Research on Driving Conditions Based on Principal Component and K-means Clustering Optimization

Huifeng Wang
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1318057134@qq.com

Jiaxiang Fang
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, China
E-mail: fangjiaxiang@st.xatu.edu.cn

Shuping Xu
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, China
E-mail: 563937848@qq.com

Feiyan Kou
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, China
E-mail: 13992096752@163.com

Huxiang Yang
Shaanxi Coal Industry Chemical Technology
Research Institute Co., Ltd.
Xi'an, China
E-mail: yanghx@sxccti.com

***Abstract*—In order to overcome the problems of traditional K-means algorithm being sensitive to the initial cluster centers and easily affected by noise points, this study proposes an enhanced K-means hybrid clustering algorithm that integrates improved principal component analysis and density optimization. By combining the distance optimization strategy with the density assessment mechanism, a data density evaluation model based on spatial distribution characteristics was established. The algorithm prioritizes data samples with large spacing in high-density areas as the initial cluster center candidate set. It realizes intelligent filtering of abnormal data points while improving the clustering quality, and selects characteristic parameters with higher principal component contribution rates to reconstruct driving conditions, and finally completes the fuel consumption characteristics verification. Experimental data show that the driving conditions constructed by this method have only a 1.17% statistical difference in the speed-acceleration joint probability distribution, and the relative error mean of key characteristic parameters remains at a low level. The research confirms that the constructed driving conditions are statistically significantly consistent with the actual road operation characteristics and can accurately characterize the essential characteristics of traffic flow in a specific area.**

Keywords-Improved Principal Component Analysis; Improved K-Means Clustering; Distance Optimization; Density Method

I. INTRODUCTION

Vehicle driving conditions, also known as vehicle operating cycles or driving cycles, refer to the mathematical representation of the speed-time curve that characterizes the vehicle's operating state in a specific traffic environment [1]. It provides core data support for fuel efficiency evaluation, emission control technology research and development, and intelligent traffic control, and directly affects the design of new energy vehicles and the accuracy of urban traffic carbon accounting [2]. Zhao Xuan's team [3] proposed a driving mode analysis method based on fuzzy C-means clustering. By integrating the time distribution characteristics of kinematic segments with multi-dimensional parameter correlation analysis, they achieved

intelligent identification and optimized reconstruction of typical driving conditions of urban electric vehicles. The operating condition curves they constructed have significant improvements in typicality indicators compared to traditional methods. Currently, K-means clustering is widely used in driving cycle synthesis. However, K-means clustering often has problems such as large dependence on the initial cluster center, isolated points, and sensitivity to noise data. Ma Fumin et al. [4] innovatively constructed a local density dynamic adaptation measurement model to accurately characterize the spatial distribution characteristics of data objects within a cluster, and based on this, designed a rough K-means clustering algorithm that integrates a local density adaptation mechanism. Yuan Yiming et al. [5] developed an optimized K-means text clustering algorithm based on density peak. This algorithm effectively overcomes the convergence instability problem caused by random initialization of centers in the traditional K-means algorithm by accurately selecting density peak points as the initial clustering centers, and significantly improves the reliability of clustering results. Although the above two methods optimize the initial cluster centers to a certain extent, they do not mention the impact of edge data and isolated points in the data set. Bao Zhiqiang et al. [6] only used an outlier removal algorithm to eliminate isolated points in the data set, but still used traditional K-means clustering to cluster the data set.

Based on the above research conclusions, this paper constructs an improved density-driven K-means clustering algorithm. The core innovation of this algorithm is to introduce density measurement indicators to screen the initial clustering centers, effectively suppress the interference of noise data on the selection of initial centers, and integrate enhanced principal component analysis technology to build a two-stage collaborative optimization framework to achieve intelligent synthesis of driving conditions.

II. Data Preprocessing

The measured data obtained in this study are from a light vehicle road operation scenario in a certain city, with a sampling rate of 1 Hz. The data dimensions include multiple source parameters such as timestamp, global positioning system (GPS) speed measurement value, geographic longitude and latitude coordinates, and instantaneous fuel consumption rate. In the actual data collection process, due to the combined influence of factors such as complex driving environment, electromagnetic signal interference, and urban building occlusion, the original sensor data generally has significant noise pollution, which manifests as multiple problems such as data distortion, abnormal value fluctuations, and signal interference [7]. Therefore, the first step in data processing is to preprocess the original data using wavelet decomposition and reconstruction [8]. The basic idea is to remove the wavelet coefficients corresponding to each frequency band and noise while retaining the wavelet coefficients of the original signal, and then perform wavelet reconstruction on the processed coefficients to obtain a pure signal.

