Construction of Driving Condition Based on Discrete Fourier Transform and Improved K-Means Clustering Algorithm

Shuping Xu
School of computer science
Xi'an University of technology
Xi'an, 710032, China
E-mail: 563937848@qq.com

Yueqiu Huang
School of computer science
Xi'an University of technology
Xi'an, 710032, China

***Abstract*—In view of the low execution efficiency and slow convergence speed of traditional clustering algorithms, the initial clustering center has a greater impact on the clustering results, which leads to the problem of reduced algorithm accuracy. This paper proposes an improved K-means algorithm (Grid-K-means), that is, the Grid density is used to determine the initial clustering center; According to the density, the grid points are sorted to eliminate the idea of noise grid points and invalid grid points, so as to improve the efficiency and accuracy of the algorithm. First, the discrete Fourier transform was used to filter the original data, and then the principal component analysis and the improved K-means clustering algorithm were used to reduce and classify the kinematics fragments respectively, so as to construct the driving conditions of the vehicle. The experimental results show that this method can effectively improve the construction accuracy and reduce the construction time, and the fitted driving conditions can effectively reflect the local actual traffic conditions.**

Keywords-Driving Cycle; Principal Component Analysis; Improved K-Means Algorithm

I. INTRODUCTION

Vehicle Driving Cycle (Driving Cycle), also known as vehicle test cycle, which describes the speed-time curve of the vehicle. It is an important general-purpose basic technology in the automotive industry, which is used to limit pollutant emissions and energy consumption, and to apply evaluation technology [1-2] for new vehicle development. In the early days, developed countries such as the United States, Japan, Europe, etc., all established vehicle driving conditions based on their cities. For example, Lin et al. [3] analyzed and compared the driving conditions constructed in California through a stochastic process; Michel et al. [4]. studied that different vehicles should be represented by different driving conditions. In recent years, domestic scholars have conducted in-depth research on the driving conditions of automobiles and have achieved some research results. For example: Gao Jianping [5], Peng Yuhui [6] and others use the K-means algorithm respectively to construct the driving conditions of passenger cars in Zhengzhou and Xiamen, but the K-means algorithm is easy to fall into a local optimal solution, which affects the stability and accuracy of the construction

conditions; Hefei university of technology, represented by professor Shi Qin [7] team, led by typical road in hefei, for example, was carried out on the road data acquisition, processing, and using the particle swarm optimization method of fuzzy clustering [8], although can avoid local optimization to some extent, but for large data, there will be a problem of poor local search ability; Cao Qian et al. [9] constructed the driving conditions of Changchun city based on Markov chain, but when the data amount was small and the error of calculating the state matrix was large, the accuracy of the final construction conditions would be resulted.

Up to now, domestic scholars generally adopt the k-means algorithm to construct their own cities. The common problem is that the k-means algorithm is sensitive to outliers, the parameter K (cluster number) is difficult to determine, and it is easy to fall into the local optimal solution. Although some other algorithms have appeared, in the case of large amounts of data, the algorithm is inefficient in execution and cannot efficiently classify large quantities of data. Therefore, this paper proposes the Grid-K-means algorithm to achieve the clustering classification effect, and the clustering effect is stable, accurate and efficient, and the constructed working condition can comprehensively reflect the local traffic situation.

II. DATA PREPROCESSING

The data collected in this paper includes vehicle running time, speed, latitude and longitude and other information, sampling frequency of 1Hz, and data is collected on different roads. The original data is denoised [10-11] by discrete Fourier transform, the steps are as follows:

Step1: set $f(x)$ as the original signal, and transform the time domain signal to the frequency domain through Fourier transform:

$$F(u)=\sum_{x=0}^{N-1} f(x)e^{-j2\pi ux/N} \tag{1}$$

In the above formula: $F(u)$ is the frequency domain signal, x and u are the discrete signal time domain variables and frequency domain variables respectively, and N is the number of discrete signals.

