Research on Driving Conditions and Fuel Consumption of Improved K-means Clustering Algorithm

Shuping Xu
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, 710032, China
E-mail: 563937848@qq.com

Xuanlv Wei
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, 710032, China

Leyi Wang
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, 710032, China
E-mail: 634877232@qq.com

Xiaodun Xiong
School of Computer Science & Engineering
Xi'an Technological University
Xi'an, 710032, China

***Abstract*—In order to solve the problem that the initial center of traditional clustering algorithm is easy to fall into local optimum and time-consuming. An improved combination optimization algorithm of principal component analysis and weighted K-means clustering is proposed. The algorithm introduces the maximum and minimum distance, weighted Euclidean distance, starting from the mean sum of the distances of the remaining clustering points, avoiding the influence of outliers and edge data. The proportion method is used to improve the principal component, and the characteristic influence factor obtained is used as the initial characteristic weight to construct a weighted Euclidean distance metric. According to the influence factors of feature contribution rate on clustering, a clustering method of feature weight influence factors is proposed. The representative feature factors are selected to highlight the clustering effect. Finally, the driving cycle of automobile is synthesized and the instantaneous fuel consumption is analyzed. The results show that: the difference value of speed acceleration joint distribution of the proposed method is only 1.05%, which saves 44.2% of the time compared with the traditional K-means clustering, and the driving cycle fitting degree is high, which can reflect the actual vehicle operation characteristics and fuel consumption.**

Keywords-Driving Cycle; Influence Factors; Feature Weight; Weighted K-Means Clustering

I. INTRODUCTION

The driving condition of a car is also called the operating cycle, which is the speed-time variation law of a vehicle in a specific environment. It is mainly used to evaluate vehicle pollutant emissions and energy consumption, and is of great value to the research and development of new vehicle models and risk assessment of traffic control [1]. Many scholars have conducted research on it, and Nguyen et al. [2] proposed a driving cycle construction process based on

DOI: 10.2478/ijanmc-2022-0011

Markov chain theory. Ding Yifeng et al. [3] used multivariate statistical methods such as principal component and cluster analysis to construct automobile road conditions. Liu Yingji et al. [4] used the characteristics of kinematics segment connection fuzzy to construct working conditions by combining principal components and fuzzy C-means clustering. Most scholars' research on driving cycle mainly focuses on the selection of K-means clustering initial center and single improved k-means clustering algorithm, but lack of research on principal component analysis and clustering combination optimization and execution time consumption. In order to achieve the ideal clustering effect and time consumption, it is still necessary to focus on the improvement of K-means clustering. Zhang Rui et al. [5] proposed OICCK-means algorithm in order to make up for the deficiency that the clustering effect of traditional K-means algorithm depends heavily on the initial clustering center. Zhang Lin et al. [6] adopted the idea of density to overcome the sensitive defect of traditional initial center. Luo Junfeng et al. [7] introduced information entropy and weighted distance to remove outliers. Zhang Yan [8] proposed an improved rough K-means clustering algorithm based on density weighting, which not only improves the clustering accuracy and reduces the number of iterations, but also weakens the interference of noise data and outliers on the results. However, the algorithm improves the clustering accuracy at the expense of efficiency cost. The algorithm puts most of the time consumption on the density of data objects, and the time complexity is too high.

Through the above analysis, this paper proposes an improved principal component analysis and improved K-means clustering combination optimization method, introduces the maximum and minimum clustering method and weighted Euclidean distance, and increases the weight of clustering eigenvalues according to the contribution factor. The results show that the clustering effect is stable, the time consumption is low, and the driving cycle constructed has strong applicability and meets the characteristics of traffic conditions.