Assume that a noisy signal can be described as:

$$S(x) = f(x) + n_1(x) \times n_2(x) \tag{1}$$

Among them, $S(x)$ is the degraded signal, $f(x)$ is the original signal, $n_1(x)$ is the additive noise, and $n_2(x)$ is the multiplicative noise.

The denoising process based on wavelet decomposition and reconstruction is described as follows:

Step 1: Decompose the noisy signal $f(x)$ into approximate component $c_{j,k}$ and detailed component $d_{j,k}$ by wavelet decomposition.

Step 2: According to the threshold δ_j, use equation (2) to process the detailed component $d_{j,k}$ of the layer j.

$$d_{j,k} = \begin{cases} d_{j,k}, & |d_{j,k}| > \delta_i \\ 0, & |d_{j,k}| \le \delta_i \end{cases} \tag{2}$$

Step 3: Use the reconstruction algorithm to

reconstruct the approximate component $c_{j,k}$ and the detailed component $d_{j,k}$ to obtain the filtered signal.

The original data and the data preprocessed by wavelet decomposition and reconstruction are shown in Fig. 1.

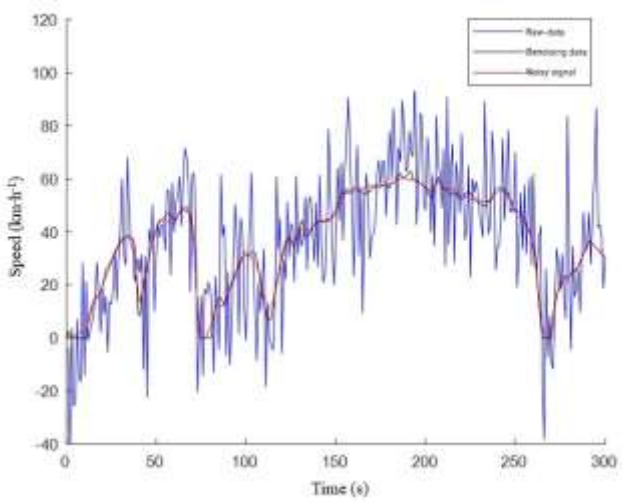

Figure 1. Comparison of noise reduction data results

The comparison between the original data and the preprocessed data shows that the wavelet decomposition and reconstruction method has a good denoising effect and can effectively improve the signal-to-noise ratio of the signal.

III. Analysis of Driving Condition Data

A. Feature Parameter Extraction and Kinematic Segment Division

Based on the analysis of relevant data and related research, 12 characteristic parameters were defined to describe the kinematic segments [8]. This paper selects 12 characteristic parameters including running time T/s, driving distance S/km, average speed $V_a/(km\cdot h^{-1})$, average driving speed $V_d/(km\cdot h^{-1})$, idling time ratio $T_i/\%$, acceleration time ratio $T_a/\%$, deceleration time ratio $T_d/\%$ cruising time ratio $T_c/\%$, speed standard deviation $V_{std}/(km\cdot h^{-1})$, average acceleration $a_a/(m\cdot s^{-2})$, acceleration standard deviation $a_{std}/(km\cdot h^{-1})$, and average deceleration $a_d/(m\cdot s^{-2})$.

The interval from the start of one idle speed to the start of the next idle speed of the car is called a kinematic segment [9]. This paper uses Python language to process and segment 1,655 kinematic segments from 195,815 pre-processed data.

B. Improved principal component analysis

Although the classical principal component analysis can effectively eliminate the differences in dimensions and magnitudes between the original variables when standardizing data, this process may also cause the characteristic differences of different indicators to be over-smoothed, resulting in potential information loss [10]. In view of the above situation, the improved principal component analysis method is as follows:

Step 1. Improve the traditional principal component dimensionless method by using the indicator mean method and indicator homogeneity method [11]. Assume that there are m objects and n indicators in the overall evaluation, and the initial indicators can form a matrix $X_{ij}=(x_{ij})_{m\times m}$. To average the matrix is to divide the original index by the average value of all indexes Y_{ij}:

$$Y_{ij}=x_{ij}/\bar{x}_j,(i=1,2,\cdots,n;\ j=1,2,\cdots,p) \quad (3)$$

Among this:

$$\bar{x}_j=\frac{1}{n}\sum x_{ij},(j=1,2,\cdots,p) \quad (4)$$

The index can be processed to make all indicators have the same effect on the whole in the same direction. In the series, let y_{ij} be the reverse index, $\min_{1\le i\le m}\{y_{ij}\}$ is the smallest number among them, and the index is processed to be:

$$y_{ij}^{'}=y_{ij}-\min_{1\le i\le m}\{y_{ij}\},(i=1,2,\cdots,m;\ j=1,2,\cdots,n) \quad (5)$$

