Received October 21, 2020, accepted November 14, 2020, date of publication November 17, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3038801

An Improved Selective Ensemble Learning Method for Highway Traffic Flow State Identification

ZHANZHONG WANG, RUIJUAN CHU, MINGHANG ZHANG, XIAOCHAO WANG, AND SILIANG LUAN
Transportation College, Jilin University, Changchun 130022, China

Corresponding author: Ruijuan Chu (1908950298@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873109, and in part by the Program of Humanities and Social Sciences, Ministry of Education, under Grant 18YJA630157.

ABSTRACT Reliable and accurate real-time traffic flow state identification is crucial for an intelligent transportation system (ITS). This identification is a prerequisite for alleviating traffic congestion and improving highway operation efficiency. In this paper, we propose an improved traffic flow state identification model that is based on selective ensemble learning (SEL). First, we adopted the fuzzy C-means (FCM) clustering method to divide the traffic flow data into three main kinds of traffic flow states and obtained the parameters that correspond to each kind of traffic flow state. Second, we applied the random subspace (RS) algorithm as the ensemble method and support vector machine (SVM) model as base learners to construct the RS-SVM ensemble model for traffic flow identification. Significantly, the discrete binary particle swarm optimization (BPSO) algorithm with global optimization search ability was employed to select the classifiers obtained by the random subspace training in the ensemble system. We experimentally validated the effectiveness of the proposed BPSO–RS-SVM-SEL approach. The research results reveal that compared with other classical traffic flow state identification methods, the proposed model has a higher maximum accuracy of 98.68%. It can be seen that our model improves the classification accuracy of traffic flow state identification and the difference in the ensemble system to a certain extent.

INDEX TERMS Traffic flow state identification, fuzzy C-means clustering, random subspace algorithm, selective ensemble learning, machine learning.

I. INTRODUCTION

With the continuous expansion of highway networks and the increase in the number of vehicles in cities and on highways, the traffic environment is deteriorating and traffic congestion is worsening. Evidence from developed countries indicates that an intelligent transportation system (ITS) is the most effective way to solve traffic congestion and improve the level of traffic management [1], [2]. Highway traffic state identification is a vital component of an ITS and can significantly realize traffic management and traffic flow guidance [3]–[5]. Obtaining accurate and timely traffic state information is necessary for individual travelers and related managers. With the current explosion of traffic flow data, identifying traffic flow states with big data technology is crucial to ensure safe travel and develop a superefficient navigation design, which may help travelers make informed travel decisions and maximize the efficiency of the limited network space and time resources.

Traffic state identification refers to the use of qualitative or quantitative indicators to judge the running state of the road under certain time and space conditions to provide different references for managers and travelers. With the development of intelligent transportation systems, traffic state identification has become a hot topic in the transportation field. In recent years, a variety of effective traffic flow state identification methods have been developed. Wang *et al.* built a model of general stochastic macroscopic traffic flow and proposed an approach to the real-time estimation of the complete traffic flow state on freeway segments based on the extended Kalman filter [6]. Nagai *et al.* investigated the traffic states and jamming transitions induced by a bus

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

(slow car) in two-lane car traffic [7]. Kong *et al.* presented an information-fusion-based approach to the estimation of urban traffic states, which integrates the federated Kalman filter and evidence theory; the test results illustrated that the proposed approach can be used in urban traffic applications on a large scale [8]. Billot *et al.* combined online traffic state estimation within a Bayesian framework, particle-filtering techniques and a parameterized first-order macroscopic mode. The experimental results showed the benefits of integrating the impact of rain in traffic state estimation [9]. Minh *et al.* proposed a real-time traffic data collection policy that is based on the "3R" philosophy, which is a unique vehicle classification method, and a reasonable traffic state quantification model for traffic state estimation. The experimental results reveal the effectiveness and robustness of the proposed solutions [10]. Antoniou *et al.* presented an approach for local traffic state estimation and prediction, which exploits available (traffic and other) information and uses data-driven computational approaches [11]. Li *et al.* proposed an ensemble learning framework to appropriately combine estimation results from multiple macroscopic traffic flow models; the framework can automatically ensemble the information from each individual estimation model based on their performance during the selected regression horizon [12]. Seo *et al.* applied conservation law for reasonable aggregation of the spacing data to acquire the traffic state and proposed a method for estimating a traffic state that is based on probe vehicle data and contains the spacing and position of probe vehicles [13]. Khan *et al.* integrated connected vehicle technology with an artificial intelligence (AI) paradigm to form a continuously variable transmission (CVT)-AI method to increase the real-time roadway traffic condition assessment accuracy [14]. Seo *et al.* conducted a survey of highway traffic state methods, which is a topic that has attracted a substantial amount of attention in recent decades, presented the current state of traffic safety evaluations (TSE) research and proposed future research directions [15]. Ryu *et al.* adapted a K-Nearest Neighbor model for the application of the proposed state vector and proposed a method for constructing traffic state vectors by using mutual information. The experimental results for real-world traffic data show that the proposed method of constructing a traffic state vector provides reasonable prediction accuracy in short-term traffic prediction [16]. Su *et al.* presented a traffic state forecasting method using adaptive neighborhood selection that is based on an expansion strategy to search manifold neighbors and obtain higher precision with manifold neighbors. The results of extensive comparison experiments indicate that the proposed model can produce more accurate forecasting results than other classic algorithms [17]. Bao *et al.* proposed a multi-index fusion clustering strategy that is based on the fuzzy C-means (FCM) clustering method to improve the identification accuracy of traffic flow states [18]. Tang *et al.* proposed a hybrid method that combines clustering methods and adopted type-2 fuzzy C-means and spatiotemporal correlation to predict future traffic trends based on an artificial neural network [19]. With the development of big data technology [20], [21], machine learning [22], [23] and deep learning [24]–[26] methods have been applied for the identification of traffic conditions and have achieved excellent results [27]–[29]. Rao *et al.* proposed an interval data-based k-means clustering method for traffic state identification at urban intersections and demonstrated the effectiveness of the proposed method in traffic state identification at urban intersections [30]. Xu *et al.* proposed a novel deep learning framework and used information from adjacent links to estimate the road traffic states [31].

Accurate real-time traffic state identification of highways is an important foundation of scientific traffic management. It is very meaningful to fully utilize the information provided by detectors on a real-world highway and to carry out a study of traffic flow state identification. By the reviews and analysis of existing traffic state identifications research methods, it can be determined that research on traffic state identification are mostly based on the single machine learning method and degrade the accuracy for traffic state identification models to a certain extent. In addition, due to the ambiguity of the traffic state, there is no clear boundary between different traffic states. Quantifying and describing the traffic states with exact values are difficult. In this study, to propose a simple but efficient traffic state identification model, we innovatively combined unsupervised learning with supervised learning and use ensemble learning to improve the performance of a single machine learning model, which enhances the accuracy and robustness of traffic state identification. More specifically, we introduce the analysis method of fuzzy clustering. The FCM clustering method, which is based on unsupervised learning, is employed to adaptively classify the traffic state, and the parameters that correspond to each category of traffic state are obtained. Ensemble learning can improve the performance of a single machine learning algorithm, which enhances the accuracy and robustness of traffic state identification. The support vector machine (SVM) model is used as the base classifier. The random subspace (RS) ensemble algorithms are utilized to divide feature subsets and train the base learner SVM. The discrete binary particle swarm optimization (BPSO) algorithm is applied to select the classifiers in the ensemble system. A hybrid highway traffic state identification model that is based on FCM, BPSO and the RF-SVM selective ensemble is constructed. This paper presents the approach of using the selective ensemble learning (SEL) method, which is devoted to traffic flow state identification on a highway, with application to an experiment to verify the performance. The research findings can deliver the real-time and accurate highway traffic state, which can provide a solid and scientific decision-making basis for traffic managers. In addition, the proposed approach for traffic flow state identification can reduce the storage overhead of the model.