Step2: adjust the value of the high frequency signal to 0 or a smaller value, while the value of the low frequency signal remains unchanged

Step3: to achieve signal smoothing and noise reduction, the frequency domain is restored to the time domain by the inverse Fourier transform, the expression is as follows:

$$f(x)=\frac{1}{N}\sum_{u=0}^{N-1} F(u)e^{j2\pi ux/N} \tag{2}$$

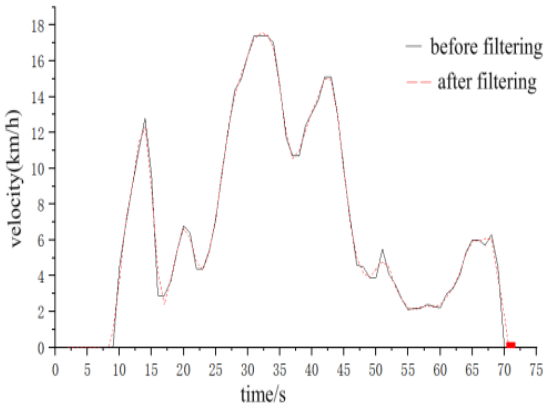

Figure 1. Contrast analysis of speed data filtering

From Figure 1, we can see that the discrete Fourier transform has a good filtering effect on the data, which can effectively eliminate the influence of noise on the data, thus avoiding the distortion in the calculated acceleration.

III. PRINCIPAL COMPONENT ANALYSIS AND IMPROVED K-MEANS CLUSTERING ALGORITHM

A. *Kinematics segment division and feature parameter extraction*

Kinematics segment refers to the vehicle speed range from the beginning of the idle state to the beginning of the next idle state [12]. As shown in Figure 2 below:

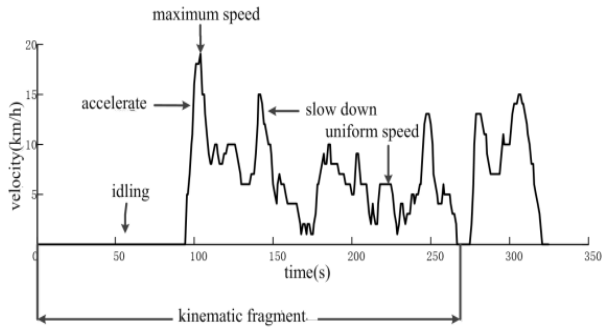

Figure 2. Schematic diagram of kinematics fragments

According to the domestic and foreign research sports state classification standards, the acceleration threshold can be selected as 0.1m/s2 to define four driving states:

1) Idling state: the engine continues to work, and its speed is zero during this process;

2) Deceleration state: the speed is not zero, and the acceleration during this continuous operation is less than or equal to -0.1m/s2;

3) Acceleration state: the speed is not zero, and the acceleration is greater than or equal to 0.1m/s2 during this continuous operation;

4) Constant speed state: the speed is not zero, and the absolute value of the acceleration during this continuous operation is less than or equal to 0.1m/s.

In order to fully and systematically reflect the driving characteristics of each feature and each operational segment, according to relevant literature[5-9], this paper selects 12 feature parameters to reflect each short-stroke feature:(segment duration) T/t ,(Driving distance) S/km ,(Average speed) $V_a/(km\bullet h^{-1})$,(Average driving speed) $V_x/(km\bullet h^{-1})$, (Idle time ratio) $T_i/\%$, (Acceleration time ratio) $T_a/\%$, (Deceleration time ratio) $T_d/\%$, (Cruise time ratio) $T_c/\%$,(Mean acceleration) $a_a/(m\bullet s^{-2})$, (Average deceleration) $a_d/(m\bullet s^{-2})$, (Standard deviation of acceleration) $a_{std}/(m\bullet s^{-2})$, (Standard deviation of velocity) $V_{std}/(km\bullet h^{-1})$.

Using Matlab to preprocess the 185726 sets of data to obtain 164089 sets of data, use R language programming to obtain 1585 kinematics segments, and then calculate the characteristic parameters of each kinematics segment respectively to obtain the number of samples (rows) x and characteristic parameters (columns), as shown in Table 1 below.