Among them, $y_{ij}^{'}$ is the sequence after y_{ij} is homogenized, Such changes will not change the

distribution of the original indicators. The improved principal component can represent more characteristic parameter information and achieve dimensionality reduction of driving conditions.

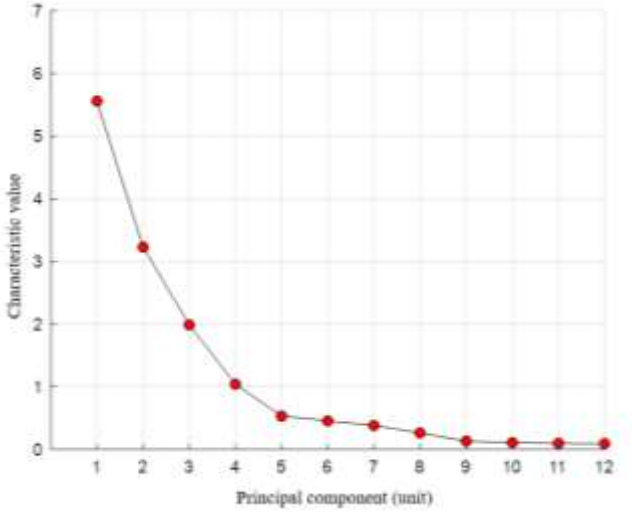

Figure 2. Lithotripsy

Fig. 2 shows that there are obvious inflection points in the variation curves of each principal component, and it is concluded from this observation that the first three principal components are selected.

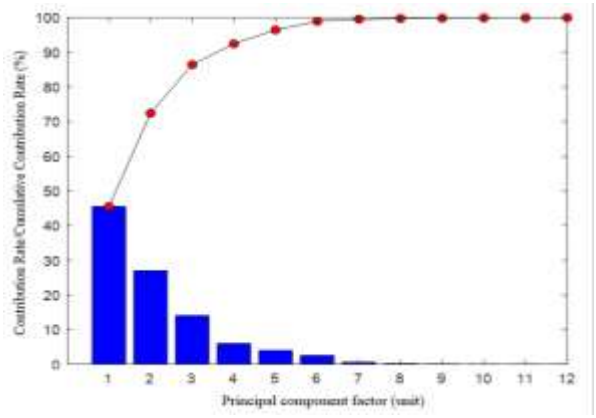

Figure 3. Contribution rate and cumulative contribution rate

As shown in Fig. 3, the cumulative contribution rate of the first three principal components has reached 85%, which basically represents all the information of the 12 characteristic parameters of the fragment, and can be used for cluster analysis. The first principal component contains 45% of the information, thus meeting the requirement of fewer principal components representing more information.

TABLE I. PRINCIPAL COMPONENT LOAD MATRIX

Characteristic parameters	M_1	M_2	M_3
Deceleration time ratio T_d/%	0.323	0.351	-0.223
Driving Clustering S /km	0.893	0.234	0.065
Run time T /s	0.782	0.251	-0.342
Acceleration time ratio T_a/%	0.396	-0.186	0.061
Cruise time ratio T_c/%	0.641	0.335	-0.075
Average speed $V_a/(km \cdot h^{-1})$	0.499	0.763	0.125
Deceleration time ratio $V_d/(km \cdot h^{-1})$	0.778	0.415	0.132
Speed standard deviation $V_{std}/(km \cdot h^{-1})$	0.498	0.333	0.054
Acceleration standard deviation $a_{std}/(km \cdot h^{-1})$	0.125	0.267	-0.077
Average acceleration $a_a/(m \cdot s^{-2})$	0.024	0.523	0.053
Average deceleration $a_d/(m \cdot s^{-2})$	0.266	-0.433	-0.059
Idle time ratio $T_i/(m \cdot s^{-2})$	0.165	-0.351	0.853

The larger the absolute value of the parameter principal component load coefficient, the higher the correlation coefficient between a parameter and a principal component, and the larger the contribution factor. According to Table 1 above, the eigenvalues of the first principal component are driving distance, segment duration, cruising time ratio, and average driving speed; the eigenvalues of the second principal component are average speed; and the eigenvalues of the third principal component are idling time ratio. From the first three principal components, it can be seen that the 12 characteristic parameter matrices of the sample are reduced to 6 characteristic parameter matrices that can represent most of the sample information.

