020, He Yue et al. [4] proposed a grid-based taxi-carrying hotspot clustering algorithm to provide information services for taxi operators and managers. In the same year, Liao Zhuhua et al. [5] proposed the taxi passenger carrying area recommendation algorithm based on sparse trajectory data. In 2017, Jiang Huijuan et al. [6] proposed a taxi-passenger data clustering algorithm based on RL-DBSCAN algorithm. In 2015, Liao Lvchao et al. [7] proposed a directed density fast clustering method (D-Optics) to extract structural information of complex road networks through clustering analysis of directed spatiotemporal data.

Based on the above research, the improved K-means clustering algorithm proposed in this paper has high adaptability to the mining of urban traffic demand.

3. Research method and data analysis

3.1. Taxi Big Data Analysis
The data used in this study were selected from the desensitized taxi operating conditions in New York City in 2019. In order to better cluster them and mine relevant features, this paper first analyzes their statistical indicators and judges the overall taxi operation characteristics of the city. From the overall trend, the overall taxi time in the first half of 2019 has been the peak state of orders. July and August are the lowest months of the year. The trend chart of monthly operating volume in 2019 is shown in Figure 1 below:

![Figure 1. Monthly Operating Volume Trends in 2019.](image)

In addition, in order to explore the change rule within a week, we take the week as the unit and the day as the analysis unit to conduct statistics, which can verify two rules: (1) the total number of orders
in the first half of the week is peak and stable. (2) The total number of orders fell off a cliff on Thursday, and the number of orders in the second half of the week began to rise from the bottom. Figure 2 shows the changing trend of operation volume for a specific week in July 2019:

![Figure 2. Trend chart of operating volume for a week in July 2019.](image)

Finally, in order to explore the relationship between commuting and taxi traffic volume, this paper takes the day as the research object and explores the distribution of taxi traffic demand over time in a day. According to the data characteristics, the following rules can be found:

1. The morning peak is not significant. From 5 a.m., the order volume picks up from the bottom. After about 8 a.m., the order volume starts to stabilize.
2. The evening peak shows a significant performance. The order volume has a small increase since 5:30 p.m., and the peak of taxi-hailing comes at about 7 p.m.
3. In the afternoon, the order volume fell at about 16:00.
4. The order quantity maintained a high level from 8 p.m. to 12 p.m. From 1 a.m., the order quantity fell off a cliff and continued to fall until about 5 a.m.

![Figure 3. Trend chart of operating volume on November 1, 2019.](image)
3.2. Model Preparation

We used data on New York City taxi operations in January 2019. The source data format declarations are shown in Table 1.

| Number | Characteristics of the data | Instructions               |
|--------|----------------------------|-----------------------------|
| 1      | trip_distance              | Distance                    |
| 2      | total_amount               | Cost                        |
| 3      | PULocationID               | Starting point              |
| 4      | DOLocationID               | Destination                 |
| 5      | profits                    | Cost per kilometer          |
| 7      | tpep_pickup_datetime       | Pick-up time                |
| 8      | tpep_dropoff_datetime      | Get off time                |
| 9      | time                       | Journey time                |

\[
time = tpep\_dropoff\_datetime - tpep\_pickup\_datetime	ag{1}
\]

In Formula 1, tpep_dropoff_datetime represents the time point of getting off the bus, and tpep_pickup_datetime represents the time point of getting on the bus. From the above formula, we can get the time, that is, the bus time. In formula (2), total_amount represents the taxi cost, and trip_distance represents the taxi distance. The taxi cost per kilometer can be obtained from the above formula.

\[
profits = \frac{\text{total}\_\text{amount}}{\text{trip}\_\text{distance}}	ag{2}
\]

In addition, if there are too many eigenvalues involved in clustering, the K-means algorithm will often fail to find the appropriate cluster structure due to the high dimension, thus resulting in the loss of interpretation of clustering results. Therefore, before clustering, this study first conducted correlation analysis on the indicators extracted above, hoping to reduce the clustering dimension by selecting some features with the low correlation coefficient, that is, small multicollinearity, so as to obtain better and more explanatory clustering results. Figure 4 shows the analysis results.

**Figure 4. Correlation coefficient matrix of characteristic data.**
Through the above analysis, we finally selected PULocationID, DOLocationID, Time, Profits. A total of four characteristics, as part of the properties of clustering.

3.3. Taxi Big Data Clustering Method Based on Improved K-Means
In the classical k-means algorithm, the position selection of K initialized centroid greatly influences the final clustering result and running time, so it is necessary to select the appropriate K centroid. If only the selection is entirely random, it may lead to slow convergence of the algorithm. Given the problem that the k-means algorithm is easily affected by the clustering center and falls into the local optimal solution, the k-means ++ algorithm selected in this paper is the optimization of the method of randomly initializing the centroid of k-means algorithm. The improvement lies in selecting the initial mean vector. The basic idea of selecting the initial centroid vector is that the distance between the initial clustering centers should be as far as possible.

3.3.1. K-Means++ algorithm steps.
(1) The initial centroid vector is selected as follows:
   Step 1: Randomly select a sample as the first clustering center C1;
   Step 2: 1. Calculate the shortest distance between each sample and the existing clustering center, denoted by \( D(x) \);
             2. Then calculate the probability of each sample being selected as the next clustering center: \( P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \). The higher the value of \( P(X) \) is, the higher the probability of being selected as the clustering center is.
             3. Finally, use the roulette wheel method to select the next clustering center;
   Step 3: Repeat Step 2 until K clustering centers are selected.
(2) Clustering is performed according to the distance between each point of the sample data and the center point;
(3) Updated to the central value of each cluster through the mean value of each cluster;
(4) Repeat processes (2) and (3) until the position of clustering center no longer changes.