II. ANALYSIS OF DRIVING CYCLE DATA

A. Data preprocessing

The data collected in this paper are the actual road driving conditions of a city light vehicle in September 2019 (sampling frequency is 1Hz), among which, the data information includes time, GPS speed measurement, longitude and latitude, instantaneous fuel consumption, etc. Using fitting interpolation method to interpolate and fit the disturbed discontinuous data, wavelet decomposition and reconstruction method to smooth the contaminated data [9] the original data was reduced from 194511 to 164039 by Matlab preprocessing

B. Feature parameter extraction and kinematic segmentation

Based on the analysis of relevant data and related research, 12 characteristic parameters are defined to describe the kinematic segments [10]. In this paper, 12 characteristic parameters including segment duration/T, travel distance/S, average speed/V_a, average driving speed/V_d, idle time ratio/ T_i, acceleration time ratio/ T_a, deceleration time ratio/T_d, cruise time ratio/T_c, speed standard deviation/V_{std}, average acceleration /a_a, average standard deviation of acceleration / a_{std}, average deceleration/a_d etc.

The interval from the start of one idling speed to the beginning of the next idling speed is called the kinematic segment [11]. This paper uses Python to develop related programs, uses stack and loop traversal data for processing, and divides

2445 kinematics segments from 164039 preprocessed data.

III. IMPROVED PRINCIPAL COMPONENT ANALYSIS

The traditional principal component uses linear technology to reduce the dimension of data, which eliminates the influence of order of magnitude and the difference information of each characteristic factor. In real life, the relationship between data is often nonlinear.

The comprehensive evaluation method with variance contribution rate as the weight can not reasonably explain the analysis results, and even the evaluation results deviate greatly from the facts [12]. Therefore, using the specific gravity method proposed in reference [13], the improved principal component can not only eliminate the dimension noise, but also can represent more feature parameter information and realize dimension reduction. The formula is as follows:

$$ZX_i = x_i / \sqrt{\sum_{i=1}^{n} x_{ij}^{\ 2}} \qquad (1)$$

In the case of dimension reduction, the improved principal component forms a matrix with the obtained number of data samples $(n)\times$ characteristic parameters (p), and select the principal component whose cumulative contribution rate reaches more than 80% for reduction and de-correlation. It can be seen from Figure 1 that the cumulative contribution rate of the first four principal components has reached 82.76%, which basically represents all the information of the 12 characteristic parameters of the fragment.

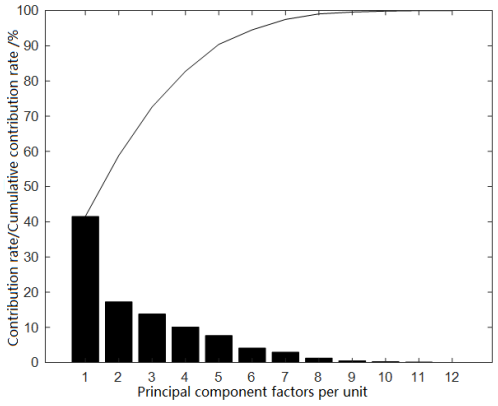

Figure 1. Contribution rate and cumulative contribution rate

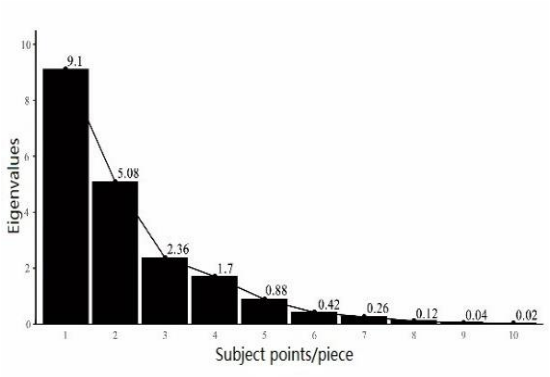

Figure 2. Gravel map

It can be seen from Figure 2 that each principal component is gradually decreasing, and there is an obvious inflection point in the change curve. It can be seen from Figure 1 that the first principal component contains 41.5% information in the improved principal component analysis results, so it meets the requirement that less principal components represent more information.