C. Improved K-means cluster analysis

The K-means algorithm is sensitive to the selection of initial cluster centers. Since the process uses a random mechanism, the initial centroids it selects may be distributed in data sparse areas or coincide with outliers. This non-ideal initial state can easily cause the algorithm to fall into a local optimal solution, thereby reducing the clustering quality [12]. In the usual optimization method, in order to make the initial cluster center better than the method of randomly selecting cluster centers in traditional algorithms, k data objects with the farthest distance or the largest density are generally selected as the initial cluster centers. However, if there is noise data in the data set, the "distance optimization method" is likely to use the noise data as the initial cluster center. The "density method" selects the k data objects with the largest density as the initial clustering centers. This method can remove isolated points of data, but it is not suitable

for non-convex data sets. This paper proposes a method that combines the "distance optimization method" and the "density method" to determine the optimal initial clustering center, and constructs a data set density measurement method.

1) Relevant definitions

a) The Euclidean distance between two points in space is:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2} \quad (6)$$

Among them, x_i , x_j are two m-dimensional data points.

b) Average distance between data objects:

$$MeanDist = \frac{1}{C_n^2} \sum d(x_i, x_j) \quad (7)$$

n is the number of data points in the data cluster, and C_n^2 is the number of logarithms obtained from n data points.

c) Given a data set $D = \{x_1, x_2, \cdots, x_n\}$ *, the density measurement function of data point* x_i *is:*

$$Dean(x_i) = \sum_{j=1}^{n} u(MeanDist - d(x_i, x_j)) \quad (8)$$

Among them, the $u(z)$ function is expressed as:

$$u(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

The density parameter of data point x_i is actually a data object inside a circle with center x_i and radius $MeanDist$.

d) The average density measurement function of the data set is defined as:

$$MeanDens(D) = \frac{1}{n} \sum_{i=1}^{n} Dens(x_i) \quad (9)$$

n is the number of data objects in the dataset a D .

e) For data point x_i *in data set* D *, if*

$$Dens(x_i) < \alpha \times MeanDens(D) \quad (10)$$

Point x_i is called an isolated point, where $0 < \alpha < 1$.

f) The distance between data object x_i *and data set* C *is the closest distance to all data points in data set* C .

$$d(x_i, C) = \min(d(x_i, x_j), x_j \in C) \quad (11)$$

2) Algorithm Idea

The improved K-means clustering process is as follows: first evaluate the density distribution function of all samples, identify and remove outliers, and then construct a high-density data subset. Then select the sample with the best density value as the first initial cluster center, and then select the sample points with the farthest distance from the previous center in the remaining high-density data as new cluster centers, until k initial centroids are established. Finally, the standard K-means clustering process is executed based on the optimized centroid configuration.

The algorithm is described as follows:

Input: Sample dataset $D = \{d_1, d_2, \cdots, d_n\}$ containing n data objects

Output: optimal k value and clustering results.

Step 1: Use $d(x_i, x_j)$ and $MeanDist$ to calculate the distance and average distance between any two data objects in data set D .

Step 2: Use $Dens(x_i)$ and $MeanDens(D)$ to calculate the density measurement function of all data objects in data set D and the average density measurement function of data set D .

Step 3: According to formula (10), determine the isolated data objects and delete them from set D to obtain set A with high density parameters.

Step 4: Select a data object with the highest

parameter density from set A as the first initial cluster center, add it to set B, and remove it from set A.

Step 5: From set A, select the data object farthest from set B as the next initial cluster center, add it to set B, and remove it from set A.

Step 6: Repeat Step 5 until the number of data objects in set B is k.

Step 7 uses traditional K-means for clustering based on k cluster centers.

3) Results Analysis

CH is used as an evaluation index to determine the optimal K value before clustering. It is a measure based on the intra-class dispersion matrix and inter-class dispersion matrix of all samples. The larger the CH is, the tighter the clusters are and the more dispersed the classes are. In this case, the clustering result is relatively better [13]. The index is defined as:

$$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)} \tag{12}$$

Where n is the number of clusters, k is the current class, $trB(k)$ is the trace of the between-class dispersion matrix, and $trW(k)$ is the trace of the within-class dispersion matrix.