The remainder of this study is arranged as follows: Section II introduces the methods and relevant theories to the research; Section III introduces the proposed model for traffic state identification; Section IV verifies whether the proposed model is effective and discusses the results; and Section V concludes the study and presents future research opportunities.

II. METHODOLOGY METHOD

A. FCM CLUSTERING MODEL

Before identifying the traffic flow state, it is necessary to cluster the original traffic parameter data to obtain the corresponding traffic state metrics. The classification of the traffic state determines the effectiveness of the traffic state identification. The traffic flow state holds fuzzy characteristics. For example, people often use fuzzy descriptions, such as congested and uncongested, to distinguish traffic conditions. There is no clear boundary between different traffic states, which hinders the quantitative description of these states. Therefore, we introduce the fuzzy cluster analysis method and use the membership function to explain the fuzzy phenomenon. The purpose of clustering is to recognize traffic state groupings of a large data set to provide a more sophisticated representation of a traffic system, specifically, when the range of available data is too large. The FCM clustering model is a representative method of fuzzy clustering analysis. This model is an unsupervised clustering method that is based on objective function minimization [32]. The method attributes clustering analysis to a constrained nonlinear programming problem and obtains fuzzy partitioning and clustering of clustered data sets via optimization [33], [34]. Therefore, we use the FCM clustering model to distinguish traffic states.

The basic idea of the FCM clustering model is based on determining where the similarity among objects that are classified in the same category is the largest and determining where the similarity among different categories is the smallest. The membership degree is the degree to which the object x belongs to the set A, which is denoted by $\mu_A(x)$. Assume $X = \{x_1, x_2, \ldots, x_n\}$ is the traffic flow data set. The objective function of the FCM clustering can be expressed as

$$J_{FCM}^{m}(U, A, X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - \alpha_i \right\| \tag{1}$$

subject to

$$s.t. \begin{cases} \sum_{i=1}^{c} u_{ij} = 1; & 1 \le j \le n \\ u_{ij} \in [0, 1]; & 1 \le j \le n, 1 \le i \le c \\ 0 < \sum_{j=1}^{n} u_{ij} < n; & 1 \le i \le c \end{cases} \tag{2}$$

where U is the membership matrix of each data point and the corresponding cluster center. $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is the cluster center. μ_{ij} is the membership degree of x_j to S_i.

Fuzzy clustering is an iterative process that minimizes the objective function. In the process of iteratively solving the minimum value of J_{FCM}^{m}, the optimal membership matrix U and cluster center V are obtained by the Lagrange multiplier method and are calculated as follows:

$$u_{ij} = 1 \Big/ \left(\sum_{k=1}^{c} d_{ij}/d_{kj}\right)^{2/(m-1)} \tag{3}$$

$$v_i = \sum_{j=1}^{n} u_{ij}^{m} x_j \Big/ \sum_{j=1}^{n} u_{ij}^{m} \tag{4}$$

where $d_{ij} = \left\| x_j - \alpha_i \right\|$ and d_{ij} is the Euclidean distance between the sample x_j and the i-th cluster center α_i.

In the iterative solution, the cluster center matrix and the membership matrix are continuously updated until $\left\| V^{(k+1)} - V^{(k)} \right\| \le \varepsilon$. At this time, the objective function reaches a minimum value and the iteration ends.

B. ENSEMBLE LEARNING BASED ON THE RS METHOD

1) RS ENSEMBLE LEARNING METHOD

In machine learning, our goal is to learn a stable model that performs well in all aspects, but the actual situation is often not ideal. Sometimes we can only obtain multiple models with preferences. With ensemble learning, a better and more comprehensive model can be obtained by combining multiple models, which are also referred to as a multiclassifier system. The premise of ensemble learning is that even if a certain weak classifier obtains incorrect results, other weak classifiers can also correct the error. By combining multiple learners, ensemble learning can often obtain significantly better generalization performance than a single learner.

Ensemble learning simultaneously trains multiple classifiers and uses a certain fusion rule to integrate the output results of each classifier to obtain a method that has better output results than a single classifier. In ensemble learning, we can acquire the best classification accuracy if the classifiers are independent of each other. For the sample set, the division in the feature space creates more individual differences than the division in the sample space, which can increase the differences in individual classifier training [35]. The RS method, which is also known as attribute bagging or feature bagging, is an integration model that is based on sample feature space sampling [36], [37]. The basic principle of the RSM is elaborated as follows: first, all the features of the training sample are regarded as a large set and are randomly sampled to form multiple different feature subspaces. Second, a new training set is built, and a machine learning classification algorithm is employed to train multiple different subspaces and obtain multiple base learners. Last, the results of the final classification are obtained by fusing the diversified base classifier results using some rules (usually majority voting rules).

The specific steps for the RS method are described as follows:

(1) For the d-dimensional data set $D = \{(x_j, t_j) \,|\, 1 \le j \le m\}$, $x_j \in X \in R^d$, t_j is the classification label. Moreover, K new k-dimensional data sets $D_i = \{(P_i(x_j), t_j) \,|\, 1 \le j \le m\}$ are generated by projection,

where P_i is a random projection, and $P_i(x_1, \cdots, x_d) = (x_{\alpha 1}, \cdots, x_{\alpha k})$. By the uniform probability distribution, k-dimensional subsets are randomly selected from subset $A = \{\alpha_1, \cdots, \alpha_k\}$.

(2) The base classifier algorithm $L$is used to train the base classifier on the training set and obtain the trained base classifier $h_i, i = 1, \cdots, K$.

(3) The base classifier $h_1, \cdots, h_K$ is synthesized by the given decision rules, and the final traffic flow state classifier C is obtained.

2) BASE LEARNER OF SVM MODEL

While many studies address ensembles of weak classifiers in the RS ensemble, studies of strong classifiers are lacking. Research shows that the combination of strong classifiers with the RS algorithm, especially integration with an SVM [38], can improve the accuracy and reduce bias and the variance of classifiers [39]–[41]. In this paper, we select the SVM model as the base learner and use the RS algorithm to combine SVM classifiers. The SVM model is a kernel learning method that applies the training set to construct a hyperplane for classifying test samples [42]–[44]. In this paper, we study the freeway traffic state identification problem as the multiclassifier, and therefore, must construct appropriate multiple classifiers. Presently, multiclassifiers are mainly constructed by combining multiple classifiers. Commonly employed methods are one-versus-one (OVO) SVMs and one-versus-rest (OVR) SVMs [45]. OVO SVMs have a higher classification accuracy than OVR SVMs and have a low computational complexity. Therefore, we choose OVO SVMs to train samples.

C. SEL BASED ON THE DISCRETE BPSO MODEL

The RS ensemble model is an approach to increase the identification accuracy by combining the results from multiple classifiers and assigns each base classifier the same weight. However, in the RS ensemble learning system, some base classifiers contribute to the ensemble system, while other classifiers provide minimal or no contribution. If all the base classifiers are ensembled, the result may not be ideal, and its computation complexity may even increase [46]. Therefore, how to select the base learner to make a greater contribution to ensemble learning has become an important research topic. A better selection strategy and improvement in the speed of the algorithm needs more research.

