ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20211201014622
Original scientific paper

A Novel Support Vector Machine Model of Traffic State Identification of Urban Expressway Integrating Parallel Genetic and C-Means Clustering Algorithm

Liyan ZHANG, Jian MA*, Xiaofeng LIU, Min ZHANG, Xiaoke DUAN, Zheng WANG

Abstract: The real-time discrimination of urban expressway traffic state is an important reference for traffic management departments to make decisions. In this paper, a parallel genetic fuzzy clustering algorithm is proposed to overcome the shortcomings of the fuzzy c-means clustering algorithm. A traffic state discrimination model is established by using the support vector machine, and the parameters of the support vector machine are optimized by using particle swarm optimization, network search and genetic algorithm, so as to obtain the parameter group that can make the training model reach the maximum accuracy. Finally, the model is verified by the measured data. The convergence speed and clustering efficiency of parallel genetic fuzzy clustering and original fuzzy c-means clustering are compared. The results show that each iteration can converge to the global minimum value, and the number of iterations is small, and the clustering efficiency is high, which lays a foundation for the subsequent training of SVM.

Keywords: fuzzy clustering; genetic algorithm; support vector machine; urban expressway

1 INTRODUCTION

Urban road traffic state recognition is an important part of modern intelligent traffic system, which can effectively solve the problem of traffic congestion in the city. The realization of traffic state recognition is of great significance to the development of intelligent traffic system. It not only allows the traffic department to understand the specific road traffic conditions and take control measures for the traffic-congested areas, but also provides feedback on the traffic conditions of each road section in time for people through the intelligent transportation system so as to provide reference route choices for people's travel. At the same time, the identification of road traffic status can also analyse the changes of traffic conditions in time and space, guide the construction of urban planning department's road network, and promote the perfection of urban road network.

At present, there are many algorithms for traffic state recognition, such as standard deviation method [1], double exponential smoothing [2], bayes [3] and so on. With the development of artificial intelligence technology, more and more artificial intelligence algorithms are applied to traffic state recognition. Neural networks, fuzzy algorithms and support vector machines are widely used [4-6]. Hawas [7] established a traffic incident detection model based on fuzzy theory, which can provide traffic information for travellers in practical application and effectively relieve traffic pressure. Ritchie et al. [8] applied the artificial neural network algorithm to the traffic parameter model for the first time. Stephanedes et al. [9] used neural network feedforward model for traffic state recognition, which improved the recognition accuracy to a certain extent. Borkar and Malik [10] processed the traffic characteristic parameters, and divided the specific state of the road into three different levels. The model established by SVM is used to describe and predict the traffic state, with high discrimination accuracy and low algorithm complexity.

In the domestic, Huang et al. [11] divided the traffic state into four categories: smooth, steady, congested and blocked, and put forward an algorithm to distinguish the traffic state of urban roads based on fuzzy c-means clustering. Dong et al. [12] analysed the traffic situation of the road network in a specific road section, and used FCM algorithm for cluster analysis to realize real-time identification of the road traffic status. Li et al. [13] proposed a method of highway traffic condition discrimination based on RBF neural network. The experiments results show that RBF neural network has higher prediction accuracy of average travel speed than BP neural network, and is more suitable for highway traffic condition discrimination. Wang et al. [14] extracted the traffic flow characteristics of the road sections at different time periods, and clustered them through the K-means algorithm to judge the matching degree between the actual traffic conditions and road grades of various types of roads. With the development of science and technology, classification methods have evolved, and SVM is the most common method. Dong et al. [15] divided the collected traffic flow parameters into three categories, corresponding to the three states of congestion, crowded and freedom respectively, the traffic state discrimination model is established by using SVM, through the analysis of three kinds of kernel function training, it is concluded that the radial basis kernel function model has high accuracy and good stability. Yu et al. [16] manipulated three kernel functions of SVM to distinguish the status of urban traffic, and pointed out the importance of normalized data, compared the classification results of the three kernel functions, and found that the classification effect of radial basis kernel function was the best. For the selection of categorical feature quantities, most researchers tend to choose vehicle speed, traffic flow, and occupancy rate. The ITS Research Center of the Department of Transportation Engineering of Tongji University analysed the basic traffic flow data based on the cluster analysis method, selected the traffic flow, average speed and occupancy rate as the classification feature quantities, and divided the road status into four categories: congestion flow, congestion flow, stable flow, smooth flow. Li et al. [17] used the method of fuzzy support vector machine to classify the urban traffic state, and only used the speed as the classification characteristic parameter, and divided the traffic state into five levels: smooth, basically smooth, crowded, congested, and blocked.