TABLE I. KINEMATICS FEATURE PARAMETER VALUES

Fragment number	T	S	V_a	V_x	T_i	T_a	T_d	T_c	a_a	a_d	a_{std}	V_{std}
1	119	203.29	6.15	7.60	0.06	0.17	0.08	0.03	0.42	-0.66	0.48	5.05
2	319	2320.7	26.19	33.63	0.14	0.26	0.19	0.11	0.30	-0.34	0.37	17.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
78	243	346	7.03	12.23	0.04	0.21	0.08	0.04	0.38	-0.61	0.39	5.01
79	167	459.12	5.97	8.04	0.06	0.19	0.05	0.05	0.44	-0.59	0.28	4.91
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1584	169	829.04	17.66	23.65	0.13	0.19	0.09	0.07	0.36	-0.50	0.48	14.8
1585	486	6799.9	50.37	52.65	0.02	0.18	0.13	0.30	0.19	-0.20	0.29	16.5

B. Principal component analysis

There are 1585 kinematics segments, each with 12 parameters, forming a 1585*12 matrix. The classification of 12-dimensional data requires a lot of work. Too much data will make the subsequent calculation steps more complicated, and some parameters are not independent of each other. Therefore, it is necessary to use principal component analysis to reduce the dimension of the data, which has reached a small number of principal components. It can reflect many original feature parameters.

This article uses SPSS for analysis based on the preprocessed data of the data set. The contribution rate and cumulative contribution rate of each principal component are shown in Table 2 below.

TABLE II. PRINCIPAL COMPONENT CONTRIBUTION RATE AND CUMULATIVE CONTRIBUTION RATE

Serial number	eigenvalue	contribution/%	Cumulative contribution/%
1	5.5605	43.23	43.23
2	3.2315	26.83	70.06
3	1.9901	16.14	86.20
4	1.0434	6.14	92.34
5	0.53476	3.11	95.45
6	0.45796	2.01	97.46
7	0.38341	1.22	98.68
8	0.21525	0.78	99.46
9	0.17450	0.23	99.69
10	0.10126	0.19	99.88
11	0.09324	0.08	99.96
12	0.02532	0.04	100

It can be seen from Table 2 that the cumulative contribution rate of the first three principal components has reached 86.2%, and the eigenvalues of these three principal components are all greater than 1. According to statistical theory [13], in principal component analysis, in general, it is sufficient to select principal components with a cumulative contribution rate of 85%, which basically represents all the information of the 12 characteristic parameters of the fragment, so the first three principal components can be selected for cluster analysis.

Through SPSS software analysis, the load matrix of each principal component can be obtained, so the correlation coefficients of the first 3 principal components and 12 characteristic parameters can be determined. As shown in Table 3 below.

TABLE III. PRINCIPAL COMPONENT LOAD MATRIX

Characteristic parameters	M_1	M_2	M_3
Segment duration T	0.132	0.231	0.745
distance S	0.293	0.134	0.045
Average speed V_a	0.719	0.463	-0.025
Average driving speed V_x	0.478	0.615	0.112
Idle time ratio T_i	0.125	-0.351	0.843
Acceleration time ratio T_a	0.694	-0.156	0.060
Deceleration time ratio T_d	0.923	0.341	-0.123
Cruising time ratio T_c	0.641	0.435	-0.045
Mean acceleration a_a	0.014	0.623	0.033
Mean deceleration a_d	0.366	-0.433	-0.052
Standard deviation of acceleration a_{std}	0.445	0.267	-0.067
Standard deviation of speed V_{std}	0.387	0.215	0.034

The principal component loading represents the correlation coefficient between the principal component and the original variable. The larger the correlation coefficient (absolute value), the more representative the principal component of the variable. According to the principal component loading matrix in the above table, some parameters corresponding to a certain principal component can be obtained. As shown in Table 4 below.

TABLE IV. THE MAIN EIGENVALUES REPRESENTED BY THE FIRST THREE PRINCIPAL COMPONENTS

Category	Eigenvalue
Primary principal component M_1	average speed, acceleration time ratio, deceleration time ratio, cruising time ratio
Secondary principal component M_2	average speed, acceleration time ratio, deceleration time ratio, cruising time ratio
Third principal component M_3	segment duration, idle time ratio

According to principal component analysis, the first, second and third principal components were selected for further analysis.