1) DISCRETE BPSO MODEL

This paper proposed a selective RS ensemble learning model that is optimized by the discrete BPSO algorithm to reduce time and memory overhead. The BPSO optimization model is a discretization method on a continuous space that is based on the PSO algorithm. Compared with other optimization algorithms, the BPSO algorithm has a global optimization ability [47]. In addition, the parameters of the BPSO algorithm are simple, and its convergence speed is faster than other algorithms.

In the BPSO model, each particle has its own position and velocity. In the d-dimensional search space, the value of all position component particles x_{ij} is 0 or 1, and the velocity component v_{ij} is the probability of $x_{ij} = 1$.

The velocity update is calculated as follows:

$$v_{ij}(t+1) = \omega \cdot v_{ij}(t) + c_1 \cdot r_1 \cdot \left(pbest_{ij}(t) - x_{ij}(t)\right) + c_2 \cdot r_2 \cdot \left(gbest_{ij}(t) - x_{ij}(t)\right) \tag{5}$$

The position update for the particle swarm is described as follows:

The sigmoid function is used to map the velocity to the interval (0,1] as the probability, which is the probability that the particle assumes a value of 1 in the next iteration:

$$s\left(v_{ij}\right) = \frac{1}{1 + \exp\left(-v_{ij}\right)} \tag{6}$$

$$x_{ij} = \begin{cases} 1 & \textit{if rand}\ () \le s\left(v_{ij}\right) \\ 0 & \textit{otherwise} \end{cases} \tag{7}$$

where $s\left(v_{ij}\right)$ is the probability of the position $x_{ij} = 1$ and *rand* () is a random number that is generated randomly from the interval [0,1]. $v_{\max}$ is the value of the maximum speed to limit the v_{ij} range, that is, $v_{ij} \in [-v_{\max}, v_{\max}]$, to avoid a value of $s\left(v_{ij}\right)$ that is too close to 0 or 1. The range of velocities limits the probability that the position is 0 or 1.

2) SEL METHOD OPTIMIZED BY THE BPSO MODEL

In this research, we apply the BPSO model to improve the RS ensemble learning method, and the RS-SEL model, which is optimized by the BPSO algorithm, is proposed. Determining the particle dimension and actual fitness value (objective function) is the key to the BPSO-RS-SEL method. For the BPSO-RS-SEL model, a particle represents an alternative to the base learner. A particle is an n-dimensional vector with n base classifiers and a value of 0 or 1, which corresponds to the base learner. For example, there are K base learners, and accordingly, the particle is a K-dimensional vector. If the k-th component of a particle is 1, then the k-th base classifier is selected. Conversely, if k is 0, then the k-th base classifier is not selected. The BPSO algorithm performs a global search according to a fitness function.

The RS-SEL model, which is optimized by the BPSO algorithm model, is constructed. The implementation steps are described as follows:

(1) Randomly generate T feature subsets $D_1, D_2, \cdots, D_T$ from the data set D via the self-help method;

(2) Use the SVM to train the data set in the corresponding eigenvector space and obtain the trained base classifier $h_1, h_2, \cdots, h_T$;

(3) Use BPSO to optimize and select the base classifiers $h_1, h_2, \cdots, h_T$ and obtain the new set of classifiers $h_1, h_2, \cdots, h_K$. In the new set of classifiers, the number of base classifiers is reduced from T to K;

(4) The base classifier $h_1, h_2, \cdots, h_K$ is synthesized by the given decision rules, and the final traffic flow state classifier C is obtained.

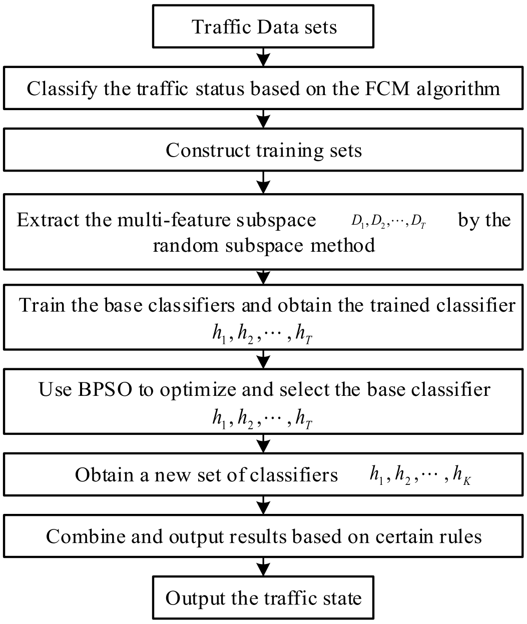

Traffic Data sets
Classify the traffic status based on the FCM algorithm
Construct training sets
Extract the multi-feature subspace $D_1, D_2, \cdots, D_T$ by the random subspace method
Train the base classifiers and obtain the trained classifier $h_1, h_2, \cdots, h_T$
Use BPSO to optimize and select the base classifier $h_1, h_2, \cdots, h_T$
Obtain a new set of classifiers $h_1, h_2, \cdots, h_K$
Combine and output results based on certain rules
Output the traffic state

FIGURE 1. FCM-BPSO-RS-SVM model.

III. HIGHWAY TRAFFIC STATE IDENTIFICATION MODEL

A. PROPOSED HIGHWAY TRAFFIC STATE IDENTIFICATION MODEL

Ensemble learning is a research hotspot in machine learning that can improve the generalization performance of a classification algorithm. In this study, we established a BPSO-RS-SEL model that is based on FCM and SVM methods to identify the highway traffic states. The core of the proposed model is clustering and classification. First, the original traffic data are clustered by the FCM clustering model to three traffic states, and their cluster centers can be obtained. Second, the SVM model is considered a base learner, and the RS ensemble learning model is applied to obtain the classification results by constructing many learners to vote on the classification results. Note that we use the BPSO model to improve the RS ensemble learning model and obtain the BPSO-RS selective ensemble method. Figure 1 shows the structure of the proposed model. The specific implementation steps are described as follow:

(1) Original data clustering. In this research, we select the traffic flow base parameters of volume, speed, and occupancy as characteristic variables for traffic flow state identification. By the FCM clustering model, the normalized traffic characteristic variable data are clustered into three categories that corresponds to three traffic states: smooth, slow, and congested. Accordingly, the numerical values of traffic flow parameters that correspond to different traffic states are obtained.

(2) Construct the training and testing sets. The training and test sets are constructed by the traffic state feature variable data obtained by the FCM clustering model. The input of the proposed model is the characteristic variable of the traffic state, and the output is the corresponding traffic state type.

(3) Train the proposed BPSO selective SVM ensemble model based on the RS method. The multifeature subspace of the original training set is extracted by the RS ensemble learning method, and the SVM base learners are trained. The BPSO algorithm is employed to select and optimize the base classifiers. The majority vote fusion method is utilized to fuse the results of the selected base classifiers that are optimized by the BPSO model to obtain the final classification results.

B. PERFORMANCE CRITERIA

In the study, we use two important criteria, the identification accuracy and difference degree of the ensemble system, to measure the performance of the proposed model.

