IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 1, 2019, accepted June 22, 2019, date of publication July 3, 2019, date of current version July 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926593

Research on Classification and Recognition of Driving Styles Based on Feature Engineering

YONGGANG LIU[1,2], (Senior Member, IEEE), JIMING WANG[1,2], PAN ZHAO[1,2], DATONG QIN[1,2], AND ZHENG CHEN[3], (Senior Member, IEEE)

[1]State Key Laboratory of Mechanical Transmissions, Chongqing University, Chongqing 400044, China
[2]School of Automotive Engineering, Chongqing University, Chongqing 400044, China
[3]Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming 650500, China

Corresponding authors: Yonggang Liu (andylyg@umich.edu) and Zheng Chen (chen@kmust.edu.cn)

This work was founded in part by the National Science Foundation of China (No. U1764259), in part by Chongqing Fundamental Research and Frontier Exploration Project (No. CSTC2018JCYJAX0409), in part by the Fundamental Research Funds for the Central Universities (No. 2018CDJDCD0001 and 2018CDQYQC0035), and in part by the EU-funded Marie Skłodowska-Curie Individual Fellowships Project under Grant 845102-HOEMEV-H2020-MSCA-IF-2018.

ABSTRACT Accurate classification and the effective recognition of driving styles are critical for improving control performance of the vehicle powertrain. In this research, a set of driving style classification and recognition methods is built based on the feature engineering. First, a specified road test is conducted considering the influence factors, and meanwhile, the corresponding driving data is collected, followed by a detailed evaluation of the driving styles. Then, the information entropy is applied to discretize the driving data, including the speed, acceleration, and opening degree of the accelerator pedal, and 44 feature quantities are extracted to characterize the driving style. By analyzing strong correlation and redundancy among the constructed feature quantities, the principal component analysis (PCA) is employed to reduce the dimension, and the fuzzy c-means (FCM) clustering algorithm is leveraged to classify the driving style. The successful classification rate reaches 92.16%, which is improved by 9.81% in comparison with traditional features. Finally, a parameter identification algorithm based on the support vector machine (SVM) is applied to identify the classified driving style, and the recognition accuracy reaches 92.86%, which is improved by 7.15% in comparison with traditional features, proving the feasibility of the proposed algorithm.

INDEX TERMS Driving style classification, feature discretization, fuzzy c-means (FCM) clustering, support vector machine (SVM).

I. INTRODUCTION

Nowadays, with the development of electrical and mechanical technologies, driving experience of automobiles has attained wide attention. In a vehicle, the powertrain control including the shifting rules design, braking and accelerating control, transmission control, and engine operation optimization, has been widely researched to improve the driving performance. In addition, the driving style identification and corresponding adjustment of the powertrain control and energy management strategies also play an important role in improving the vehicle driving performance [1], [2]. An effective identification algorithm cannot only improve the driver's operation experience, but also supply the reference for optimal control of the powertrain, thereby reducing the fuel consumption and greenhouse gas (GHG) [3]–[5].

In terms of driving styles, a main task is to conduct the classification and recognition. Currently, a variety of researches have emerged to improve the identification precision based on various algorithms. Typical classification methods include the subjective definition and objective division based on advanced algorithms [6], [7]. Currently, there are two common subjective definitions for the driving style. A typical subjective definition method is the subjective questionnaire, in which drivers are asked to fill in a special table [8]–[10]. Another method is called the level rules, which classify the drivers according to few thresholds of operating actions, such as the jerk and throttle position [7], [11]. However, the subjective evaluation method based on simple logics may lead to certain subjectivity, and the driving style cannot be objectively defined. With the development of data mining and modern communication technologies, more and more driving data can be collected and numerous machine learning algorithms are employed to classify driving styles with improvement of

The associate editor coordinating the review of this manuscript and approving it for publication was Rui Xiong.

rationality and accuracy [12], [13]. In [14], [15], the driver's behavioral characteristics are studied by collecting information from on-board GPS sensors and applying three different approaches, i.e., the DP-means algorithm, hidden Markov model (HMM), and behavioral topic extraction. As such, the contextual scene detection is conducted and different behaviors in each trip are identified. In [16], a fuzzy synthetic evaluation method is introduced to classify thirty driving styles into three different types, i.e., cautious, moderate and aggressive. In order to overcome uncertainty of the driving behavior and time-consuming limitation of manual marking when identifying the driving styles, a semi-supervised learning algorithm is proposed based on the marked data points [17].

Currently, a variety of existing advanced algorithms are employed to recognize the driving styles, which can be broadly divided into two categories: model-based and learning-based [18]. A direct identification manner is to build a driver model capable of characterizing the basic driving behaviors. In [19], the HMM is leveraged to model and predict the driving behavior because of its strong capabilities of describing latent state in dynamic and stochastic processes and dealing with time-series data. In [20], [21], three driving styles are modeled and identified through the HMM based on the braking characteristics. The other way is to directly analyze the driving data using pattern recognition or data-mining methods. In [22], a Bayesian nonparametric learning method based on the hidden semi-Markov model is introduced to extract primitive driving patterns according to the time-series driving data. In [23], an inverse reinforcement learning (RL) method is harnessed to learn individual driving styles for autonomous vehicles. To shorten calculation time and improve recognition precision of driving styles, a k-means clustering support vector machines (SVM) method is developed to classify the driving style into two types, i.e., aggressive and moderate [24].

Actually, a preliminary process when conducting classification and identification of the driving style is to determine characteristic variables. Choosing effective signals is crucial since any further actions and subsequent results would largely depend on it [25]. At present, characteristic parameters commonly employed to highlight the driving styles include the vehicle speed, acceleration, opening degree of the accelerator pedal, and jerk [16], [26]. Since the collected feature quantities, such as the vehicle speed and acceleration, are a series of continuous data, the first step is to translate them into a number of feature variables. Many researches are performed to discretize them based on statistical values of their standard deviation, mean values and maximum values [4], [16], [27]. References [14], [16], [28] indicate that the time percentage with different speed interval accounts for an important influence on classification of driving patterns and styles, and therefore it can be incorporated with additionally acquired signals and statistical information to facilitate the classification. However, the selection of the speed range is usually determined based on experience. Few researches have studied the distribution law of the speed based on the long-time driving data with proper manners. To the best knowledge of authors, there are few studies focusing on selection and construction of characteristic parameters of the driving style, which can be potentially beneficial to identification and classification of the driving styles. Motivated by this, the feature engineering algorithm is firstly proposed in this study to classify the driving styles by incorporating their influence factors. Feature engineering refers to the process of transforming raw data into the training data of the target model, and its main purpose is to acquire better training features and pave the road for more precise identification [29]. In this paper, the characteristics of the driving style are statistically analyzed, and then the information entropy is applied to discretize the speed, acceleration and opening degree of the accelerator pedal, so as to further excavate the distribution law of these characteristics. Considering the strong correlation and redundancy among those features, the principal component analysis (PCA) is conducted to reduce the dimension of the constructed features, and a set of features are selected that can effectively characterize the driving style. Based on this, the fuzzy c-means (FCM) clustering and SVM are leveraged to classify and identify the driving styles. Experimental results demonstrate feasibility of the proposed classification and pattern recognition algorithm.