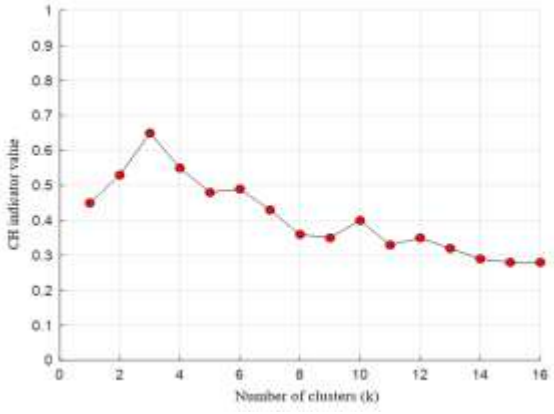

Figure 4. Relationship between cluster number and CH index

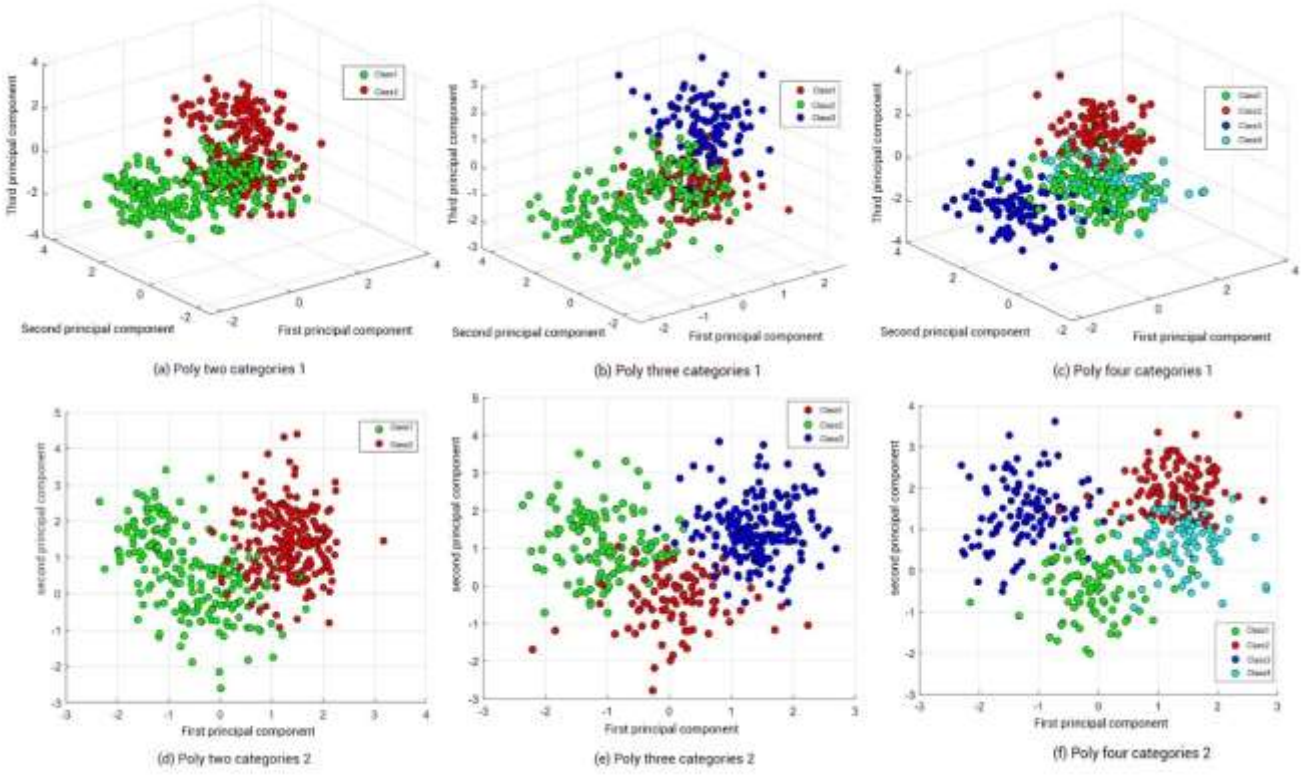

Figure 5. K-means clustering results

In order to determine the appropriate number of clusters, this paper first clusters into 2, 3, and 4 categories, and the clustering results are shown in Fig. 5. At this time, the clustering k value cannot be clearly determined. At this time, CH can be used as an evaluation indicator. The processing results of CH values under different clustering states are shown in Fig. 4. It can be observed that 3 categories are clustered when the CH value is the largest.