TABLE I. PRINCIPAL COMPONENT LOADING MATRIX

Characteristic parameter	M_1	M_2	M_3	M_4
Deceleration time ratio T_d	0.423	0.341	-0.723	0.248
Distance traveled S	0.893	0.134	0.045	0.432
Fragment duration T	0.432	0.231	-0.142	0.768
Acceleration time ratio T_a	0.394	-0.156	0.060	0.491
Cruise time ratio T_c	0.341	0.835	-0.045	-0.138

Average velocity V_a	0.499	0.763	0.025	0.255
Average driving speed V_d	0.778	0.315	0.112	0.358
Speed standard deviation V_{std}	0.198	0.033	0.034	0.189
Accelerate standard deviation a_{std}	0.145	0.267	-0.067	-0.121
Average acceleration a_a	0.014	0.223	0.033	0.024
Average deceleration a_d	0.566	-0.433	-0.052	0.315
Idle Time Ratio T_i	0.125	-0.351	0.843	0.467

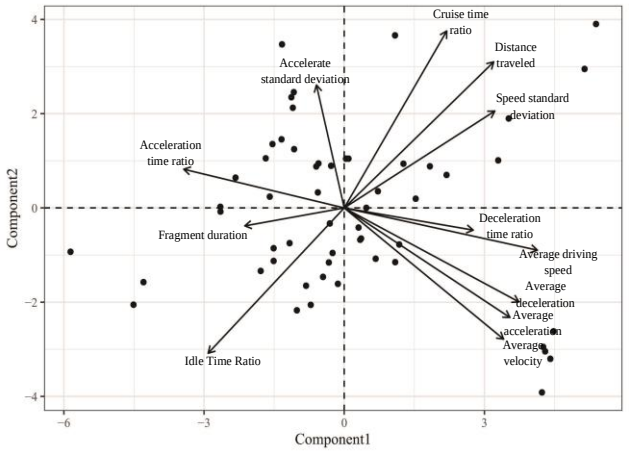

Figure 3. Principal component analysis scatter plot

When the absolute value of the principal component load factor of the selected parameter is larger, the correlation coefficient between a parameter and a principal component is higher [14]. From Figure 3, we can see the correlation of each eigenvalue directly. According to the above table 1, the first principal component eigenvalues are driving distance, average deceleration and average driving speed, and the correlation coefficients are 3.15, 2.08 and 3.69, respectively, so they have great correlation with driving distance and average driving speed; The second principal component eigenvalues have the average speed and cruise time ratio, and the correlation coefficients are 2.75 and 3.84 respectively, so they have a greater correlation with the cruise time ratio; the third principal component eigenvalues have the idle time ratio and deceleration time ratio, and the correlation coefficients are 3.06 and 2.85 respectively, so they have a greater correlation with the idle time ratio; The fourth principal component eigenvalue has fragment duration, and the correlation coefficient is 2.43, which indicates that it has a strong correlation with fragment duration. Through the analysis of IPCA, the first four principal components can reflect the characteristics of the original segment, and the 12 characteristic parameter matrices of the population sample are compressed into one eight characteristic parameter matrix which can represent the vast majority of sample information.

IV. IMPROVED K-MEANS CLUSTERING ANALYSIS

A. Outlier processing

The actual test species will have more or less interference, which often produces outliers or noises, which will affect the clustering effect. Here, we construct a residual point distance mean sum method to eliminate the influence of noise and outliers [15]. For the i point in the data, the sum of distances between each point and other points is S_i, and the sum of distances is H .When $S_i > H$, point i is regarded as an isolated point. Among them, the sample data is, the data dimension is, and the calculation is as follows:

$$S_i = \sum_{j-1}^{n} \sqrt{\sum_{h-1}^{d} (x_{ih} - x_{jh})^2} \tag{2}$$

$$H = \sum_{i-1}^{n} \frac{S_i}{n} \tag{3}$$

B. Maximum and minimum distance

1) The maximum and minimum distance of the remaining data in the cluster and dataset is defined as:

$$D_{\max} = \mathrm{Max}(d) \tag{4}$$

Among them, d is a set consisting of the minimum value of the distance between each cluster and the remaining data in the data set.