The main contributions are attributed to the following two aspects: 1) The information entropy is applied to discretize the acquired driving data into different intervals. In this manner, the distribution law of the driving data is more clearly expressed. Based on the discretization results, 34 feature quantities that can effectively characterize the driving style are constructed. 2) The FCM and SVM are harnessed to construct an efficient model thereby classifying and identifying the driving style effectively.

The remainder of this article is structured as follows: Section II details the road data acquisition process. Section III analyzes the collected data based on the statistical methods. In Section IV, the FCM clustering algorithm is proposed to classify the driving style, and the SVM is applied to identify the driving style, followed by validation by the experimental results. Finally, main conclusions are drawn in Section V.

II. EXPERIMENT DESIGN AND DATA COLLECTION

The study of driving styles relies on precise acquisition of road experiment data. Note that we neglect the uncertainty induced by the road slope and weather variation. In this study, a total of 51 drivers, numbered from 1 to 51 hereinafter, are selected that age from 20 to 50 years old. Their actual driving years range from 1 to 20. Meanwhile, three professional driving assessors are engaged in evaluating each driver's operating style. The evaluation rules are as follows: during the test, the experimenter drives the vehicle from the starting point to the end according to the usual driving habits, while the professional evaluator rides on the vehicle, and evaluates the driving style of the experimenter by observing the intensive process of the experimenter's operation and the

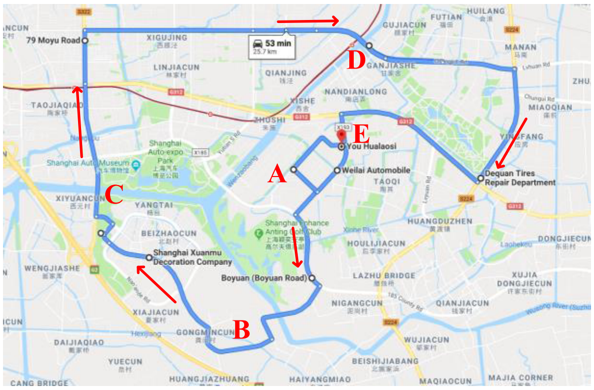


FIGURE 1. Experiment path.

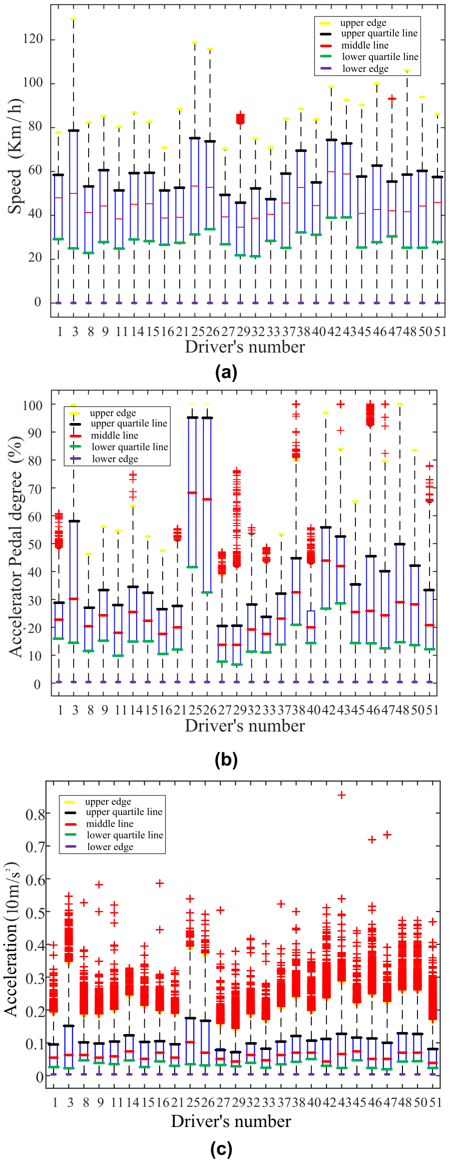
TABLE 1. Subjective evaluation of driving styles.

Subjective evaluation	Driving Style	Driver's number
1	Aggressive	3, 12, 13, 25, 26, 38, 42, 43, 45, 46
0	General	1, 2, 5, 6, 14, 15, 19, 20, 22, 23, 24, 30, 34, 37, 39, 40, 41, 44, 47, 48, 49, 50, 51
-1	Conservative	4, 7, 8, 9, 10, 11, 16, 17, 18, 21, 27, 28, 29, 31, 32, 33, 35, 36

subjective feeling of the professional evaluator during the vehicle driving process. Finally, the current driving style was chosen as the most evaluated driving style by three evaluators.

In order to consider influences of the road type, driving condition and other factors that can affect the driving style, the experimental driving route is carefully designed, as shown in Fig. 1. The total length of the designed route is around 25 Km and the whole driving duration is about 40 minutes. To satisfy diversity of road conditions, the selected route involves the rural highway, national freeway and urban congested conditions, and includes the curve road, straightway and rampway. As can be seen in Fig. 1, sections AB and DE are the typical rural highways, section BC is the urban congested road, and section CD belongs to the freeway. Meanwhile, the selected path includes 22 curves. In order to make the collected data include more road conditions and better show the characteristics of the corresponding road conditions, this paper chooses 12 o'clock a day as the starting time of the experiment.

To reach a more convincible classification of the driving style, three experts participated in evaluating the driving style [30], [31], and judged the drivers' style by observing their operations according to their own evaluation. All the drivers' styles are divided into three grades, with −1, 0 and 1 being conservative, general and aggressive, respectively. After collecting all the test data, the most voted value is labeled as the style of the current driver. Since the study primarily focuses on the impact of longitudinal driving behavior influenced by the driving style, the data collected mainly includes the speed, longitudinal acceleration and opening degree of the acceleration pedal with acquisition frequency of 10 Hz. After experiment, 51 sets of driver's data are acquired, as listed in Table 1, in which there are 10 aggressive styles, 23 general styles and 18 conservative styles. We can find that most of drivers prefer to drive the car normally or conservatively for safety.



(a)

(b)

(c)

FIGURE 2. Box plot analysis. (a) Speed box plot; (b) Accelerator pedal position; (c) Acceleration box plot.

In the next step, data analysis and classification are conducted to feature the key factors of the driving style.

III. PREPROCESSING OF EXPERIMENTAL DATA

The collected data is analyzed by the box-line figures, and the information entropy algorithm is employed to discretize the vehicle speed, acceleration and opening degree of the accelerator pedal. In this study, 44 feature variables representing the driving style are constructed and the dimension of variables is reduced by the PCA.

A. DATA ANALYSIS

In order to compare differences among drivers, the box-line figures are plotted to depict the location and dispersion of each driver's habit, as shown in Fig. 2. Five marks existing in each line from top to bottom are the upper edge, upper quartile line, middle line, lower quartile line, and lower edge, respectively. Fig. 2 shows the distribution of the speed, acceleration pedal position and acceleration of each driver.