1) ACCURACY

The model accuracy is the proportion of correctly identified states. A larger accuracy value indicates that the method is more sensitive to the traffic state and can detect the traffic states more accurately. The accuracy is computed by the following formula:

$$\begin{aligned} &Accurary \\ &\quad = \frac{number\ of\ instances\ correctly\ classified}{total\ number\ of\ instances} \times 100\% \end{aligned} \tag{8}$$

2) DIFFERENCE DEGREE OF THE ENSEMBLE SYSTEM

The difference measure is a unique evaluation standard in ensemble learning. The larger is the difference, the stronger is the generalization ability of the system. The most common approaches in the literature were employed Q-statistic (Q), Correlation coefficient (ρ), Disagreement (dis) and Double Fault Measure (DF). The following necessary related symbols must be introduced: T indicates the base learners; C_i with $C_j\,(i, j = 1, 2, \ldots, T,\ i \neq j)$ are two different classifiers; $N^{11}\left(N^{00}\right)$ indicates the number of test samples that classifiers C_i and C_j can correctly (incorrectly) classify; and $N^{10}\left(N^{01}\right)$ is the number of test samples that can be correctly classified by classifier $C_i\left(C_j\right)$, while classifier $C_j\,(C_i)$ is incorrectly classified. The total number of samples is given by $N = N^{11} + N^{00} + N^{10} + N^{01}$.

a: Q-STATISTIC

For a sample, if two base learners have the same performance for the same sample, that is, they are simultaneously correctly classified or misclassified, $N^{10} = N^{01} = 0$, that is, $Q_{ij} = 1$, the difference between them is the lowest. Otherwise, if two base learners have different classification results for every sample, $N^{11} = N^{00} = 0$, that is, $Q_{ij} = -1$. In this case, the difference is the highest. Formally,

$$Q_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}} \tag{9}$$

b: CORRELATION COEFFICIENT ρ

If $\rho = 1$, the difference is the lowest; otherwise, $\rho = -1$ is the highest difference. Mathematically,

$$\rho = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{\left(N^{11}+N^{10}\right)\left(N^{01}+N^{00}\right)\left(N^{11}+N^{01}\right)\left(N^{10}+N^{00}\right)}} \tag{10}$$

c: THE DISAGREEMENT dis

dis focuses on samples with different classification results from the two classifiers. The larger the number of samples is, the higher the difference degree is. *dis* can be computed as

$$dis_{ij} = \left(N^{10} + N^{01}\right) \Big/ N \tag{11}$$

d: DOUBLE FAULT MEASURE DF:

DF concerns the classification errors of classifiers C_i and C_j. If each sample is misclassified by the two classifiers, the accuracy and difference of the ensemble system are the lowest. Mathematically, this situation is expressed as follows:

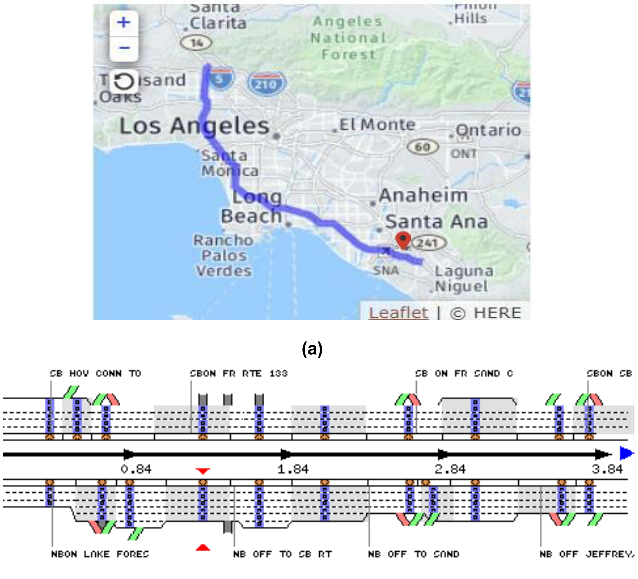
$$DF_{ij} = \frac{N^{00}}{N} \tag{12}$$

IV. EXPERIMENTAL VERIFICATION

To evaluate the performance of the proposed method, extensive experiments are performed. For comparison, the RS-SVM model, SVM method and KNN method are also utilized to perform the experiments in similar application scenarios, in which the proposed approach demonstrates the effectiveness and merit in comparison with existing approaches.

A. EXPERIMENTAL DATA DESCRIPTION

Extensive experiments were performed to quantitatively verify the performance of our proposed method. For this study, we adopted data from the California Department of Transportation Caltrans Performance Measurement System (PeMS) that was collected from detector No. VDS-1209092 on the I405-N freeway in the city of Irvine. The PeMS can collect, filter, process, aggregate, and examine traffic data in real time. We utilized the datasets collected from May 1 to May 25, 2019 in 5-min intervals to train and modify the identification models. The datasets, which consist of a total of 7200 data points, contain information such as traffic flow, speed, and occupancy rate. The datasets were divided into two parts. The first part (May 1, 2019, to May 20, 2019) was used as training data to train the proposed model. The second part, from May 21, 2019, to May 25, 2019, was used as testing datasets. Figure 2 illustrates the data collection location and provides relevant details. Figure 3 reveals the raw traffic flow, speed, and occupancy rate data from May 1 to May 25, 2019.



(a)

(b)

FIGURE 2. TI405-N freeway position location of observation loop detector No. VDS-1209092: (a) I405-N freeway position location; and (b) location of loop detector No. VDS-1209092.

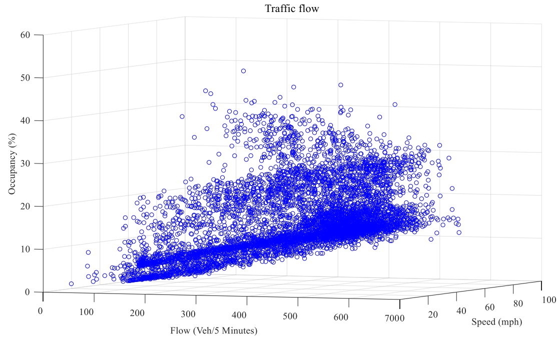


FIGURE 3. Raw traffic data of VDS-1209092 in 5-min intervals from May 1 to May 25, 2019.

TABLE 1. Cluster centers determined by FCM clustering.

Traffic states	Flow (veh/5-min)	Speed (mph)	Occupancy (%)
smooth	52.21	66.87	4.39
slow	272.92	54.09	13.35
congested	455.64	36.16	18.37

B. HIGHWAY TRAFFIC STATE CLUSTERING BASED ON THE FCM CLUSTERING MODEL

According to the three-phase theory, the traffic flow states can be divided into three types: smooth, slow, and congested. The FCM clustering model is applied to cluster the sample traffic flow data. The cluster centers of the three traffic flow states are shown in Table 1. The traffic parameters of different traffic states are quite different, which conforms to the operation characteristics of the traffic flow in different states.

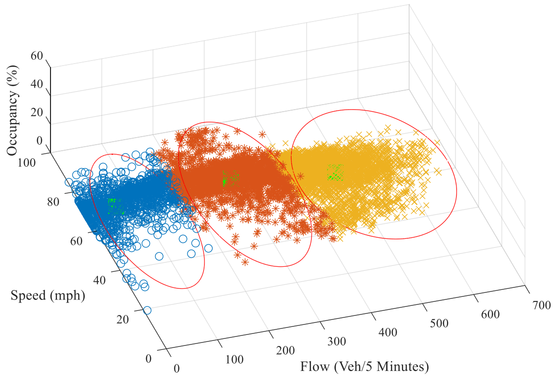


FIGURE 4. Traffic state identification results based on the FCM clustering model.