3.4. Example analysis
3.4.1. Example overview. In this paper, the data of taxi passengers in New York City is used as the experimental test data set of the algorithm. The data set used in the experiment was taxi ridership data for January 2019, covering any area in the five boroughs of New York, and included data on the ridership of 7.76 million cabs.

The experimental implementation is carried out in Python language, using the computer software PyCharm, and under the computing environment of Intel i7-10510U CPU, MX250 GPU, and 8GB memory.

| Number | PULocationID | DOLocationID | time       | profits      |
|--------|--------------|--------------|------------|--------------|
| 1      | 231          | 79           | 0.221666667| 7.263157895 |
| 2      | 231          | 238          | 0.485      | 4.621004566 |
| ……    | ……           | ……           | ……         | ……           |
| 7760005| 161          | 249          | 0.503055556| 6.1875       |
| 7760006| 113          | 229          | 0.250555556| 4.75         |

Table 2. Characteristic values of the dataset.
3.4.2. Results analysis. In this paper, the boarding and disembarking points of passenger load data were analyzed by clustering, and the data were divided into 9 clustering clusters according to the elbow rule. The visualization of clustering results was shown in Figure 5.

![Figure 5. Cluster analysis results.](image)

After cluster analysis, the 9 cluster centers are shown in Table 3, combined with Table 3 and Table 4. In the areas of No. 1, 2, 3 and 6 clustering centers, there are a lot of data of taxi passengers and a great demand for taxis. For taxi operators, they need to invest more taxis in these areas. For the city managers, the public transportation in areas 1, 2, 3 and 6, they need to increase the input in transportation.

| Cluster number | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|----------------|----|----|----|----|----|----|----|----|----|
| PULocationID   | 149| 242| 153| 72 | 148| 238| 64 | 63 | 238|
| DOLocationID   | 238| 242| 150| 150| 62 | 150| 238| 60 | 61 |

**Table 4. Number of cases in each cluster.**

| Cluster number | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|----------------|------|------|------|------|------|------|------|------|------|
| Number of cases (1000) | 1212 | 1194 | 1396 | 651  | 749  | 1047 | 459  | 450  | 508  |

**Table 5. Distance between cluster centers of each cluster.**

| number | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|--------|------|------|------|------|------|------|------|------|------|
| 1      | 0.00 |      |      |      |      |      |      |      |      |
| 2      | 93.08| 0.00 |      |      |      |      |      |      |      |
| 3      | 88.09| 128.00| 0.00 |      |      |      |      |      |      |
| 4      | 116.93| 193.30| 81.00| 0.00 |      |      |      |      |      |
| 5      | 176.00| 203.07| 88.14| 116.28| 0.00 |      |      |      |      |
| 6      | 125.16| 92.09| 85.00| 166.00| 125.87| 0.00 |      |      |      |
| 7      | 85.00| 178.04| 125.16| 88.36| 195.02| 194.99| 0.00 |      |      |
| 8      | 197.69| 255.27| 127.28| 90.45| 85.02| 196.79| 178.00| 0.00 |      |
| 9      | 198.12| 81.10| 123.07| 188.35| 90.00| 89.00| 248.20| 175.00| 0.00 |
The distance between the centers of each cluster subcluster is obtained according to the sum of the center distances of each dimension subcluster. The distance matrix of the center distances of subcluster is shown in Table 5 above. The diagonal element of this matrix is the distance from the center of the subcluster itself, so it is specified as 0. As can be seen from Table 5, the distance from the center of the subcluster is highly dependent on pulocationID and the attribute values of DolocationID.

4. Conclusions
This paper uses New York City taxi passenger data based on the improved k-means clustering algorithm to cluster the extracted data, and studies the operation characteristics of taxis in large cities. In view of the traditional k-means algorithm has a vast defect -- the convergence depends heavily on the initialization of the clustering center, leading to the local optimal solution. Therefore, a taxi-passenger data clustering method based on the improved K-means clustering algorithm was proposed. The experiment shows that the K-means ++ algorithm has a good effect on the clustering analysis of taxi-carrying hot spots, which can better reflect the needs of taxi-carrying hot spots. It should be noted that the algorithm is slow in handling massive data, and the optimization degree of the algorithm needs to be further improved. In the future experimental research, we will continue to improve and optimize the algorithm to improve the running speed of the algorithm.

References
[1] Pan Wei, Zhou Xiaoying et al. Attribute Weighted Clustering Algorithm Based on Semi-supervised K-means [J]. Computer Applications and Software, 2016, 34(3): 189-194.
[2] Wei Kangyuan, He Qing, Xu Qinshuai. K-means Clustering Based on Improved Gravity Search Algorithm [J]. Application Research of Computers, 2019, 36 (11):3240-3244.
[3] Cheng Jing, Liu Jiajun et al. Analysis of spatial-temporal characteristics of taxi travel volume in Beijing based on time series clustering method [J]. Journal of Geo-information Science, 2016, 18 (9):1227-1239.
[4] He Yue, Wang Chongchang. Research on Mining Taxi Passenger Hot Spot Based on Temporal and Spatial Clustering [J]. Geomatics and Spatial Information Technology, 2020, 43(1):99-102.
[5] Liao Zhuhua, Zhang Jian et al. Taxi passenger carrying area recommendation based on coefficient trajectory data [J]. Acta Electronica Sinica, 2020, 11(11):2178-2185.
[6] Jiang Huijuan, Yu Yang. Improved DBSCAN Algorithm for Fine Extraction of Taxi Passenger Hot Spot [J]. Geospatial Information, 2017, 15(10):16-21.
[7] Liao Lvchao, Jiang Xinhua et al. Directed Density Method for Clustering Trajctional Data of Floating Vehicles [J]. Journal of Geo-information Science, 2015, 17(10):1152-1161.