2) d_k is the minimum value of the distance between each cluster and the remaining data in the data set,

$$d_k = \mathrm{Min}(\sum_{k=1}^{m}(X_{ik} - X_{jk})^2) \tag{5}$$

Among them, X_i is the cluster center, X_j is the remaining data in the data set, and m is the dimension of the data.

3) Determine whether to select the initial candidate center as the optimized candidate center.

$$\mathrm{Max}(\mathrm{Min}(\mathrm{D}_t)) > \theta \left\| v_1 - v_2 \right\| \tag{6}$$

Among them, v_1, v_2 are the points that first become the candidate centers after optimization, θ is the parameter, which can be 0.5.

4) The criterion function of the $K-means$ algorithm for clustering is the error sum of square criterion function.

$$J_c = \sum_{i=1}^{k}\sum_{p\in C_i}(\left\| P - M_i \right\|)^2 \tag{7}$$

Among them, M_i is the mean value of all data in class C_i, p is each data in class C_i, and J_c is the function of sample and cluster center.

C. Weighted Euclidean distance

$\omega = [\omega_1, \omega_2, \cdots, \omega_n]^T \in R^{n\times d}$, The weight ω is introduced to distinguish the relationship between the sample data and the cluster center,

$$\sigma d_{\omega}(x_j, c_i) = \sqrt{\sum_{m=1}^{d}\omega_{jm}(x_{jm} - c_{im})^2} \tag{8}$$

$$\omega_{jm} = \frac{x_{jm}}{\frac{1}{n}\sum_{j=1}^{n}x_{jm}} \tag{9}$$

The initial new weight is as follows:

$$\mathrm{W_{iNew}} = \mathrm{W_i}(1 + \frac{\mathrm{A_{init}} - \mathrm{A_i}}{\mathrm{A_{init}}}) \tag{10}$$

Among them, the clustering accuracy is

$$\mathrm{A_i} = \frac{N_{cor}}{N}\% \tag{11}$$

Among them, $\omega_j = (\omega_{j1}, \omega_{j2}, \cdots, \omega_{jd})^T$ is d dimensional vector, x_{jm} is the m component of the j sample, $\frac{1}{n}\sum_{j=1}^{n}x_{jm}$ the average of the sum of the m component of each data object in the sample data set. It can be seen that ω is a weight that can reflect the overall distribution characteristics of the sample Value [5].

D. Feature weighted K-means clustering algorithm

1) By processing the noise and outliers, a new data set is obtained, and the related feature list is obtained.

2) The improved principal component analysis calculates the contribution factor of each feature to obtain the initial weight.

$$W=(W_{1X_1},W_{2X_2},\cdots,W_{nXn}) \quad (12)$$

3) The maximum minimum distance multi center clustering algorithm iteratively implements the proposed clustering center selection method to determine k *initial clustering centers.*

4) Based on the weighted features and the initial clustering center, K-means is executed to obtain K *clusters.*

5) Calculate the initial clustering accuracy A_{init}

6) For each feature in ω*, K-means clustering without the feature is performed, and the clustering accuracy is calculated;If* $A_i<A_{init}$ *, increase its weight* W_{iNew} *; otherwise remove the feature.*

7) Normalize the weights, perform K-Means clustering based on the new weights, and calculate the clustering accuracy A_{init}*;*

If $A_{final}>A_{init}$, accept the new weight and set $A_{init}=A_{final}$; otherwise, keep the old weight unchanged.

According to the above working condition data, the improved K-means algorithm is used for processing. First, edge data and outliers are detected, and abnormal points are eliminated. As shown in Figure 4 below, cluster 1 is a normal clustered point. Cluster 2 is the outlier of edge data. As can be seen in Figure 5, the edge data is relatively distant from most normal points, and most of the edge data are outliers, which can be eliminated.