Reference [14] used two typical driving condition characteristic parameters, average vehicle speed and idling time ratio, to conduct clustering research. This study innovatively focused on the cruising time ratio and average driving speed, which have a higher contribution, as the core analysis dimensions. The original data points shown in Fig. 6 are scattered point distribution, in which the red marked area clearly circles the isolated samples and outliers that significantly deviate from the main distribution trend.

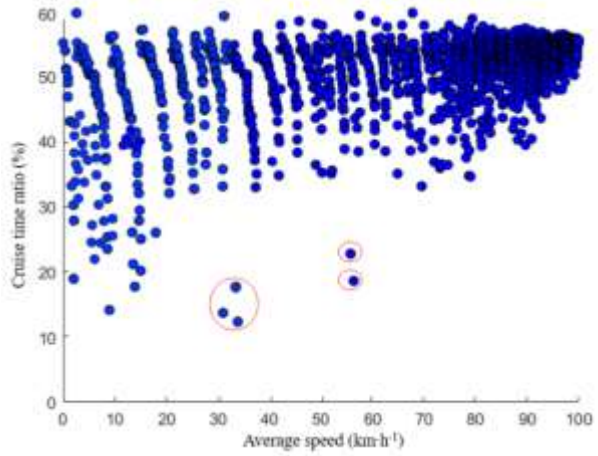

Figure 6. Scatter plot of segments in two-dimensional feature space

This paper clusters the kinematic segments into three categories. As shown in Fig. 7, the cluster centers of the first, second and third categories are (14, 38), (52, 45) and (86, 54) respectively, considering the general urban traffic conditions: the first category is the relatively crowded urban area, with relatively low average speed and cruising time, and more frequent starts and stops; the second category is the relatively unobstructed urban suburbs, with relatively high average speed and cruising time, and fewer starts and stops; the third category is the unobstructed high-speed segment, with high average speed and cruising time, and fewer starts and stops.

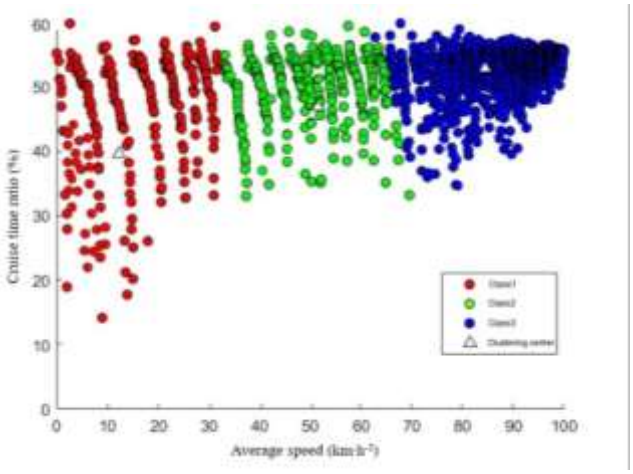

Figure 7. Clustering results of fragments in two-dimensional feature space

IV. Driving Condition Construction and Analysis

A. Working condition construction

According to the proportion of the total time of each time segment to the driving conditions of all data sets, the time taken by each segment in the final constructed condition can be calculated. As shown in Fig. 8, this paper constructs it according to the time of 1,200s of the general typical driving condition.

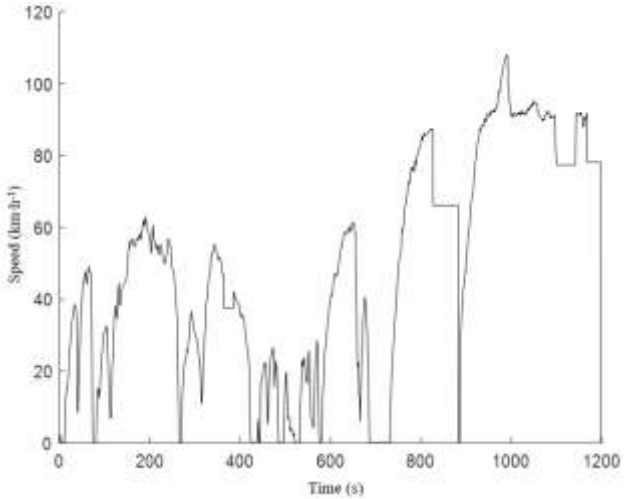

Figure 8. Synthetic driving conditions

As can be seen from Fig. 9, most of the operating points are concentrated in the medium and low speed range, and the distribution of acceleration is relatively reasonable, which can show the actual acceleration and deceleration of the car.