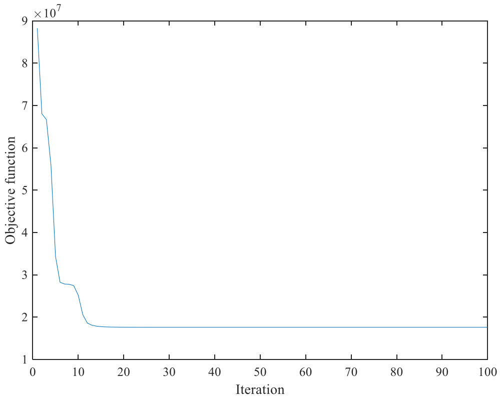


FIGURE 5. Iteration curve of the FCM clustering model.

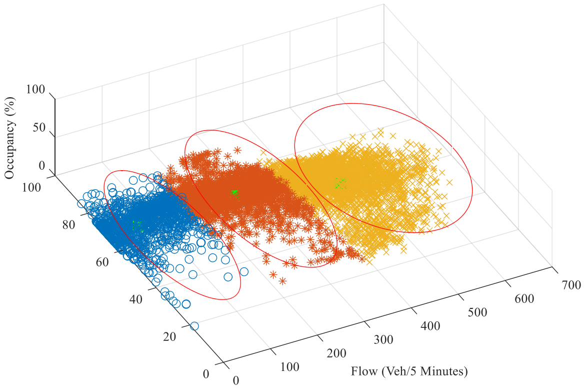


FIGURE 6. Traffic state identification results based on the k-means clustering model.

The clustering results are displayed in Figure 4. The relationship between the number of iterations and the objective function value of the FCM clustering model is shown in Figure 5. To confirm the performance and superiority of the FCM clustering model, the classical k-means clustering model is employed to cluster the same highway traffic flow experimental data[48]. The clustering results are displayed in Figure 6. The boundaries between different traffic flow states are unambiguous, and the data overlaps are substantially reduced in Figure 4, which indicates the effectiveness of the FCM clustering model.

To quantitatively compare the performance of the FCM clustering model and k-means clustering model, we apply the Davies-Bouldin Index (DBI), which is an internal evaluation index of clustering effectiveness, to measure the clustering effect [49]. The smaller is the DBI, the smaller is the in-class distance, the greater is the between-class distance, and the better is the clustering effect. By calculating the DBI of the FCM clustering model and k-means clustering model, we can determine that the DBI value of the FCM clustering model is 0.59 and the DBI value of the k-means clustering model is 0.87. The results suggest that the FCM clustering model is significantly better than the k-means clustering model.

Via the FCM clustering model, traffic is divided into three categories, namely, smooth, slow, and congested. When the traffic flow is in the smooth state, the traffic speed is very high, and the driver can control the vehicle at a speed near the speed limit. In the slow traffic state, the amount of road traffic increases, and the mutual interference between vehicles increases. Furthermore, the speed of the traffic flow decreases, and the phenomenon of car following begins to appear. In congested traffic conditions, as the road traffic flow increases, the interference between vehicles intensifies and becomes increasingly intense and the instability of the road traffic gradually deteriorates.

C. HIGHWAY TRAFFIC STATE IDENTIFICATION RESULTS AND ANALYSIS

After the original traffic data are clustered to three traffic states via the FCM clustering model, the SVM model is considered a base learner, and the RS ensemble learning model is applied to identify the traffic flow states.

Specifically, for the base learners of the SVM models, we use the Gaussian radial basis function (RBF) as the kernel function. We adopt 10-fold cross-validation to train the RS-SVM model. Two important parameters are considered in the RS-SVM model: the subspace dimension and the number of subspaces (i.e., number of learners). We set the subspace dimension to 3. The relationship between the traffic state classification accuracy and the number of subspaces is shown in Figure 7. Figure 7 shows that as the number of subspaces increases, the classification accuracy improves. However, the more the base classifiers are ensembled, the more the computational cost of the system will increase. When the number of subspaces is 30, the classification accuracy of the model reaches the maximum. With an increase in the number of subspaces, the classification accuracy undergoes minimal change. Therefore, the number of base classifiers is chosen to be 30, and we achieve 97.78% classification accuracy.

However, in the RS-SVM ensemble learning system, some base classifiers contribute to the ensemble system, while other classifiers provide minimal or no contribution among the 30 base classifiers. If all the base classifiers are ensembled, the result may not be ideal, and its computation complexity may even increase. Therefore, we use the BPSO

TABLE 2. Description of the significance parameter for comparison models.

Model	Main parameters	Classification accuracy
Proposed BPSO-RS-SVM-SEL model	number of learners=30 learning rate=0.1	98.68%
RS-SVM ensemble learning model	number of learners=30 learning rate=0.1	97.78%
SVM model	Gaussian RBF kernel function parameter=2.2 error penalty parameter=10	94.72%
RS-BP ensemble learning model	number of learners=30 learning rate=0.1	93.40%
RS-KNN ensemble learning model	number of learners=30 learning rate=0.1	93.96%
Bagging-SVM ensemble learning model	number of learners=30 learning rate=0.1	95.56%
AdaBoost-SVM ensemble learning model	number of learners=30 learning rate=0.1	96.39%
KNN model	number of neighbors=50 distance metric: Euclidean distance weight=equal	92.78%
BP neural network model	learning rate=0.1 number of hidden nodes=10	90.35%

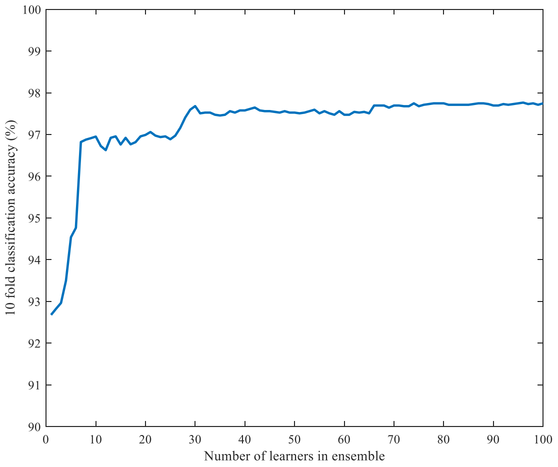


FIGURE 7. TRelationship between traffic state classification accuracy and the number of subspaces.