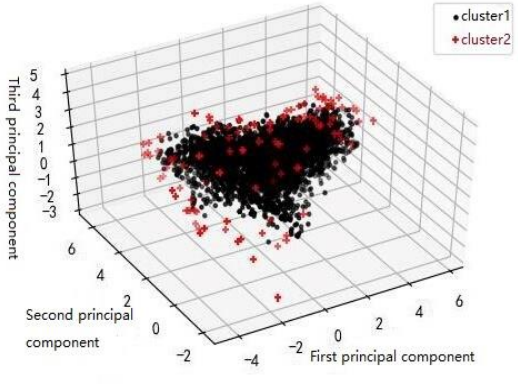

Figure 4. Scatter plot of edge data points of working conditions

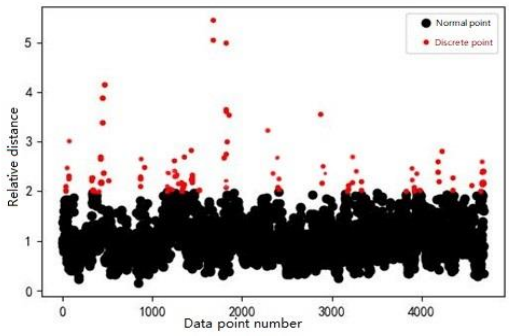

Figure 5. Relative distance comparison of outliers

According to the above-mentioned improved principal component analysis, the contribution factor and the characteristic value with high correlation are used to draw the three-dimensional graph, as shown in Figure 6. In this paper, the average speed, driving distance and cruise time ratio are selected to represent each point of clustering.

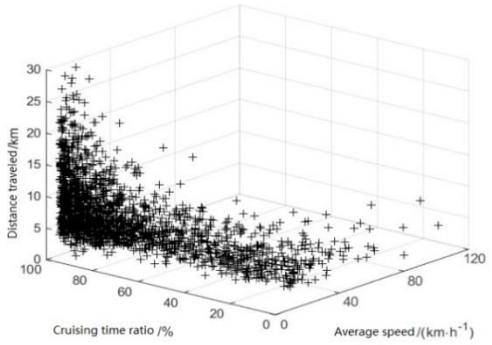

Figure 6. Three-dimensional scatter plot of working conditions

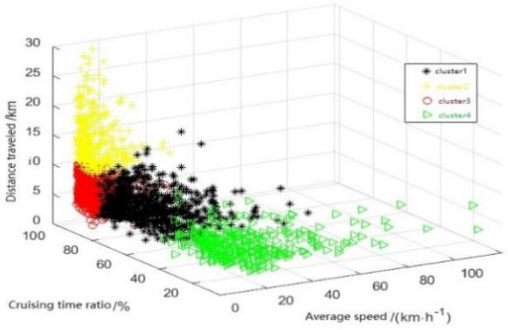

Figure 7. Working condition cluster analysis scatter plot

The improved K-Means clustering algorithm divides the kinematic segments into four categories, which are represented by cluster 1, cluster 2, cluster 3 and Cluster 4. It can be seen from Figure 7 that the first type is downtown area, where the vehicles start and stop frequently and the speed is low, and the average speed, cruise time ratio and driving distance are low; the second type is the living area, which is congested, with more start and stop times, and lower average speed, cruise time ratio and driving distance; the third type is suburban area, with smooth road conditions, less starting and stopping times, average speed, cruise time ratio and driving distance The fourth type is high-speed area, with smooth traffic, less start and stop times, high average speed, cruise time ratio and driving distance.

V. DRIVING CYCLE CONSTRUCTION AND FUEL CONSUMPTION ANALYSIS

A. Construction and verification of working conditions

According to the proportion of the total time of various time segments in the driving cycle of all data sets, the time taken by each driving cycle in the final construction cycle can be calculated [16]. This paper takes 1400s to construct vehicle driving cycle, as shown in Figure 8 below. The first type of low speed segment, the second type of medium speed segment, and the third type of medium high speed segment. The fourth type of high-speed video.

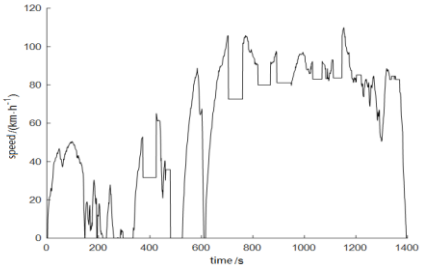

Figure 8. Synthetic driving conditions

From the speed and acceleration to verify the difference between the constructed driving cycle and the experimental data [11], this is a relatively standard verification method. Matlab software is used to calculate the speed acceleration joint distribution matrix of the vehicle driving cycle data, as shown in Figure 9.