As can be seen from Fig. 2 (a), the speed distribution of drivers with different driving styles is different. Aggressive drivers such as No. 3, No. 25 and No. 26 have wider box lines, which indicates that their speed distribution range is larger, and the median line and the upper quartile line are also higher than other drivers, which indicates that the range with higher speed is larger. Compared with general drivers such as No. 14 and No. 15, conservative drivers such as No. 8, No. 11 and No. 16 have lower middle and low quartile lines and lower quartile lines, indicating that their speeds are more distributed in low speed areas. It can be seen that different styles of drivers have different speeds at different speeds. Similar phenomenon occurs in the distribution of the accelerator pedal position and acceleration. For the throttle opening, the distribution law of different drivers under different throttle opening is more different.

B. DISCRETIZATION OF CONTINUOUS ATTRIBUTION

Discretization of continuous attribution is essential for data mining, and discretization results can directly correlate to the learning efficiency [32]. The main purpose of discretization is to divide the continuous attribution into a number of unequal ranges among units, each of which corresponds to a discretized value. Amongst them, some of characteristic values represent continuous attribution. According to the existing research and expert experience [15], [33], the vehicle speed, acceleration and opening degree of the accelerator pedal show great influence on the driving style. In this study, the information entropy is applied for discretization of continuous attribution, and the percentage of the data set in each interval is calculated to construct the feature quantity characterizing the driving style.

For the continuous attribution of each variable in the database, its range can be divided into a number of intervals, and each interval requires at least one sample. Thus, m samples can only be divided into m intervals, i.e., $O(m)$. If there are maximum amounts of attribute values in the frequency distribution, then the entropy can be maximized. The following two conditions, i.e., minimization of the dimensions and minimum loss of key information of the attribute value, should be attained in terms of the discretization. Actually, the information entropy represents the average amount of information provided by each event. The information entropy of a discrete random variable can be defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

where x represents an event and $p(x)$ denotes its probability. Assume that the discrete value of a continuous random variable X can be represented by a separate interval, and the total number of intervals is k. In each interval, the probability of X is $p(i)$, and the entropy of the distribution of k discrete intervals is $H(p_k)$. Reference [32] demonstrates that if the two adjacent intervals, i.e., i and $i+1$, are combined and meanwhile $H(P_k) - H(P_{k-1})$ is minimized, then the merged probability is monotonically non-decreasing. $H(p_k)$ is a concave function with respect to k, and reducing the interval number k can minimize the change amount of $H(p_k)$.

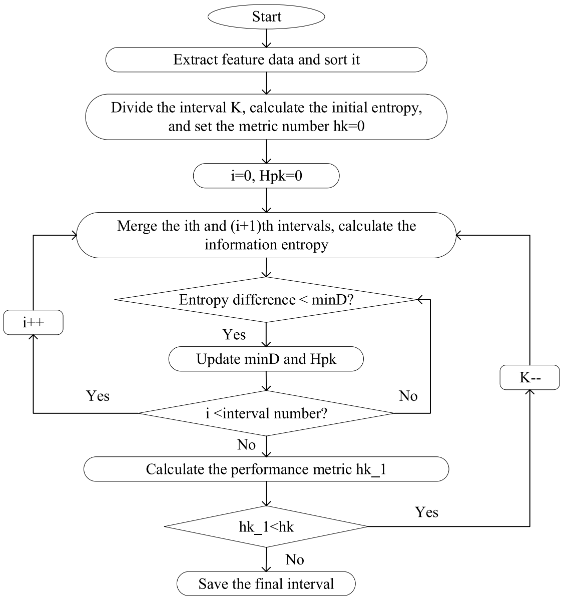

Start
Extract feature data and sort it
Divide the interval K, calculate the initial entropy, and set the metric number hk=0
i=0, Hpk=0
Merge the ith and (i+1)th intervals, calculate the information entropy
Entropy difference < minD?
Yes
Update minD and Hpk
i++
Yes
No
i <interval number?
No
Calculate the performance metric hk_1
hk_1<hk
Yes
K--
No
Save the final interval

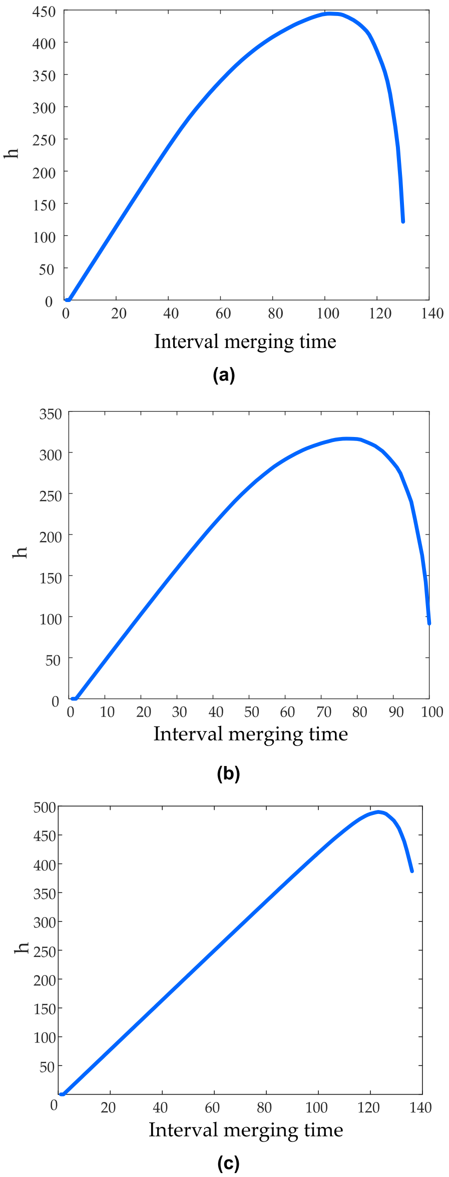
FIGURE 3. Flowchart of feature discretization.

Based on it, this paper proposes an entropy-based discretization method for continuous attribution values. The detailed processing flowchart is shown in Fig. 3.

For K non-repetitive attribution values, they are first sorted by size, and then divided into a number of intervals, each of which corresponds to a non-repetitive value. In the next step, the probability of each interval is calculated and the point of demarcation is saved. Based on (1), the entropy $H(p_k)$ is calculated and then the adjacent intervals are combined to minimize the entropy difference before and after merging. Meanwhile, the entropy after merging is saved and the point of demarcation is reset. In the above process, when two adjacent intervals are merged, the interval with the smallest entropy difference is selected as the interval of the merge. The merging steps are repeated until the ending condition is reached. In this manner, the discrete values and dividing points of all continuous attributes are determined.

Based on the above analysis, determination of the stopping point is critical to optimize the combined results. According to the characteristics of information entropy function, this study designs an optimal stopping point judgment method as follows.