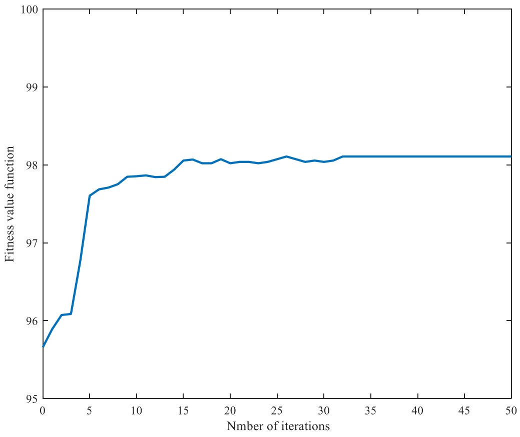


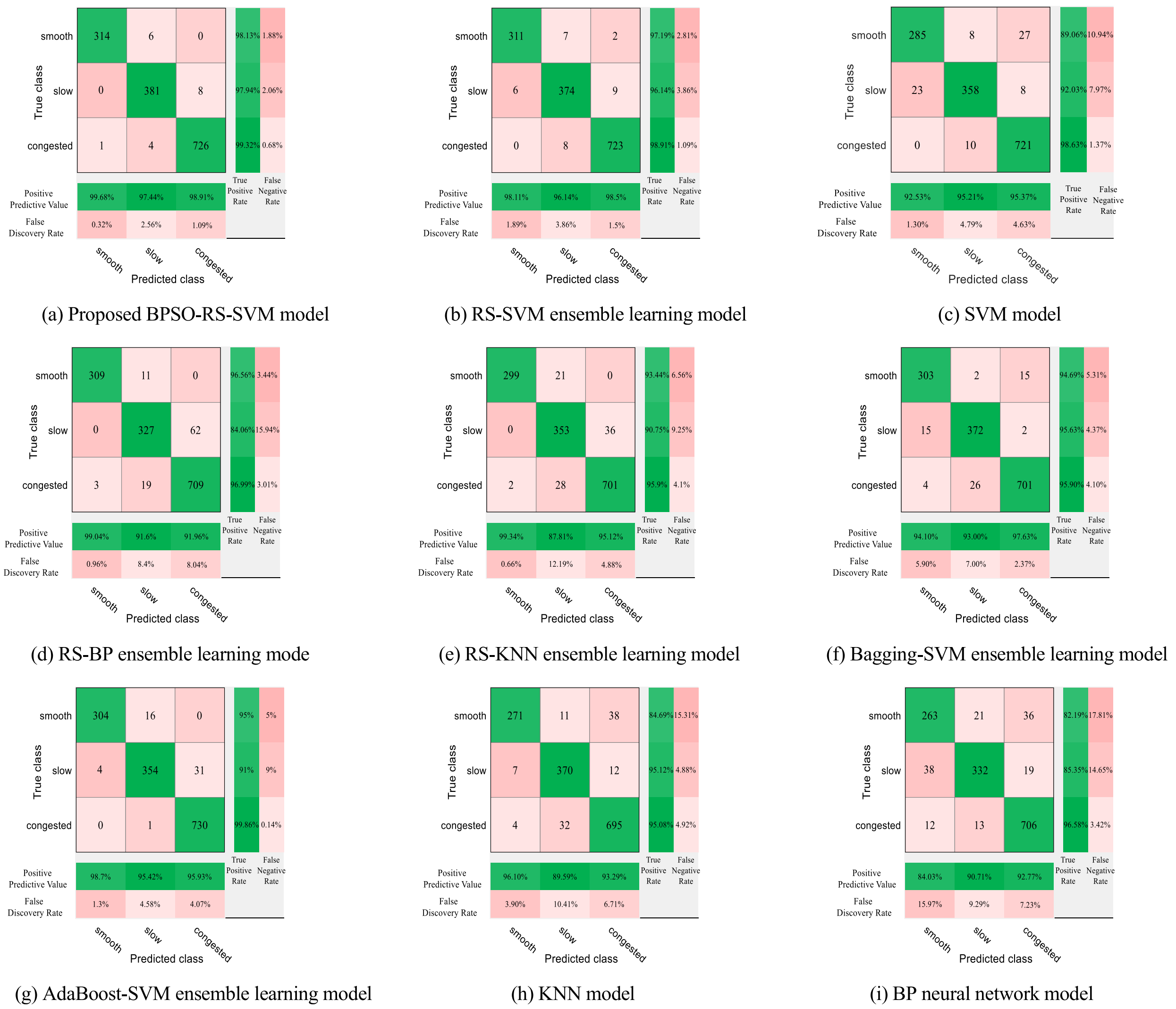
FIGURE 8. Iteration curve.

model to optimize the results obtained by the RS-SVM ensemble learning model. Reserving typical samples and reducing the number of subspaces, the generalization performance and training efficiency of the classifier are guaranteed. For the BPSO model, the accuracy of the system classification is used as the fitness value function, and the number of iterations is 50. We also set the learning factor to $c_1 = c_2 = 2$, and the inertia weight is set to $w\_\min = 0.1$, $w\_\max = 0.6$. The relationship between the number of iterations and the fitness value in the optimization process is shown in Figure 8. We obtain the number of classifiers, which is significantly reduced from 30 to 12, and 18 classifiers are reduced after optimization by the BPSO algorithm. More importantly, the improved model BPSO-RS-SVM model has the classification precision of 98.68%, which is better than the RS-SVM model, which has a classification precision of 97.78%. The experimental results show that the proposed BPSO-RS-SVM-SEL model effectively improves the precision of classification and obviously reduces the complexity in training.

1) COMPARISON MODELS

To confirm the performance and superiority of the proposed BPSO–RS-SVM-SEL model for highway traffic state identification, the same highway traffic flow experimental data were employed for modeling. A total of 8 alternative forecasting models were constructed for a comparative analysis with the proposed model. Table 2 shows the instructions of the comparison models and the classification accuracy.

The performances of these models are also compared by means of a confusion matrix, as presented in Figure 9(a)-9(i). In this research, we use the confusion matrix plot to understand how the classifier performed in each class. The confusion matrix can help to identify the areas where the classifier has performed poorly. For the confusion matrix plot, the rows show the true class, and the columns show the predicted class.



(a) Proposed BPSO-RS-SVM model

(b) RS-SVM ensemble learning model

(c) SVM model

(d) RS-BP ensemble learning mode

(e) RS-KNN ensemble learning model

(f) Bagging-SVM ensemble learning model

(g) AdaBoost-SVM ensemble learning model

(h) KNN model

(i) BP neural network model

FIGURE 9. Confusion matrix for accuracy comparison of each model.

The diagonal cells show where the true class and predicted class match. If these cells are green, the classifier has performed well and correctly classified the observations of this true class.

Figure 9 and Table 2 show that the proposed model has the highest classification accuracy of 98.68%, which indicates that the proposed model can effectively improve the classification precision of the traffic states. The detailed analysis and discussion are presented as follows.

(1) Compare the RS-SVM model to the proposed model (BPSO-RS-SVM model). As shown in Figure 9(a) and Figure 9(b), the SEL model can effectively increase the accuracy and obviously reduce the complexity in training.

(2) In Figure 9(b), 9(d), and 9(e), we compare the RS-BP ensemble learning model and RS-KNN ensemble learning model with the RS-SVM ensemble learning model and determine that the SVM classification model is significantly better than the BP and KNN models. This finding suggests that the SVM has a higher accuracy than that of the BP and KNN models and has broad application prospects in highway traffic state identification.

(3). In Figure 9(b), 9(e) and 9(f), a comparison of the RS-SVM ensemble learning model, Bagging-SVM ensemble learning model and AdaBoost-SVM ensemble learning model reveals that the integration effect based on the RS-SEL model is better than that of the AdaBoost and Bagging traffic flow state learning model with the highest classification accuracy.

From Figure 9 and Table 2, we can determine that the BPSO-RS-SVM model has the highest accuracy rate for traffic state identification of all the models, which is 98.68%, followed by the RS-SVM model. This finding indicates that the SEL model can improve the accuracy rate. We determine that the performance of the RS-SVM method is much better than that of the SVM model, which indicates that ensemble learning can improve the performance of a single model.