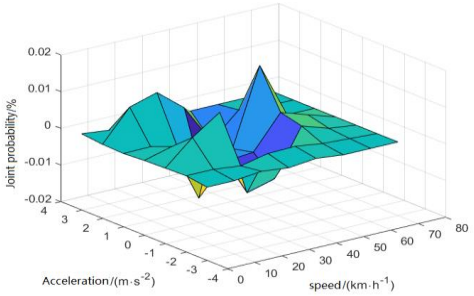

Figure 9. SAFD difference between experimental data and synthetic conditions

As can be seen from Figure 9 above, the joint velocity acceleration difference distribution of the experimental data and the improved clustering algorithm in this paper is within the ±1.2% range, and the calculated distribution difference value ($SAFD_{diff}$) is 1.05%，while the difference value ($SAFD_{diff}$) of the speed acceleration joint distribution between the experimental data and TKM is 0.97%. Therefore, the driving cycle constructed in this paper meets the driving characteristics of light vehicles, meets the development requirements of vehicle driving cycle construction, and has strong applicability.

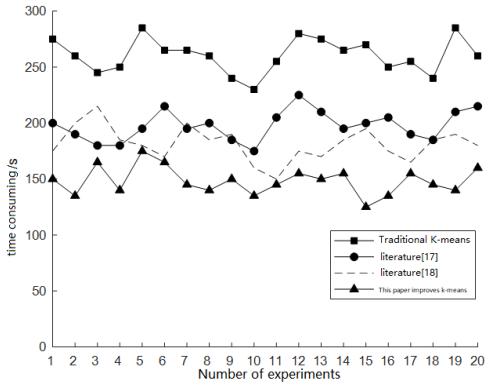

Figure 10. The results of the running time of the four methods

This paper uses the working condition construction method of literature [17] and literature [18]. According to the data of this paper, the improved principal component algorithm of this paper is combined with the four algorithms respectively, and 20 experiments are performed, as shown in Figure 10. The results show that, In this paper, the improved K-means clustering algorithm can not only weaken the influence of noise points on the initial center, but also greatly shorten the clustering time based on the stable clustering effect.

TABLE II. FOUR METHODS TO COMPARE THE RESULTS OF THE EXPERIMENT

Clustering method	The number of wrong samples	Average running time / s	Average accuracy /%	$SAFD_{diff}$ /%
k-means	184	260.5	89	1.98
Literature[17]	121	202.75	97	1.54
Literature[18]	98	181.5	99	1.25
The algorithm in this paper	101	145.25	98	1.05

The results of programming using Matlab are shown in Table 2 above. The comparison of the four algorithms in terms of the number of error-clustering samples, average running time, average correct rate and $SAFD_{diff}$, the improved K-means algorithm in this paper performs better in clustering performance and time consumption. The average running time is 44.2% less than that of traditional K-means clustering.

B. Fuel consumption analysis

As shown in figures 11 and 12, the instantaneous fuel consumption is large at low speed, medium and low speed, the torque fluctuation in the region is larger than that in the high speed region, the instantaneous fuel consumption rate in the high speed region is relatively stable, and the instantaneous fuel consumption rate in the low speed region and medium speed region is obviously increased. It can be observed in Figure 13 that the instantaneous fuel consumption increases briefly at low speed, and then the fluctuation trend is roughly consistent with the driving speed. As can be seen from Figure 14, the engine speed is mainly distributed in 1500-2500r / min under driving condition, and the opening of accelerator pedal is concentrated in 0.12-0.18, indicating that the driving condition is in medium high speed state.