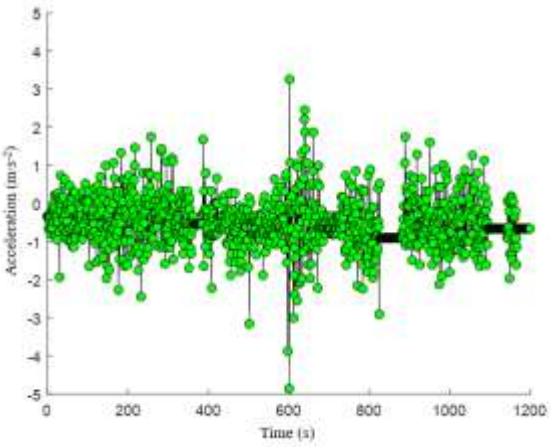

Figure 9. Time and acceleration diagram

As can be seen from Fig. 10, the acceleration is mainly distributed in the speed range of 0-40 $km \cdot h^{-1}$ and around 80 $km \cdot h^{-1}$. During low speed and high acceleration, the instantaneous fuel consumption has a significant bulge, which may be caused by the driver's improper operation.

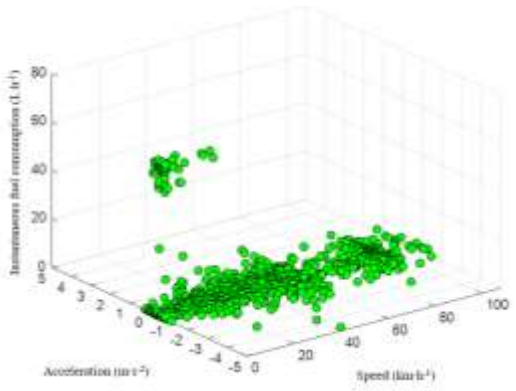

Figure 10. Scatter plot of instantaneous fuel consumption of speed and acceleration

B. Working condition verification and fuel consumption analysis

The smaller the distribution difference value $SAFD_{diff}$ is, the higher the commonality between the constructed working condition and the actual data is [15].

$$SAFD_{diff} = \frac{\sum_i (SAFD_{cycle}(i) - SAFD_{data}(i))^2}{\sum_i SAFD_{data}(i)^2} \tag{13}$$

$SAFD_{cycle}$ is the $SAFD$ of a cycle, and $SAFD_{data}$ is the $SAFD$ of all data.

TABLE II. COMPARISON OF THE METHOD IN THIS PAPER AND TRADITIONAL K-MEANS RESULTS

Method	Eigenvalue mean relative error /%	Cluster average accuracy /%	Average time /s	SAFD^{diff}/%
Traditional K-means clustering	8.1	93	215	2.12
Clustering of this article	4.6	98	127	1.17

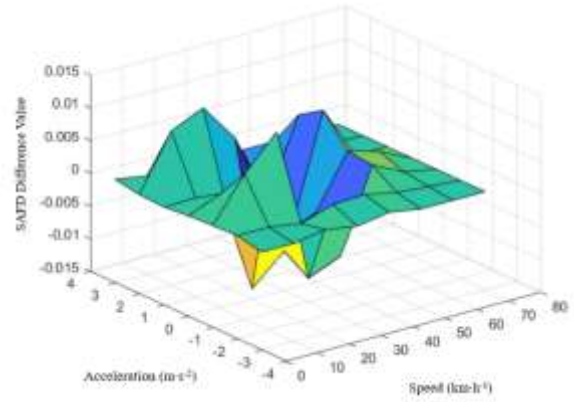

Figure 11. SAFD difference between experimental data and synthetic conditions

The velocity-acceleration joint distribution is to separate the velocity values and acceleration values into equal intervals and further calculate the proportion of the working condition data in different intervals [16]. As shown in Fig. 11, the combined speed-acceleration difference between the original data and the constructed driving condition is distributed within the range of $\pm 1.2\%$, and the calculated distribution difference value ($SAFD_{diff}$) is 1.17%. Therefore, the driving condition constructed in this paper meets the driving characteristics of light vehicles and has strong applicability.