Since $H(p_k)$ is a concave function of p and monotonically increases as k, the increase rate gradually decreases when k approaches its maximum value. Meanwhile, when the entropy reaches the maximum point v_1, the corresponding number of intervals can also attain the maximum value. From



(a)

(b)

(c)

FIGURE 4. Feature Discretization. (a) Velocity discretization; (b) The discretization of the opening degree of accelerator pedal; (c) Acceleration discretization.

this point of view, if a line L is plotted from the starting point v_2 of the entropy function curve to v_1, all the points should locate above L. Consequently, the furthest point v_0 can be found in the inflection point and at this moment, an optimal balance can be reached between the entropy loss and moderate interval amounts.

In this manner, v_0 can be considered as the termination point when merging the neighboring intervals. By drawing a vertical line$H$ from any point on the function curve to L, we can attain:

$$h = (k_{\max} - 1)H(p) - H_{\max}(p)(k - 2) \tag{2}$$

where $k_{\max}$ denotes the maximum interval number, which is also the number of non-repetitive attribution and $H_{\max}(p)$ means the corresponding maximum value. Obviously, the maximum value of L is located at v_0, by which the optimal interval amounts can be solved. Detailed discretization results can be shown in Fig. 4.

TABLE 2. Discretization results of speed, acceleration and accelerator pedal degree.

Item	Interval discretization		
Speed	(0, 15.5)	(15.5, 20.5)	(20.5, 27.5)
	(27.5, 31.5)	(31.5, 35.5)	(35.5, 40.5)
	(40.5, 43.5)	(43.5, 47.5)	(47.5, 52.5)
	(52.5, 56.5)	(56.5, 64.5)	(64.5, 72.5)
	(72.5, $v_{\max}$)		
Acceleration	(0, 0.055)	(0.055, 0.085)	(0.085, 0.115)
	(0.115, 0.195)	(0.195, 0.26)	(0.26, 0.305)
	(0.305, 0.365)	(0.365, 0.475)	(0.475, $a_{\max}$)
Opening degree of the accelerator pedal	(0, 4.5)	(4.5, 8.5)	(8.5, 12.5)
	(12.5, 15.5)	(15.5, 17.5)	(17.5, 19.5)
	(19.5, 23.5)	(23.5, 26.5)	(26.5, 31.5)
	(31.5, 37.5)	(37.5, 44.5)	(44.5, $p_{\max}$)

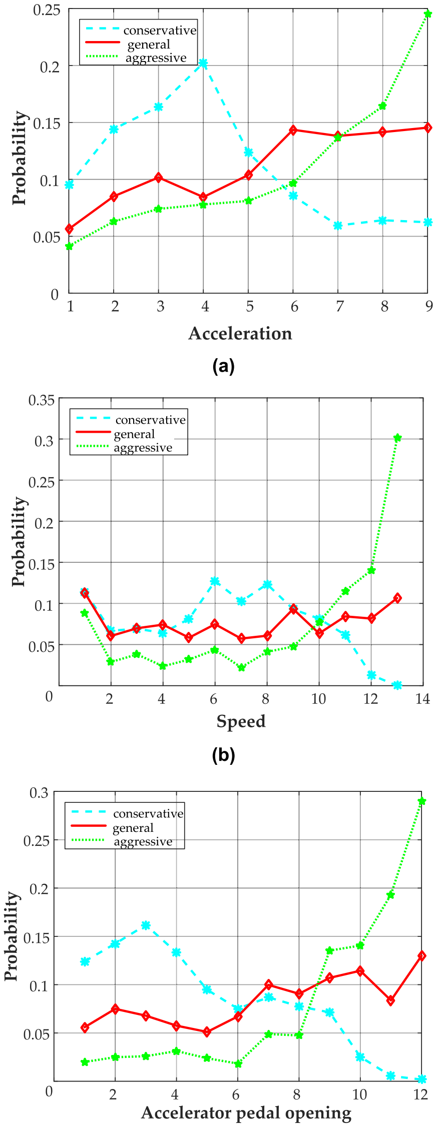
As can be seen from Fig. 4 (a), in the process of the speed discretization, when the interval merging time is 101, h reaches a maximum value of 431.5, and the interval discretization result is optimal. The number of generated intervals is 13. Similarly, for discretization of the acceleration and opening degree of the accelerator pedal, the interval merging times of 124 and 94 are finally determined as the interval emerging points, h equals 489.9 and 251.9, and the numbers of generated intervals are 9 and 12, respectively. Table 2 shows the final discretization results of these three variables.

Consequently, the discretization features can be calculated, as:

$$i_j = \frac{T_j}{T_{trip}} \times 100\% \tag{3}$$

where T_j denotes the time of interval vehicle speed, acceleration and opening degree of the accelerator pedal listed in Table 2, and T_{trip} means the cycle total time.

Based on the determined interval, probability distributions of the opening degree of the accelerator pedal, vehicle speed and acceleration are calculated and shown in Fig. 5. It can be observed that the conservative driver's speed distribution is higher from 5 to 9 and lower from 11 to 13. The probability of the general driver is almost uniformly distributed in each interval. For the aggressive driver, the distribution is significantly higher from 10 to 13. For the acceleration probability, the distribution of conservative drivers accounts for a larger proportion from 1 to 4, whereas the distribution of aggressive drivers becomes larger from 7 to 9. The distribution law of the opening degree of the accelerator pedal is similar to that of the acceleration. Therefore, it can be concluded that the discretization method proposed in this paper can effectively characterize the distribution law of the raw data. The built features are qualified in distinguishing different styles of drivers, thereby facilitating more effective identification among different types of drivers.



(a)

(b)

(c)

FIGURE 5. Feature Discretization.

C. FEATURE PARAMETER DIMENSION REDUCTION

Related findings in [4], [16], [27] show that most studies select ten variables, including the mean values and standard deviation of the vehicle speed (v_m, v_{std}), acceleration (a_m, a_{std}), degree of the jerk (j_m, j_{std}), opening degree of the accelerator pedal (p_m, p_{std}) and opening degree rate of the accelerator pedal (pd_m, pd_{std}), as traditional features to classify and identify the driving styles. By considering these 10 traditional variables and adding 34 new ones, 44 characteristic variables are selected to characterize the driving styles, as detailed in Table 3. The 34 new variables include the percentage of the interval vehicle speed to the total speed range (v_1 to v_{13}), the percentage of the interval acceleration to the total acceleration interval (a_1 to a_9) and the percentage of the interval opening degree of the accelerator pedal to the total opening degree interval (p_1 to p_{12}). Nevertheless, too many parameters will no doubt increase complexity of the clustering model. In fact, there exists strong coupling among these vectors of data, such as the speed, acceleration and opening degree of the accelerator pedal. Consequently, the PCA is employed to simplify the dimension of the constructed feature variables by explaining the strong correlation and deleting redundant characteristic variables.

TABLE 3. Driving styles parameters.