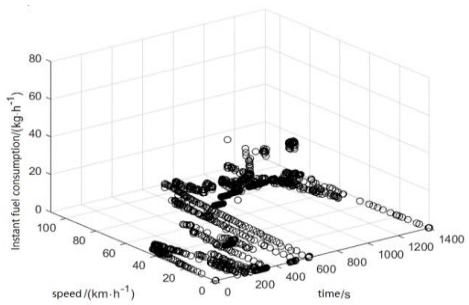

Figure 11. The relationship between driving time and speed instant fuel consumption

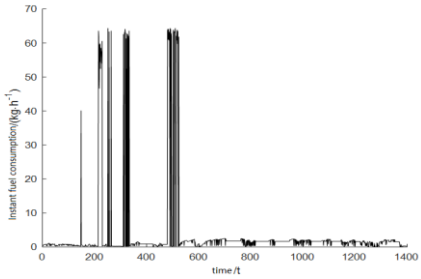

Figure 12. Relationship between driving time and instantaneous fuel consumption

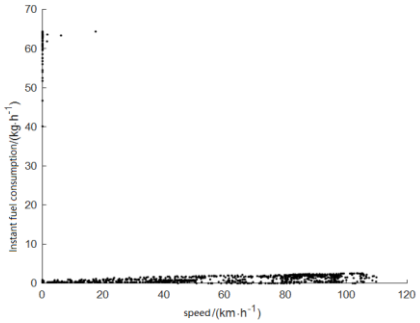

Figure 13. The relationship between driving speed and instantaneous fuel consumption

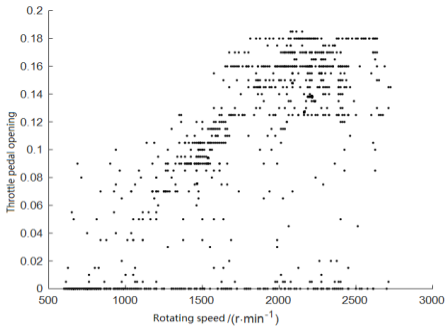

Figure 14. The relationship between driving speed and accelerator pedal opening

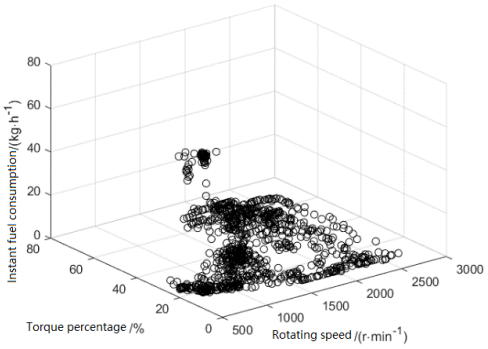

Figure 15. Instantaneous fuel consumption off for driving time and speed

It can be observed from Figure 15 above that the instantaneous fuel consumption is mostly concentrated in the speed of 1000-1500r / min, and the percentage of torque is 10% - 30%, which indicates that this part is composed of high-speed, medium speed and low-speed driving conditions. There are a few relatively concentrated areas in the speed range of 1500-2500r / min. it can be observed that this part is the instantaneous fuel consumption generated under the condition of high engine speed and low torque percentage, which may be due to driving It is caused by the extreme operation of the driver.

VI. CONCLUDING REMARKS

This paper proposes an improved optimization algorithm for the combination of principal components and feature-weighted K-means clustering, and introduces the residual point clustering mean method to eliminate outliers and reduce clustering time. The maximum minimum distance method can optimize the candidate initial centers, so that K-means avoids falling into the local optimal solution, so as to achieve a good clustering effect. According to the contribution rate of the eigenvalue contribution factor to the cluster, the initial feature weight is obtained, and a weighted Euclidean distance metric is proposed. Select characteristic values such as cruise time ratio, travel distance, average speed and so on with larger contribution factors, and then increase the weight to perform cluster analysis to construct vehicle driving conditions. The improved clustering algorithm proposed in

this paper still has room for improvement. The weighted density K-means clustering algorithm can be proposed on the basis of the algorithm in this paper. You can also consider directly removing outliers in the data preprocessing part of this paper to reduce the running time of subsequent clustering. You can also add more dimensional feature information.