C. Improved K-means clustering algorithm

For traditional clustering algorithm clustering, the initial clustering center has a greater impact on the clustering results, which is easy to reduce the accuracy of the algorithm, unable to handle noisy data, and the clustering results are unstable and relatively inefficient when processing large-scale data sets [14-15]. This paper proposes a Grid-K-means clustering algorithm based on Grid improvement, which introduces the idea of grid division into K-means algorithm, reduces the number of manually set initial parameters, speeds up the ability to process data, and reduces the time complexity. This improves the overall efficiency of the algorithm. The main steps of the algorithm are as follows:

Step1. Let the data set be S, and each sample point $x_i = \{x_{i1}, x_{i2}, \ldots x_{im}\}$ has m attributes (the data set is S has m dimensions)

Step2. Read all the sample points in the data set, and record the maximum and minimum values of each dimension $E_{xi} = \{x_{i\max}, x_{i\min}\}$

Step3. Calculate the data range of each dimension, $Range_i = x_{i\max} - x_{i\min}$, $i = 1,2,\ldots n$

Step4. set the grid number of each dimension (GridNum), $GridNum = k \times 9$,among them, $GridNum \le \sqrt{n/2}$

Step5. Determine the grid step size of each dimension in the data set: $Steps_i = Range_i / Gridnum$, $i = 1,2,\ldots n$

Step6. Convert the data set into grid points, $Grid_{ij} = \frac{x_{ij} - x_{i\min}}{Steps_i}$ $i = 1,2,\ldots n;\ j = 1,2,\ldots m$

Step7. set the dynamic threshold, if the grid density is less than the threshold, it is noise, and if the grid density is equal to zero, it is invalid grid points $Threshold = n/(3 \times GridNum^n)$

Step8. Sort the grid points, select K high-density grid points as the initial cluster centers ($Center_1$), and select other k-1 high-density grid points outside the interference radius R as the remaining initial cluster centers ($Center_2, Center_3 \ldots, Center_k$), The initial cluster center interference radius $R = a \times (GridNum/(k+1))$, among them $a \in [1,2]$.

Step9. Iteratively and circularly calculating the grid clustering center until the grid clustering center no longer changes ; $Center_k' = (\sum Grid_{ij} \times Grid_{i0}) / \sum Grid_{i0}$

Step10. Reversely calculating the corresponding cluster center of sample points $v = v_1, v_2, \ldots v_k$, the formula is as follows: $x_{ij} = Grid_{ij} \times Steps_i + x_{i\min}$

In order to verify the performance of the improved algorithm Grid-K-means proposed in this article, simulation experiments select 3 sets of data (20,000, 80,000, and 100,000) for verification, and the accuracy of clustering results and the performance of different algorithms are compared respectively, as shown in the following table 5 below.

TABLE V. TABLE 5 COMPARISON OF ACCURACY AND PERFORMANCE OF DIFFERENT ALGORITHMS

Data set	k-means		k-means++		Grid-K-means	
	Accuracy	Elapsed time	Accuracy	Elapsed time	Accuracy	Elapsed time
20000	92.76%	25.82s	93.16%	34.86s	95.78%	25.83s
80000	88.32%	43.83s	89.32%	49.16s	92.36%	40.69s
100000	87.24%	58.12s	87.65%	67.94s	90.12%	50.38s

It can be seen that the performance improvement of grid-K-means algorithm is not significant when dealing with small data sets; but as the amount of data increases, the accuracy of the clustering results and the clustering speed of Grid-K-means are both Far beyond k-means, and k-means++ algorithm.

The kinematics fragments are divided into 3 categories by using the improved K-means algorithm. Table 6. shows the average eigenvalues of each category after clustering. As can be seen from Table 6, the first category is traffic congestion, because the average speed and average driving speed are low and the idle time ratio is high; The second type is smooth road conditions, because the average speed and the average driving speed are higher, the idle time ratio is in the middle, and the cruising time ratio is in the middle; The third category is high-speed road conditions, because the average speed and average driving speed are the highest, idle time ratio is the lowest.

TABLE VI. AVERAGE FEATURE VALUE OF EACH CATEGORY AFTER CLUSTERING

Clustering categories	v_a	T_c	v_x	a_a	T	T_i
Class 1	6.23	0.12	11.06	0.16	26.15	0.59
Class 2	16.25	0.25	18.26	0.59	35.56	0.23
Class 3	23.03	0.49	30.51	0.45	48.12	0.18

IV. CONSTRUCTION AND COMPARATIVE ANALYSIS OF DRIVING CONDITIONS

A. Construction of working conditions

According to the construction of the city, the duration of the working condition is about 1200s. According to the time curve, the kinematic segments are selected from the three categories respectively to be 56,89,225, forming the final representative working condition. Fitting the representative working condition time-speed graph and working condition time-acceleration graph, as shown in Figure 3 and Figure 4.