Characteristic Parameters	Symbol
Average speed	v_m
Standard speed difference	v_{std}
Average acceleration	a_m
Acceleration standard deviation	a_{std}
Average pedal opening	p_m
Pedal opening standard deviation	p_{std}
Average jerk	j_m
Jerk standard deviation	j_{std}
Average pedal opening change rate	pd_m
Pedal opening change rate standard deviation	pd_{std}
Interval speed percentage	$v_1 - v_{13}$
Interval acceleration percentage	$a_1 - a_9$
Interval pedal opening percentage	$p_1 - p_{12}$

The key purpose of the PCA is to transform raw data into several new independent components with minimization of the information loss. These principal components can represent most information of the raw data and are usually expressed by linear combination of raw variables. Supposing that p feature parameters are set to characterize the driving styles, denoted as $x = (x_1, x_2, \cdots, x_p)'$, and assuming that the expectation and covariance of x exist, expressed as $E(x) = \mu$ and $var(x) = \Sigma$, we can attain a linear transformation, as

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = a_1'x \\ y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p = a_2'x \\ \vdots \\ y_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p = a_p'x \end{cases} \tag{4}$$

where $\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_p}$ are unit vectors. By sorting $\boldsymbol{a_1}$, the variance of y_1 can be maximized. Supposing that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are eigenvalues of Σ and $t_1, t_2, \cdots, t_p$ are corresponding unit orthogonal eigenvectors, we can calculate the variance of y_1, as:

$$var(y_1) = \sum_{i=1}^{p} \lambda_i (\boldsymbol{a}_1{}'\boldsymbol{t}_i)^2 \leq \lambda_1 \sum_{i=1}^{p} (\mathbf{a}_1{}'\boldsymbol{t}_i)^2 = \lambda_1 \boldsymbol{a}_1{}'\boldsymbol{a}_1 = \lambda_1 \tag{5}$$

From (5), it can be known that when $\boldsymbol{a}_1$ equals $\boldsymbol{t}_1$, $y_1 = \boldsymbol{t}_1{}'\boldsymbol{x}$ has the largest variance and the maximum value is λ_i. Now, $y_1 = \boldsymbol{t}_1{}'\boldsymbol{x}$ is called the first principal component. Similarly, λ_i and its main component $y_i = \boldsymbol{t}_i{}'\boldsymbol{x}$ can be found. The results

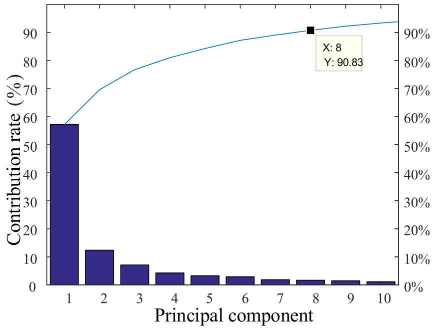


FIGURE 6. Contribution rate of the PCA.

obtained by the PCA in terms of the characteristic parameters describing the driving style are shown in Fig. 6.

As can be seen in Fig. 6, the cumulative contribution rate of the first eight principal components is 90.83%, which satisfies the requirement of more than 85%. As such, the first eight principal components are selected to characterize the driver's style, which provides a solid basis for the cluster analysis of the driving style.

IV. CLASSIFICATION AND RECOGNITION OF DRIVING STYLE

In this study, the FCM is employed to classify the driving style, followed by the identification conducted by the SVM.

A. CLASSIFICATION OF DRIVING STYLE

The cluster analysis can automatically classify the sample data, and an advantage is that it does not need to determine the classification criteria in advance. The FCM exhibits superior advantages in resolving classification uncertainty and fuzzy problems. The samples with concentrated distribution can be divided into multiple categories simultaneously to obtain the optimal clustering result.

The FCM is mainly based on the c-means algorithm, and the square function of the weighted error is substituted with the intra-class error square function, then the objective function can be calculated, as:

$$J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{\omega}\mu_{ik}^{\delta}d_{ik}^{2} \tag{6}$$

where c means the amount of classes, ω denotes the sample amount, U is the membership matrix of each center of the sample, V is the vector of each cluster center, and μ_{ik} expresses the membership degree of the center of the sample. d_{ik} is the norm distance of the sample to the center, which is generally expressed in terms of the Euclidean distance. δ presents the fuzzy weighted index, which is usually two by default.

The purpose of the FCM algorithm is to find a set of membership matrix and cluster center to ensure that the objective function reaches the minimum value, which is essentially an optimization problem. Detailed formulations are expressed as:

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}\left(d_{ik}/d_{jk}\right)^{\frac{2}{\delta-1}}} \tag{7}$$

$$v_i=\sum_{k=1}^{\varphi}\mu_{ik}^{\delta}x_k\Bigg/\sum_{k=1}^{\varphi}\mu_{ik}i=1,2\ldots c \tag{8}$$

where x_k is the kth sample vector, v_i is the ith cluster center vector. By iterating (7) and (8), a set of optimal U and V can be found. According to the dimensionality reduction results of the PCA, the total 51 drivers' styles are clustered by the FCM, and the results are shown in Fig. 7 and Table 4, respectively. Fig. 7 shows the variation of the objective function with iteration during the clustering process. When the iteration reaches 15, the objective function begins to stabilize, indicating that the clustering process takes effect. Table 4 shows the results after clustering.

Combining with the driving data, we can find that the third type of driver presents characteristics including higher vehicle speed, faster variation rate of the vehicle speed, larger acceleration, larger degree of jerk, and larger throttle opening and changing rate of the throttle opening. However, the characteristics values of the first type of driver are relatively small, and meanwhile the second one locates in the middle.

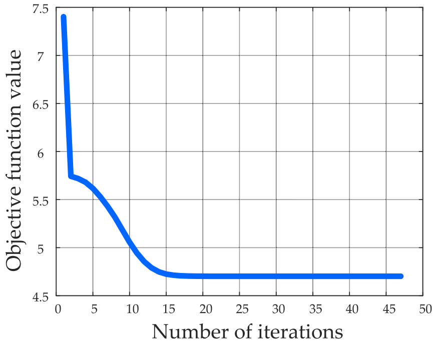


FIGURE 7. Objective function value based on optimized feature quantities.

TABLE 4. Classification results of driving styles.

Class	Driving Style	Driver's number
First	Conservative	4, 7, 8, 10, 11, 16, 17, 18, 21, 27, 28, 29, 31, 32, 33, 35, 36, 40
Second	General	1, 2, 5, 6, 9, 12, 14, 15, 19, 20, 22, 23, 24, 30, 34, 37, 39, 41, 44, 45, 47, 48, 49, 50, 51
Third	Aggressive	3, 13, 25, 26, 38, 42, 43, 46

TABLE 5. Comparison of classification results based on optimized feature quantities.

Item Type	Actual Sample Quantity	Number of samples			Accuracy	Average accuracy
		C	G	A		
C	18	17	1	0	94.44%	92.16%
G	23	1	22	0	95.65%	
A	10	0	2	8	80.00%	

Note that C, G and A represent conservative, general and aggressive driving style, respectively.

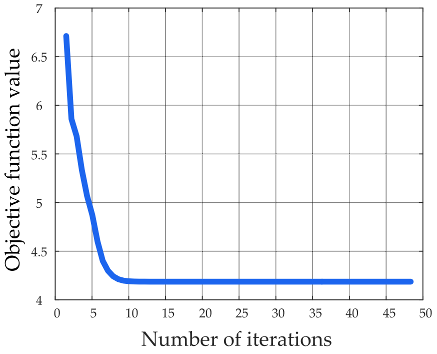


FIGURE 8. Objective function value based on traditional feature quantities.