TABLE 3. Performance comparisons of the BPSO-RS-SVM and RS-SVM models.

Model	Q-statistic	ρ	*dis*	*DF*
BPSO-RS-SVM	-0.017	-0.011	0.471	0.403
RS-SVM	0.029	0.022	0.462	0.416

2) DIFFERENCE DEGREE OF ENSEMBLE SYSTEM ANALYSIS

The key in ensemble learning is to effectively generate individual learners with strong generalization ability and great diversity. This paper employed Q-statistic, ρ, *dis*, and DF statistics to measure the diversity in the proposed BPSO-RS-SVM SEL model. Specifically, the proposed BPSO-RS-SVM-SEL model was compared with the RS-SVM ensemble learning model. The results are shown in Table 3; these results indicate that the proposed method obtains the best performance, because its values of Q-statistic, ρ, and DF are lower. According to the calculation principle of the Q-statistic, ρ, and DF, the smaller is the value, the higher is the difference degree of the ensemble learning system. The values of the Q-statistic and ρ do not exceed 0 and are near -1. In addition, we discover that the *dis* of the proposed method has a much larger value than that of the RS-SVM ensemble learning model. Obviously, the BPSO-RS-SVM SEL model achieves an effectiveness that is significantly better than that of the RS-SVM model, and the robustness of the proposed method is the best among the models. The proposed method can improve the degree of the system difference and enhance the model generalization ability.

V. CONCLUSION

In this paper, we proposed an improved SEL model for highway traffic flow state identification. The model combines unsupervised learning with supervised learning to effectively improve the accuracy of classification and reduces the complexity in training. First, the FCM clustering method was employed to divide the original traffic flow data into three kinds of traffic flow states and obtained the parameters that correspond to each kind of traffic state. Second, we applied the RS algorithm as the ensemble method and SVM model as the base learner to construct the RS-SVM ensemble model for traffic flow identification. Significantly, the BPSO algorithm with the ability of a global optimization search was used to select the classifiers obtained by the random subspace training in the ensemble system. The proposed method was tested on the collected data via California's Freeway Performance Measurement System and compared with several other models. By analyzing the identification results of different comparison models, it can be determined that the number of base learners is reduced from 30 to 12 in the ensemble learning system. Compared with other classical traffic state identification methods, the proposed model has a higher maximum accuracy of 98.68%. We showed that the model proposed in this paper is superior to other classification models in two aspects: 1. The FCM clustering model was utilized to divide the original traffic flow data, which reduced the identification error compared with the K-means clustering model; 2. The proposed BPSO-RS-SVM-SEL model can improve the identification accuracy of traffic state identification and the difference in the traffic flow state system to a certain extent. In future studies, we plan to analyze a highway traffic flow identification method that considers space, weather, accidents, and other factors.

REFERENCES

[1] X. Xu, Y. Liu, W. Wang, X. Zhao, Q. Z. Sheng, Z. Wang, and B. Shi, "ITS-frame: A framework for multi-aspect analysis in the field of intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2893–2902, Aug. 2019.

[2] W. Wang, G. Tian, M. Chen, F. Tao, C. Zhang, A. AI-Ahmari, Z. Li, and Z. Jiang, "Dual-objective program and improved artificial bee colony for the optimization of energy-conscious milling parameters subject to multiple constraints," *J. Cleaner Prod.*, vol. 245, Feb. 2020, Art. no. 118714.

[3] H. Yu, N. Ji, Y. Ren, and C. Yang, "A special event-based K-nearest neighbor model for short-term traffic state prediction," *IEEE Access*, vol. 7, pp. 81717–81729, 2019.

[4] X. Ma, S. Luan, C. Ding, H. Liu, and Y. Wang, "Spatial interpolation of missing annual average daily traffic data using copula-based model," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 3, pp. 158–170, Jun. 2019.

[5] H.-T. Wu and G.-J. Horng, "Establishing an intelligent transportation system with a network security mechanism in an Internet of vehicle environment," *IEEE Access*, vol. 5, pp. 19239–19247, 2017.

[6] Y. Wang and M. Papageorgiou, "Real-time freeway traffic state estimation based on extended Kalman filter: A general approach," *Transp. Res. B, Methodol.*, vol. 39, no. 2, pp. 141–167, Feb. 2005.

[7] R. Nagai, T. Nagatani, and N. Taniguchi, "Traffic states and jamming transitions induced by a bus in two-lane traffic flow," *Phys. A, Stat. Mech. Appl.*, vol. 350, nos. 2–4, pp. 548–562, May 2005.

[8] Q.-J. Kong, Z. Li, Y. Chen, and Y. Liu, "An approach to urban traffic state estimation by fusing multisource information," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 499–511, Sep. 2009.

[9] R. Billot, N.-E.-E. Faouzi, J. Sau, and F. De Vuyst, "Integrating the impact of rain into traffic management: Online traffic state estimation using sequential Monte Carlo techniques," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2169, no. 1, pp. 141–149, Jan. 2010.

[10] Q. T. Minh and E. Kamioka, "Traffic state estimation with mobile phones based on the '3R' philosophy," *IEICE Trans. Commun.*, vol. E94-B, no. 12, pp. 3447–3458, 2011.

[11] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 89–107, Sep. 2013.

[12] L. Li, X. Chen, and L. Zhang, "Multimodel ensemble for freeway traffic state estimations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 3, pp. 1323–1336, Jun. 2014.

[13] T. Seo and T. Kusakabe, "Probe vehicle-based traffic state estimation method with spacing information and conservation law," *Transp. Res. C, Emerg. Technol.*, vol. 59, pp. 391–403, Oct. 2015.

[14] S. M. Khan, K. C. Dey, and M. Chowdhury, "Real-time traffic state estimation with connected vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1687–1699, Jul. 2017.

[15] T. Seo, A. M. Bayen, T. Kusakabe, and Y. Asakura, "Traffic state estimation on highway: A comprehensive survey," *Annu. Rev. Control*, vol. 43, pp. 128–151, 2017.

[16] U. Ryu, J. Wang, T. Kim, S. Kwak, and U. Juhyok, "Construction of traffic state vector using mutual information for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 96, pp. 55–71, Nov. 2018.

[17] Z. Su, Q. Liu, J. Lu, Y. Cai, H. Jiang, and L. Wahab, "Short-time traffic state forecasting using adaptive neighborhood selection based on expansion strategy," *IEEE Access*, vol. 6, pp. 48210–48223, 2018.

[18] D. Bao, "A multi-index fusion clustering strategy for traffic flow state identification," *IEEE Access*, vol. 7, pp. 166404–166409, 2019.

[19] J. Tang, L. Li, Z. Hu, and F. Liu, "Short-term traffic flow prediction considering spatio-temporal correlation: A hybrid model combing type-2 fuzzy C-means and artificial neural network," *IEEE Access*, vol. 7, pp. 101009–101018, 2019.

[20] Z. Wang, R. Chu, W. Wu, Q. Li, Z. Cai, N. Cao, and M. Gu, "Identification and optimization models for a freight-integrated transportation corridor with line importance and freight communication capability," *IEEE Access*, vol. 7, pp. 11114–11126, 2019.

[21] G. Tian, Y. Ren, Y. Feng, M. Zhou, H. Zhang, and J. Tan, "Modeling and planning for dual-objective selective disassembly using and or graph and discrete artificial bee colony," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2456–2468, Apr. 2019.