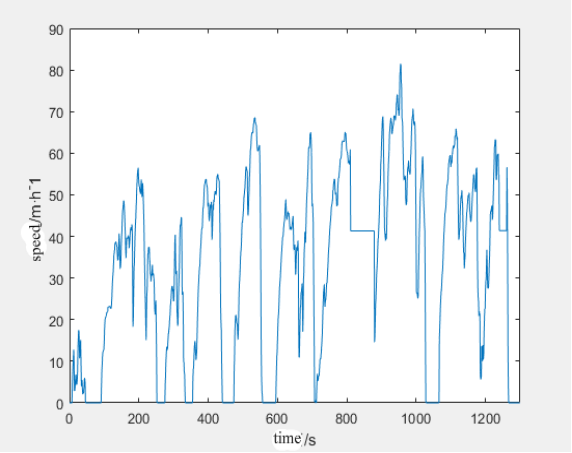

Figure 3. Time-speed curve of representative working conditions

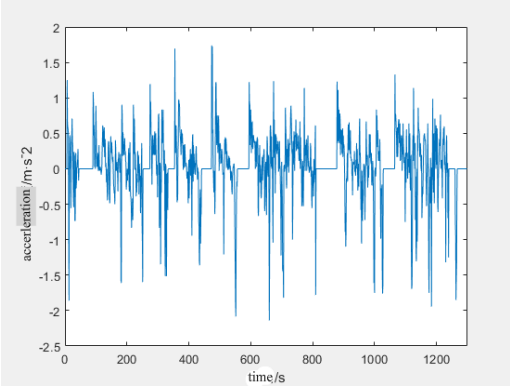

Figure 4. Time-acceleration curve of representative working conditions

B. Verification of Working Conditions

1) Error Analysis and Verification based on characteristic parameters

Comparing the characteristic parameters of the collected overall sample data with the characteristic parameters calculated based on traditional k-means, k-means++, and Grid-K-means, as shown in Table 7 below. Among them, E_s is the average relative error between the fitting representative working condition and measured values of all characteristic parameters of the test [16].

TABLE VII. CHARACTERISTIC PARAMETER RELATIVE ERROR

Characteristic parameters	Overall sample data	k-means		k-means++		Grid-K-means	
		Fitted value	Relative error/%	Fitted value	Relative error/%	Fitted value	Relative error/%
v_a	18.91	18.11	4.23	18.34	3.01	18.60	1.64
a_a	1.85	1.95	5.41	1.88	1.62	1.86	0.54
a_d	-2.36	-2.45	3.81	-2.43	2.97	-2.39	1.27
T_i	0.52	0.58	11.54	0.55	5.77	0.53	1.92
T_a	0.49	0.41	16.32	0.44	10.20	0.47	4.08
T_d	0.26	0.31	19.23	0.29	11.54	0.28	7.69
E_s			10.09		5.85		2.86

The average relative error of k-means is 10.09%, the average relative error of k-means++ is 5.85%, and the average relative error of all characteristic parameters of driving conditions constructed based on the improved k-means algorithm (Grid-K-means) is only 2.86%, and the relative error of the other eigenvalue parameters is less than 8%, so it shows that the Grid-K-means method can more accurately represent the driving conditions than the traditional method.

2) K-S test based on independent sample acceleration distribution

The driving acceleration of the whole sample is divided into three categories [17], （-∞,-0.64） m/s2，[-0.64,0.64] m/s2,（0.64,+∞）m/s2 ,and the sample speed of each type of acceleration is divided into six categories,[0,10） m/s，[10,20） m/s，[20,30） m/s，[30,40） m/s，[40,50） m/s，[50, +∞）m/s. In order to further test the similarity between acceleration distribution of representative working conditions and test data, K-S test of independent samples is carried out, and the test results are shown in Table 8 below. From this, we found that the acceleration distribution of the improved clustering algorithm Grid-K-means fitting working conditions is better than that of k-means++, which can better reflect the distribution frequency of acceleration.