TABLE 6. Comparison of classification results based on traditional features.

Class	Driving Style	Driver's number
First	Conservative	1, 4, 6, 7, 8, 10, 11, 16, 17, 18, 21, 27, 28, 29, 30, 31, 32, 33, 36, 40
Second	General	2, 5, 9, 12, 14, 15, 19, 20, 22, 23, 24, 34, 35, 37, 38, 39, 41, 44, 45, 47, 48, 49, 50, 51
Third	Aggressive	3, 13, 25, 26, 42, 43, 46

Therefore, these three drivers belong to conservative, general and aggressive types, respectively.

The classification results of the driving styles based on the FCM are compared with the results of subjective evaluation, as shown in Table 5. It can be found that Nos. 12 and 45 who are defined as aggressive drivers are classified into the general type; No. 40 who belongs to the general type is classified as a conservative type; and No. 9 who is assumed as the conservative type is classified into a general type. The overall classification accuracy of the driving style is 92.16%.

In order to validate whether the discretized features have a positive effect in clustering the driving styles, ten traditional features (top 10 feature variables listed in Table 3) that have not been optimized are selected for comparison. After the dimension reduction by the PCA, the first eight principal components, of which the accumulated contribution rate is 99.64%, are selected and the driving style is clustered by the FCM. Fig. 8 shows the variation of the objective function during the clustering process, and Table 6 lists the classification result after the clustering process.

As can be seen, the third type of the driver's style is characterized by the high average speed, large acceleration, large degree of jerk, and large opening degree of the accelerator pedal, therefore it belongs to an aggressive type. In addition, the characteristics of the first type are low, consequently it is a conservative type; and the second type of driver belongs to the general type. The classification results are shown in Table 7.

It can be observed from Table 7 that for the aggressive types, there are 3 misclassifications for Nos. 12, 38, and 45; and for the general driver, there are 4 misclassifications, i.e., Nos. 1, 6, 30 and 40. The overall classification accuracy rate is 82.35%. The two classification results show that both conservative and aggressive drivers are easily clustered into general drivers, whereas the general drivers are more likely to be clustered into conservative drivers. To a certain extent, this is because that most of aggressive drivers are particularly aggressive when driving the vehicle. As shown in Fig. 2, drivers such as Nos. 3, 25, and 26 are much superior to other drivers in all indicators. At the same time, the clustering result based on all the features is 9.81% higher than the average recognition rate of the traditional features, and the algorithm can effectively classify the behavior of Nos. 1, 6, 30 and 35. However, these drivers are clustered incorrectly by partial features, and all features can effectively correct the false clustering results. It proves that the feature discretization based on the information entropy can effectively optimize the traditional feature quantity, thereby improving the classification accuracy of the driving style. Based on the above discussion, the information entropy can better discretize the characteristics of the speed, acceleration, and opening degree of the accelerator pedal and the obtained discrete interval can effectively represent distribution rules of the raw data. The constructed characteristic values can supplement extra information of the traditional algorithm, thus enhancing clearer classification of the driving style and improving the classification precision.

TABLE 7. Comparison of Classification Results Based on Traditional Features.

Item Type	Actual Sample Quantity	Number of samples			Accuracy	Average accuracy
		C	G	A		
C	18	16	2	0	88.89%	82.35%
G	23	4	19	0	82.61%	
A	10	0	3	7	70%	

Notes that C, G and A represent conservative, general and aggressive driving style, respectively.

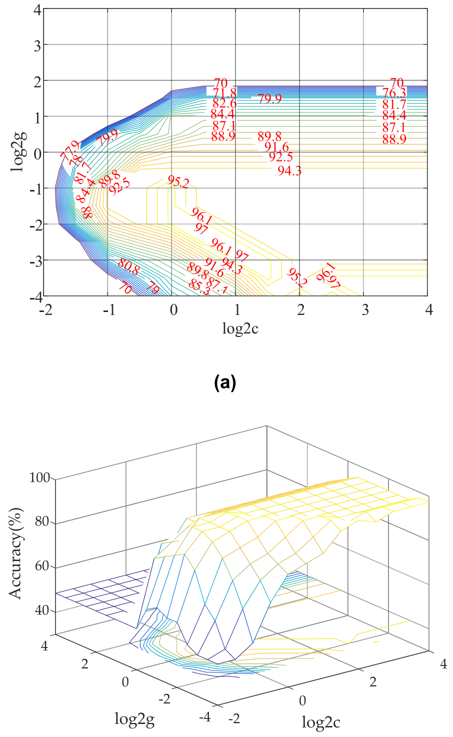
B. RECOGNITION OF THE DRIVING STYLE

The SVM can well adapt to the high-dimensional space and is suitable and effective to solve high nonlinear problems. Moreover, it can provide satisfactory generalization for pattern classification, and consequently it is qualified for classifying the driving styles. The key idea of SVM is to maximize the interval between the support vector and classification hyperplane.

Given training vectors $x_i \in R^h$, $i = 1, \ldots, l$ in two classes, and a vector $y \in R$, $y_i \in \{-1, 1\}$, the support vector can find the solution of the following optimization problem [34], as

$$\begin{aligned}&\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{I} \xi_i,\\ &\textit{subject to } y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i,\\ &\qquad \xi_i \geq 0, \quad i = 1, \ldots l. \end{aligned} \tag{9}$$

where C is a penalty factor to balance the accuracy and complexity of the model, and ξ_i is the slack non-negative variable. Here, the Lagrange factor is introduced to resolve this optimization problem, and one can easily obtain the



(a)

(b)

FIGURE 9. SVC parameter selection graph (Optimal c = 1, g = 0.25 CV Accuracy = 100%). (a) Contour map; (b) 3D view discretization.

following Wolfe dual format of the primal quadratic programming problem, as

$$\begin{aligned} &\min_{\alpha_i} \frac{1}{2}\sum_{i,j=1}^{I}\alpha_i\alpha_j y_i y_j k(x_i x_j) - \sum_{i=1}^{I}\alpha_i,\\ &\textit{subject to}\ 0 \le \alpha_i \le c, \quad i = 1,\ldots i,\\ &\qquad y^T\alpha = 0. \end{aligned} \tag{10}$$

SVM works in the feature space F via the nonlinear mapping function $\varphi : R^h \mapsto F$, which can be defined implicitly by a kernel function $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$. At the optimal point of (10), we can get either $\alpha_i = 0$, $0 < \alpha_i < C$, or $\alpha_i = C$. The input vectors for $\alpha_i > 0$ are indexed as the support vectors. These is only one important information from the perspective of classification, as they define the decision boundary, and yet the rest of the inputs may be ignored. For a binary classification problem, the decision function of the SVM [34] can be expressed as

$$f(x) = \text{sgn}(\sum_{i=1}^{N_S}\alpha_i k(x_i, x) + b \tag{11}$$

where α_i is the corresponding weight of support vector x_i, x is the input pattern to be classified, and N_s is the number of support vectors and b is the bias.