ACKNOWLEDGMENT

The authors wish to thank the cooperators. This research is partially funded by the Project funds in Shaanxi province University Student Innovation and Entrepreneurship Fund Project (S S202010702115X) and the Project funds in engineering laboratory project (GSYSJ2018013).

REFERENCE

[1] Yuan Su-fen. Research on driving conditions of urban vehicles and optimal matching of transmission system[D]. Wuhan University of Technology, 2013.

[2] Nguyen, Nghiem, Le, et al. Development of the typical driving cycle for buses in Hanoi, Vietnam. 2019, 69(4):423-437.

[3] Ding Yi-feng, Li Jun, Liu Yu. Experimental study on actual road driving conditions of heavy diesel vehicles[J]. Automotive Engineering, 2017, 39(12): 1438-1443.

[4] Liu Ying-ji, Xia Hong-wen, Yao Yu, et al. Vehicle driving condition formulation method combining principal component analysis and fuzzy c-means clustering [J]. Highway and Transportation Science and Technology, 2018, 35(03): 79-85.

[5] Zhang Rui, Wang Yi-wu, Zhu Xiao-long, et al. Research on K-means algorithm for optimizing initial center based on UPGMA [J]. Computer Technology and Development,2018,28(02):50-53+58.

[6] Zhang Lin, Chen Yan, Ji Ye, et al. Research on a density-based K-means algorithm [J]. Application Research of Computers, 2011, 28(11): 4071-4073+4085.

[7] Luo Jun-feng, Suo Zhi-hai. A density-based k-means clustering algorithm [J]. Microelectronics and Computer, 2014, 31(10): 28-31.

[8] Zhang Yan. Clustering algorithm based on rough set and genetic algorithm [D]. Shaanxi Normal University, 2010.

[9] Ding Yi-feng, Li Jun, Gai Hong-chao, et al. Application of wavelet transform in vehicle speed data processing for construction of driving conditions [J]. Science Technology and Engineering, 2017, 17(28): 274-279.

[10] Li A-wu, Zhang Cui-ping, Wang Yang, et al. Research on the construction of driving conditions and emission values of light vehicles in Taiyuan City [J]. Chinese Science and Technology Papers, 2017, 12(22): 2537-2542.

[11] Peng Yu-hui, Zhuang Yuan. Combinatorial optimization clustering and Markov chain construction method of urban sanitation vehicle driving conditions[J]. Journal of Fuzhou University (Natural Science Edition), 2019, 47(04): 502-508.

[12] Chen Zhao-ming, Wang Wei, Zhao Ying, et al. Improved Principal Component Analysis and Multiple Regression Integration of Hanfeng Lake Water Quality Assessment and Prediction [J]. Environmental Monitoring Management and Technology, 2020, 32(04): 15-19 .

[13] Liu Qing-yuan, Li Yong, Pu Xun-chi, et al. Application research of improved principal component analysis method in reservoir water quality evaluation[J]. Sichuan Environment, 2017, 36(06): 116-122.

[14] Yuan Su-fen. Research on driving conditions of urban vehicles and optimal matching of transmission system[D]. Wuhan University of Technology, 2013.

[15] Zhang Jie, Zhuo Ling, Zhu Yun-you. Improvement and application of a K-means clustering algorithm [J]. Application of Electronic Technology, 2015, 41(01): 125-128+131.

[16] Song Yi-fan. Construction of urban road vehicle driving conditions in Shenzhen based on clustering and Python language [D]. Chang'an University, 2018.

[17] Gao Jian-ping, Gao Xiao-jie. Construction of actual driving conditions of vehicles based on improved fuzzy C-means clustering method [J]. Journal of Henan University of Science and Technology (Natural Science Edition), 2017, 38(06): 21-27+4-5.

[18] Liu Bing-jiao, Shi Qin, Qiu Duo-yang, et al. Construction of driving conditions and accuracy analysis based on improved ant colony algorithm [J]. Journal of Hefei University of Technology (Natural Science Edition), 2017, 40(10): 1297- 1302.