V. CONCLUSIONS

In view of the inherent defects of information distortion in the dimensionless processing of traditional principal component analysis, this study proposes a dual optimization strategy of "mean preservation-trend synchronization" to achieve standardized improvement while maintaining the original difference characteristics of the variables. In view of the limitation of the K-means clustering algorithm being sensitive to the initial center, an intelligent optimization mechanism for initial clustering centers based on local density distribution is constructed. The statistically representative initial centroids are selected by the density measurement value of data objects, which effectively eliminates noise interference and improves clustering stability. This paper constructs an optimization method of improved principal component and improved K-means combination to synthesize automobile driving conditions. The verification results show that the difference rate between the synthetic conditions generated by the proposed method and the original data in the joint distribution space of speed-acceleration is only 1.17%, and the reconstruction accuracy of the conditions reaches 98.83%. The driving conditions synthesized by the proposed method are significantly better than those of traditional methods and are closer to actual traffic conditions.

REFERENCES

[1] YU Yefeng, ZHANG Chen, GAO Zhan, et al. Optimization of Driving Cycle Development Based on Multi-Objective Genetic Algorithm [J]. Chinese Internal Combustion Engine Engineering, 2023, 44(05): 57-65.

[2] HAN Rui, SHI Pengwei, DING Qingguo, et al. Construction of driving conditions for light vehicles on urban roads in Harbin [J]. Technology & Economy in Areas of Communications, 2025, 27(01): 50-58.

[3] Zhao X, Yu Q, Ma J, et al. Development of a Representative EV Urban Driving Cycle Based on ak-Means and SVM Hybrid Clustering Algorithm [J]. Journal of Advanced Transportation, 2018, 2018(1): 1890753.

[4] Ma Fumin, Lu Ruiqiang, Zhang Tengfei. Rough K-means clustering algorithm based on local density adaptive metric [J]. Computer Engineering and Science, 2018, 40(01): 184-190.

[5] Yuan Yiming, Liu Hongzhi, Li Haisheng. Improved K-Means text clustering algorithm based on density peak and its parallelization [J]. Journal of Wuhan University (Science Edition), 2019, 65(05): 457-464.

[6] Bao Zhiqiang, Zhao Yuanyuan, Hu Xiaotian, et al. A new K-Means clustering algorithm that is not sensitive to outliers[J]. Modern Electronic Technology, 2020, 43(05): 109-112.

[7] Ding Yifeng, Li Jun, Gai Hongchao, et al. Application of Wavelet Transform to Vehicle Speed Data Processing for Construction of Driving Conditions [J]. Science Technology and Engineering, 2017, 17(28): 274-279.

[8] Zeng Xiaorong, Kong Lingwen, Yang Xueyi, et al. Application of Principal Component Analysis to Vehicle Driving Conditions [J]. Automobile Practical Technology. 2014(05): 5-9.

[9] Peng Yuhui, Zhuang Yuan. Construction Method of Urban Sanitation Vehicle Driving Conditions Based on Combination Optimized Clustering and Markov Chain[J]. Journal of Fuzhou University (Natural Science Edition). 2019, 47(04): 502-508.

[10] Shang Liqun, Wang Shoupeng. Application of Improved Principal Component Analysis Method in Comprehensive Evaluation of Thermal Power Units [J]. Power System Technology, 2014, 38(07):1928-1933.

[11] Fang Rui. Research on Wuhan City Competitiveness Based on Improved Principal Component Analysis[D]. Huazhong University of Science and Technology, 2012.

[12] Yanling D, Qun L, Shuyin X. An improved initialization center k-means clustering algorithm based on distance and density [C]//American Institute of Physics Conference Series. 2018, 1955(1): 40-46.

[13] Zhang Yuanxiang. Research on the method of determining the optimal number of clusters in cluster analysis [D]. Anhui University, 2020.

[14] Fotouhi A, Montazeri-Gh M. Tehran driving cycle development using the k-means clustering method[J]. Scientia Iranica, 2013, 20(2): 286-293.

[15] Nguyen Y L T, Nghiem T D, Le A T, et al. Development of the typical driving cycle for buses in Hanoi, Vietnam [J]. Journal of the Air & Waste Management Association, 2019, 69(4): 423-437.

[16] Ye Chenchen, Zhang Hongkun, Fan Luyan, et al. Experimental Research on Urban Road Conditions of Passenger Cars in Shenyang City [J]. Science Technology and Engineering, 2017, 17(21): 241-247.