In this paper, the radial basis function is applied as the kernel function to establish the model. According to the principle of the SVM, the penalty factor C and kernel function parameter g are crucial to influence prediction precision of the model. In this study, the grid search algorithm is employed to find the ladder values for C and g. For the fixed C and g, the training set is used as the original data set to obtain the predicted mean square error by the K-fold cross validation (K-CV) method. Finally, C and g that minimize the mean square error (MSE) of the prediction result of the training set are selected as the optimal parameters. It is necessary to point out that multiple sets of C and g may exist corresponding to the smallest MSE, and we select the group with the smallest C as the final parameter, since larger C may lead to occurrence of over learning.

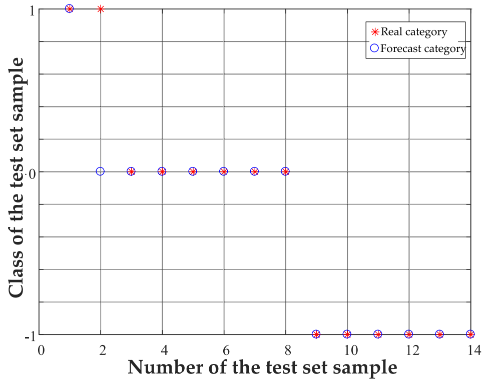


FIGURE 10. The driving style recognition result based on the SVM (accuracy = 92.8571%).

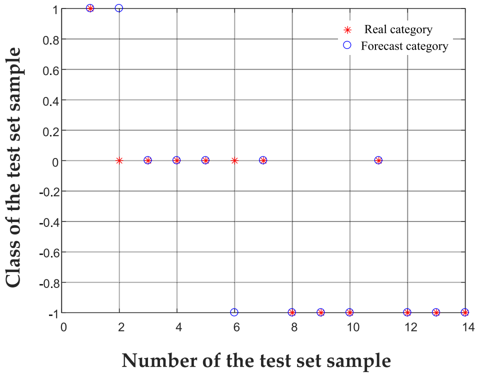


FIGURE 11. The driving style recognition result based on traditional feature quantities (accuracy = 85.71%).

Based on the FCM algorithm, the clustered drivers are renumbered, in which we can find samples 1 to 8, 9 to 29 and 30 to 51 belong to aggressive, general and conservative types, respectively. In this paper, 70% data is used for training and the remaining 30% data is applied for validation [34]. We selected samples 1 to 6, 14 to 26 and 34 to 51 as the training set. Based on the K-CV method, C and g finally converge to 1 and 0.25 after trial and error.

By updating the SVM model according to the optimal C and g, the final validation results are shown in Fig. 10. As can be seen, the red dots index real categories, and blue dots denote the categories identified based on the algorithm. The accuracy of the overall prediction result is 92.86%, and only sample 2 is recognized incorrectly. As such, the effectiveness of the built strategy is validated.

In order to verify whether the constructed stylistic features of driving style have a positive effect on the classification results. For the same training set, test set and parameter optimization SVM model, the traditional features (top 10 feature variables listed in Table 3) are applied to recognize driving style, and the recognition results are shown in Figure 11.

From the figure, it can be seen that the correct rate of driving style recognition by using traditional features is 85.71%, and the number 2 and 6 drivers are misidentified. Therefore, compared with the traditional feature variables, the proposed feature quantities will improve the recognition accuracy by 7.15%.As such, the effectiveness of the built strategy is validated.

V. CONCLUSION

This paper applies the feature engineering to classify and identify the driving styles. First, the driving data that characterizes the driving style through the designed road test are collected. Then, based on the information entropy, the velocity, acceleration and opening degree of the accelerator pedal are discretized to construct 44 feature quantities that essentially reflect their distribution characteristics. By comparing the distribution of drivers with different styles under the constructed features, we can see that the constructed features can effectively describe the distribution of the original data and distinguish the three types of driving styles. Subsequently, the PCA is introduced to reduce the dimension of the constructed feature quantities. In this manner, a set of feature quantities that can effectively characterize the driving style are obtained, and the FCM algorithm is applied to classify the driving style. The results show that the classification accuracy of the constructed feature quantities is 9.81% higher than the traditional algorithm. After that, the driving style is identified by the SVM algorithm, of which the main parameters are optimized, and the recognition accuracy of the constructed feature quantities is 7.15% higher than the traditional features. The validation results prove the effectiveness of the proposed feature-based driving style classification method. Therefore, we conclude that the proposed algorithm can supply a solid foundation for development of advanced algorithms, by which the powertrain controlling performances can be improved with consideration of the driving styles.

Next step, more identification and experiment in terms of the driving styles will be carried out and the corresponding energy management strategies of hybrid electric vehicles will be researched by incorporating identification results of the driving styles.

REFERENCES

[1] X. Rui, Y. Zhang, H. He, Z. Xuan, and M. G. Pecht, "A double-scale, particle-filtering, energy state prediction algorithm for lithium-ion batteries," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1526–1238, Feb. 2017.

[2] X. Rui, C. Huan, W. Chun, and S. Fengchun, "Towards a smarter hybrid energy storage system based on battery and ultracapacitor—A critical review on topology and energy management," *J. Cleaner Prod.*, vol. 202, pp. 1228–1240, Nov. 2018.

[3] T. Lee and J. Son, "Relationships between driving style and fuel consumption in highway driving," *Religious Stud.*, vol. 20, no. 3, pp. 316–323, 2011.

[4] G. Kedar-Dongarkar and M. Das, "Driver classification for optimization of energy usage in a vehicle," *Procedia Comput. Sci.*, vol. 8, pp. 388–393, Jan. 2012.

[5] P. Themann, J. Bock, and L. Eckstein, "Optimisation of energy efficiency based on average driving behaviour and driver's preferences for automated driving," *IET Intell. Transport Syst.*, vol. 9, no. 1, pp. 50–58, 2015.

[6] M. Ishibashi, M. Okuwa, S. Doi, and M. Akamatsu, "Indices for characterizing driving style and their relevance to car following behavior," in *Proc. SICE Annu. Conf.*, Sep. 2007, pp. 1132–1137.

[7] Y. L. Murphey, R. Milton, and L. Kiliaris, "Driver's style classification using jerk analysis," in *Proc. IEEE Workshop Comput. Intell. Vehicles Veh. Syst.*, Apr. 2009, pp. 23–28.

[8] T. B. A. Orit, M. Mario, and G. Omri, "The multidimensional driving style inventory-scale construct and validation," *Accident Anal. Prevention*, vol. 36, no. 3, pp. 323–332, May 2004.

[9] H. H. van Huysduynen, J. Terken, J.-B. Martens, and B. Eggen, "Measuring driving styles: A validation of the multidimensional driving style inventory," in *Proc. 7th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, 2015, pp. 257–264.

[10] O. Taubman-Ben-Ari and V. Skvirsky, "The multidimensional driving style inventory a decade later: Review of the literature and re-evaluation of the scale," *Accident Anal. Prevention*, vol. 93, pp. 179–188, Aug. 2016.