[22] L. Li, S. He, J. Zhang, and B. Ran, "Short-term highway traffic flow prediction based on a hybrid strategy considering temporal–spatial information," *J. Adv. Transp.*, vol. 50, no. 8, pp. 2029–2040, Dec. 2016.

[23] J. An, L. Fu, M. Hu, W. Chen, and J. Zhan, "A novel fuzzy-based convolutional neural network method to traffic flow prediction with uncertain traffic accident information," *IEEE Access*, vol. 7, pp. 20708–20722, 2019.

[24] F. Basso, L. J. Basso, F. Bravo, and R. Pezoa, "Real-time crash prediction in an urban expressway using disaggregated data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 202–219, Jan. 2018.

[25] L. Mou, P. Zhao, H. Xie, and Y. Chen, "T-LSTM: A long short-term memory neural network enhanced by temporal information for traffic flow prediction," *IEEE Access*, vol. 7, pp. 98053–98060, 2019.

[26] Z. Duan, Y. Yang, K. Zhang, Y. Ni, and S. Bajgain, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31820–31827, 2018.

[27] X. Chen, X. Cai, J. Liang, and Q. Liu, "Ensemble learning multiple LSSVR with improved harmony search algorithm for short-term traffic flow forecasting," *IEEE Access*, vol. 6, pp. 9347–9357, 2018.

[28] Z. Jiang, X. Chen, and Y. Ouyang, "Traffic state and emission estimation for urban expressways based on heterogeneous data," *Transp. Res. D, Transp. Environ.*, vol. 53, pp. 440–453, Jun. 2017.

[29] A. Nantes, D. Ngoduy, A. Bhaskar, M. Miska, and E. Chung, "Real-time traffic state estimation in urban corridors from heterogeneous data," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 99–118, May 2016.

[30] W. Rao, J. Xia, W. Lyu, and Z. Lu, "Interval data-based k-means clustering method for traffic state identification at urban intersections," *IET Intell. Transp. Syst.*, vol. 13, no. 7, pp. 1106–1115, Jul. 2019.

[31] D. Xu, C. Wei, P. Peng, Q. Xuan, and H. Guo, "GE-GAN: A novel deep learning framework for road traffic state estimation," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102635.

[32] A. Stetco, X. J. Zeng, and J. Keane, "Fuzzy C-means plus plus: Fuzzy C-means with effective seeding initialization," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7541–7548, Nov. 30, 2015.

[33] K. Q. Zou, J. Hu, W. L. Li, and L. H. Yu, "FCM clustering based on ant algorithm and its application," *Int. J. Innov. Comput. Inf. Control*, vol. 5, no. 12b, pp. 4819–4824, Dec. 2009.

[34] P. Wang, W. Xu, Y. Jin, J. Wang, L. Li, Q. Lu, and G. Wang, "Forecasting traffic volume at a designated cross-section location on a freeway from large-regional toll collection data," *IEEE Access*, vol. 7, pp. 9057–9070, 2019.

[35] S. B. Kotsiantis, "A random subspace method that uses different instead of similar models for regression and classification problems," *Int. J. Inf. Decis. Sci.*, vol. 3, no. 2, p. 173, 2011.

[36] P. Teisseyre, R. A. Kłopotek, and J. Mielniczuk, "Random subspace method for high-dimensional regression with the r package regRSM," *Comput. Statist.*, vol. 31, no. 3, pp. 943–972, Sep. 2016.

[37] X. Li and H. Zhao, "Weighted random subspace method for high dimensional data classification," *Statist. Interface*, vol. 2, no. 2, pp. 153–159, 2009.

[38] S. Lu and Y. Liu, "Evaluation system for the sustainable development of urban transportation and ecological environment based on SVM," *J. Intell. Fuzzy Syst.*, vol. 34, no. 2, pp. 831–838, Feb. 2018.

[39] C. Zhao, H. Zhao, G. Wang, and H. Chen, "Improvement SVM classification performance of hyperspectral image using chaotic sequences in artificial bee colony," *IEEE Access*, vol. 8, pp. 73947–73956, 2020.

[40] F. Ali, P. Khan, K. Riaz, D. Kwak, T. Abuhmed, D. Park, and K. S. Kwak, "A fuzzy ontology and SVM-based Web content classification system," *IEEE Access*, vol. 5, pp. 25781–25797, 2017.

[41] J. Cao, G. Lv, C. Chang, and H. Li, "A feature selection based serial SVM ensemble classifier," *IEEE Access*, vol. 7, pp. 144516–144523, 2019.

[42] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemble SVM using spark," *IEEE Access*, vol. 6, pp. 59657–59671, 2018.

[43] X. Feng, X. Ling, H. Zheng, Z. Chen, and Y. Xu, "Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2001–2013, Jun. 2019.

[44] R. B. Sharmila, N. R. Velaga, and A. Kumar, "SVM-based hybrid approach for corridor-level travel-time estimation," *IET Intell. Transp. Syst.*, vol. 13, no. 9, pp. 1429–1439, Sep. 2019.

[45] A. Jegorowa, J. Górski, J. Kurek, and M. Kruk, "Initial study on the use of support vector machine (SVM) in tool condition monitoring in chipboard drilling," *Eur. J. Wood Wood Products*, vol. 77, no. 5, pp. 957–959, Sep. 2019.

[46] Z. Chen, L. Zhang, G. Tian, and E. A. Nasr, "Economic maintenance planning of complex systems based on discrete artificial bee colony algorithm," *IEEE Access*, vol. 8, pp. 108062–108071, 2020.

[47] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-AdaBoost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, Mar. 2016.

[48] R. K. Esfahani, F. Shahbazi, and M. Akbarzadeh, "Three-phase classification of an uninterrupted traffic flow: A k-means clustering study," *Transportmetrica B, Transp. Dyn.*, vol. 7, no. 1, pp. 546–558, Dec. 2019.

[49] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.

ZHANZHONG WANG received the B.S. degree in transportation planning and management from the Jilin University of Technology, Changchun, China, in 1986, and the M.S. degree in transportation planning and management and the Ph.D. degree in carrying tools applied engineering from Jilin University, Changchun, in 1989 and 2007, respectively. He was with the Jilin Province Transportation Administration Bureau from 1989 to 1997. He was also with Jilin Ji Transport Group Company Ltd., from 1997 to 2002. He is currently a Professor with Jilin University. His research interests include transport resources optimization technology, mainly in direction of production logistics operation, transportation economy, and national economy evaluation, integrated transportation, and traffic big data analysis.

RUIJUAN CHU is currently pursuing the Ph.D. degree with the School of Transportation, Jilin University, Changchun, China. Her research interests include integrated freight transportation corridors, data mining, machine learning, and traffic big data analysis.

MINGHANG ZHANG is currently pursuing the master's degree with the School of Transportation, Jilin University, Changchun, China. His research interests include multimodal transport, data mining, and machine learning.

XIAOCHAO WANG is currently pursuing the Ph.D. degree with the School of Transportation, Jilin University, Changchun, China. Her research interests include intelligent transportation and traffic big data analysis.

SILIANG LUAN is currently pursuing the Ph.D. degree with the School of Transportation, Jilin University, Changchun, Jilin, China. She is also a Guest Researcher in urban planning with the Department of Built Environment, Eindhoven University of Technology, Eindhoven, The Netherlands. Her main research interests include traffic safety, machine learning, travel behaviour, and traffic management.