TABLE VIII. SAMPLE K-S TEST

Method		Acceleration distribution（m/s2）		
		（-∞,-0.64）	[-0.64,0.64]	（0.64,+∞）
Grid-K-means	K-S value	0.68	0.57	0.53
	Similarity level	0.89	0.96	0.99
k-means++	K-S value	0.71	0.26	0.45
	Similarity level	0.79	0.44	0.82

V. CONCLUSION

In order to effectively construct vehicle driving conditions based on the city's own data, this paper uses discrete Fourier transform to preprocess the data, eliminate the noise in the data, and use principal component analysis to reduce the dimension of the data, and propose an improved k-means (Grid-K-means) optimization method to cluster the data, replace the initial parameter setting with the idea of dynamic grid division to improve clustering efficiency, speed up data processing, and reduce time complexity. Through experimental analysis: the average relative error between the driving conditions and the characteristic parameters of the experimental data is only 2.86%. The K-S test is used to verify and analyze the similarity between the fitting conditions and the experimental data. The results show that the constructed driving conditions are more accurate and more accurate and can better reflect the overall driving characteristics.

ACKNOWLEDGMENT

This research is partially funded by the Project funds in Shaanxi province University Student Innovation and Entrepreneurship Fund Project (S202210702077).

REFERENCE

[1] Li Mengliang, Li Wei, Fang Maodong, Liu Xiangmin, Du Chuanjin, Zhang Xingquan. The Parse Method of Actual Driving Cycle of Vehicle on Road [J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 2003(01):69-72.

[2] Gao Jianping, Sun Zhongbo, Ding Wei, Xi Jianguo. Development and precision research of vehicle driving conditions [J]. Journal of Zhejiang University (Engineering Edition), 2017,51(10):2046-2054.

[3] Jie Lin, Debbie A Niemeier. An exploratory analysis comparing a stochastic driving cycle to California's regulatory cycle [J]. Atmospheric Environment, 2002, 36(38).

[4] Michel Andr é. The ARTEMIS European driving cycles for measuring car pollutant emissions [J]. Science of the Total Environment, 2004, 334.

[5] Gao Jianping, Ding Wei, Sun Zhongbo, Xi Jianguo. Construction of passenger car driving conditions in Zhengzhou [J]. Mechanical design and manufacturing, 2018(08):102-105.

[6] Peng Yuhui, Yang Huibao, LI Mengliang, Qiao Xueqi. Research on Construction Method of Urban Road Vehicle Driving Conditions Based on K-means Cluster Analysis [J]. Automotive Technology, 2017(11):13-18.

[7] Zhang Rui. Construction and Research of Urban Road Car Driving Conditions [D]. Hefei: Hefei University of Technology, 2009

[8] Shi Qin, Wang Nannan, Chou Duoyang. Application of Particle Swarm Optimization Fuzzy Clustering Method in Vehicle Driving Conditions [J]. Chinese Management Science, 2011, 19(02):110-115

[9] Cao Qian, Li Jun, Liu Yu, Qu Dawei. Construction of passenger car driving conditions based on Markov chain[J]. Journal of Jilin University (Engineering Edition), 2018, 48(05):1366-1373.

[10]Cao Qian, Li Jun, Liu Yu, Qu Dawei. Construction of driving conditions based on big data and Markov chain [J]. Journal of Northeastern University (Natural Science Edition), 2019, 40(01):77-81.

[11]DingYifeng, LI Jun,Gai Hongchao,Chen Hao. The Application of Wavelet Transform in Processing Vehicle Speed Data Processing [J]. Science Technology and Engineering, 2017, 17(28):274-279.

[12]Miao Qiang, Sun Qiang, Bai Shuzhan, Yan Wei, Li Guoxiang. The Application of Wavelet Transform in Processing Vehicle Speed Data Processing[J]. Journal of China Highway, 2016, 29(11):161-169.

[13]Fan Jincheng, MEI Canglin. data analysis. Beijing: Science Press, 2010.

[14]Xu, Tian-Shi, Chiang, Hsiao-Dong,Liu, Guang-Yi, et al.Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data [J]. IEEE Transactions on Power Delivery, 2017, 32(2):609-616.

[15]Wen Peng. Research on Improved K-means Algorithm and New Clustering Effectiveness Index in Cluster Analysis [D]. Anhui University, 2019.

[16]Shi Qin, Chou Duoyang, Wu Jing. Research on driving conditions based on principal component analysis and FCM clustering [J]. Environmental Science Research, 2012, 25(01):70-76.

[17]Jiang Ping, Shi Qin, Chen Wuwei. Construction method of urban road driving conditions based on Markov [J]. Journal of Agricultural Machinery, 2009, 40(11):26-30.