[11] W. J. Zhang, S. X. Yu, Y. F. Peng, Z. J. Cheng, and C. Wang, "Driving habits analysis on vehicle data using error back-propagation neural network algorithm," *Comput. Control Inf. Educ. Eng.*, 2015, p. 55.

[12] Y. Ma, Z. Li, Y. Li, H. Li, and R. Malekian, "Driving style estimation by fusing multiple driving behaviors: A case study of freeway in China," *Cluster Comput.*, vol. 2, pp. 1–11, Jan. 2018.

[13] G. Li, S. E. Li, B. Cheng, and P. Green, "Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 113–125, Jan. 2017.

[14] M. Brambilla, P. Mascetti, and A. Mauri, "Comparison of different driving style analysis approaches based on trip segmentation over GPS information," in *Proc. IEEE Int. Conf. Big Data*, Jan. 2018, pp. 3784–3791.

[15] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2015, pp. 2641–2646.

[16] D. Chu, Z. Deng, Y. He, C. Wu, C. Sun, and Z. Lu, "Curve speed model for driver assistance based on driving style classification," *IET Intell. Transp. Syst.*, vol. 11, no. 8, pp. 501–510, Oct. 2017.

[17] W. Wang, J. Xi, A. Chong, and L. Li, "Driving style classification using a semisupervised support vector machine," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 5, pp. 650–660, Oct. 2017.

[18] W. Wang, J. Xi, and H. Chen, "Modeling and recognizing driver behavior based on driving data: A survey," *Math. Problems Eng.*, vol. 1, pp. 1–20, Feb. 2014.

[19] E. Tadesse, W. Sheng, and M. Liu, "Driver drowsiness detection through HMM based dynamic modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, Jun. 2014, pp. 4003–4008.

[20] C. Deng, C. Wu, N. Lyu, and Z. Huang, "Driving style recognition method using braking characteristics based on hidden Markov model," *Plos One*, vol. 12, no. 8, 2017, Art. no. e0182419.

[21] X. Meng, K. K. Lee, and Y. Xu, "Human driving behavior recognition based on hidden Markov models," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2007, pp. 274–279.

[22] W. Wang, J. Xi, and Z. Ding, "Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[23] J. Wu, Y. Du, G. Qi, and M. Xu, "Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis," *IET Intell. Transp. Syst.*, vol. 9, no. 8, pp. 792–801, 2015.

[24] W. Wang and J. Xi, "A rapid pattern-recognition method for driving types using clustering-based support vector machines," in *Proc. Amer. Control Conf.*, Jul. 2016, pp. 5270–5275.

[25] C. M. Martinez, M. Heuke, F. Y. Wang, G. Bo, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 666–676, Mar. 2018.

[26] W. Han, W. Wang, X. Li, and J. Xi, "Statistical-based approach for driving style recognition using Bayesian probability with kernel density estimation," *IET Intell. Transp. Syst.*, vol. 13, no. 1, pp. 22–30, Jan. 2018.

[27] A. Augustynowicz, "Preliminary classification of driving style with objective rank method," *Int. J. Automot. Technol.*, vol. 10, no. 5, pp. 607–610, Oct. 2009.

[28] E. Ericsson, "Independent driving pattern factors and their influence on fuel-use and exhaust emission factors," *Transp. Res. D, Transp. Environ.*, vol. 6, no. 5, pp. 325–345, 2001.

[29] M. R. Anderson and M. Cafarella, "Input selection for fast feature engineering," in *Proc. IEEE Int. Conf. Data Eng.*, May 2016, pp. 577–588.

[30] C.-C. Chang. (2008). *LIBSVM: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[31] A. C. Holman and C. Havârneanu, "The Romanian version of the multidimensional driving style inventory: Psychometric properties and cultural specificities," *Transp. Res. F, Psychol. Behav.*, vol. 35, pp. 45–59, Nov. 2015.

[32] M. G. Rahman and M. Z. Islam, "Discretization of continuous attributes through low frequency numerical values and attribute interdependency," *Expert Syst. Appl.*, vol. 45, pp. 410–423, Mar. 2016.

[33] F. Dai, J. W. Zhang, and T. L. Lu, "Modelling and recognition of a driver's starting intentions," *Proc. Inst. Mech. Eng. D, J. Automobile Eng.*, vol. 226, no. 5, pp. 623–633, May 2012.

[34] R. Tempo, G. Calafiore, and F. Dabbene, "Statistical Learning Theory," in *Randomized Algorithms for Analysis and Control of Uncertain Systems With Applications*. New York, NY, USA: Springer, 2013, pp. 123–134.

YONGGANG LIU received the B.S. and Ph.D. degrees in automotive engineering from Chongqing University, Chongqing, China, in 2004 and 2010, respectively. He was a joint Ph.D. student and a Research Scholar with the University of Michigan-Dearborn, MI, USA, from 2007 to 2009.

He is currently a Professor and a Dean Assistant with the School of Automotive Engineering, Chongqing University. He is also a Committee Member of the Vehicle Control and Intelligence Society, Chinese Association of Automation (CAA). He has led more than 10 research projects, and published more than 50 research journal papers. His research interests include optimization and control of intelligent electric and hybrid vehicles, and integrated control of vehicle automatic transmission systems.

JIMING WANG received the B.S. degree in automotive engineering from the North University of China, Shanxi, China, in 2016. He is currently pursuing the M.S. degree in mechanical engineering with Chongqing University, Chongqing, China. His research interest includes intelligent control for dual clutch transmissions.

PAN ZHAO received the B.S. degree in mechanical engineering from the Hebei University of Technology, Tianjin, China, in 2017. She is currently pursuing the M.S. degree in automotive engineering with Chongqing University, Chongqing, China. Her research interest includes intelligent control for dual clutch transmissions.

DATONG QIN received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Chongqing University, Chongqing, China, in 1982, 1984, and 1993, respectively, and the joint Ph.D. degree from Tohoku University, Sendai, Japan, in 1989.

He is currently a Professor with the State Key Laboratory of Mechanical Transmissions and the School of Automotive Engineering, Chongqing University. He has conducted more than 60 projects and has published more than 200 peer-reviewed journal papers and conference proceedings. His research interests include the control and application of mechanical transmission and vehicle power transmission. He was a recipient of the Changjiang Scholars Program of China and two first prizes of Provincial-Level Scientific and Technological Progress Awards, in 2008 and 2010.

ZHENG CHEN received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in control science engineering from Northwestern Polytechnical University, Xi'an, China, in 2004, 2007, and 2012, respectively.

He was a Postdoctoral Fellow and a Research Scholar with the University of Michigan, Dearborn, MI, USA, from 2008 to 2014. He is currently a Professor with the Faculty of Transportation Engineering, Kunming University of Science and Technology, Kunming, Yunnan, China. He has conducted more than 20 projects and has published more than 80 peer-reviewed journal papers and conference proceedings. His research interests include battery management systems, battery status estimation, and energy management of hybrid electric vehicles. He was a recipient of the Yunnan Oversea High Talent Project, China, and the second place of the IEEE VTS Motor Vehicles Challenge, in 2017 and 2018.