The above documents have provided a foundation for expressway traffic status recognition, but most of them only use clustering or classification, which has a large amount of data and high parameter dimension. If the parameters are not pre-processed, it is easy to cause problems such as large calculation, long program running time, or inaccurate classification results.

Therefore, this paper adopts a strategy of clustering before classification to build a novel support vector machine model of traffic state identification of urban expressway integrating parallel genetic and c-means clustering algorithm. In view of the shortcomings of fuzzy c-means clustering, this paper improves the fuzzy c-means clustering algorithm, inserts the fuzzy c-means clustering into the process of genetic algorithm, and proposes a parallel genetic fuzzy c-means clustering algorithm, which effectively solves the problem that the fuzzy c-means clustering algorithm is sensitive to the initial value. Because there is a large amount of calculation in the method of traffic state discrimination based on Euclidean distance, the classification model of traffic state is established by using support vector machine, and the parameters of support vector machine are optimized in three ways: grid search method, genetic algorithm and particle swarm optimization algorithm, so as to obtain the parameter group that can make the training model reach the maximum accuracy. Finally, the model is verified by the measured data. The parallel genetic fuzzy c-means clustering and the original fuzzy c-means clustering are compared according to convergence speed, sensitivity to initial value and clustering effectiveness. The results show that the selection of initial value has no effect on parallel genetic fuzzy clustering, and each iteration can converge to the global minimum. The number of iterations is very small. The clustering efficiency is also higher than the original fuzzy c-means clustering. This method provides a good foundation for the subsequent training of SVM and saves a lot of time.

2 PARALLEL GENETIC FUZZY CLUSTERING ALGORITHM

2.1 Genetic Algorithm

Genetic algorithm (GA) [18] is a random search algorithm proposed by J. H. Holland in 1962, which simulates the evolution process of organisms and follows a mechanism of heredity, variation and survival. It is a global optimization algorithm widely used in evolutionary algorithm, which provides an effective way for us to solve optimization problems. When using genetic algorithm, coding design should be carried out for specific problems to represent some solutions in the solution space of the problem, and then the optimal solution according to the principle of survival is searched through using selection, crossover and mutation to simulate the process of biological evolution. The following is the specific process of genetic algorithm:

Step 1: Code chromosomes;

Step 2: Set the evolution algebra to 0 and initialize population;

Step 3: Calculate fitness values for all individual;

Step 4: Select higher adaptability individuals to cross and mutate for producing new individuals and forming new species groups;

Step 5: Stop iteration if the termination condition is satisfied, otherwise the evolutionary algebra will increase loop steps 3-5.

There are two main factors that affect the performance of genetic algorithm: on the one hand, the fitness function will affect the convergence direction of the algorithm; on the other hand, the crossover and mutation operators, to a certain extent, will expand the target solution set, increase the search time, and affect the rapid optimization of the algorithm.

2.2 Fuzzy C-Means Clustering Algorithm

Cluster analysis is an unsupervised classification method that can divide a dataset without classification labels into clusters [19]. Fuzzy c-means clustering (FCM) is a clustering algorithm based on objective function. Dunn [20] proposed c-means algorithm for the first time, and Bezdek [21] improved it later. In fuzzy clustering, each data point belongs to a certain category to some extent, and membership degree is used to express the degree that each data point belongs to a certain cluster [22].

$X=\{x_1,x_2,\cdots,x_n\}\subset R^p$: data sample set in the feature space; R^p: the feature space; *n*: the number of samples; *c*: the number of clusters; $C=\{c_1,c_2,\cdots,c_n\}\subset R^p$: vector set of feature space consists of *c* cluster center vectors; u_{ij}: Membership degree of the sample *j* belonging to the Centre *i*; $U=\left[u_{ij}\right]$: *c*×*n* matrix [23]. The objective function of FCM algorithm is:

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|x_j-c_i\right\|^2,\sum_{j=1}^{n}u_{ij}=1,j=1,2,\cdots,n \tag{1}$$

constraint condition:

$$\begin{cases}0\le u_{ij}\le 1\\ \sum_{i=1}^{c}u_{ij}=1\end{cases} \tag{2}$$

where: *J*: the objective function; *m*: the weighted index is also called the smoothing index, $m\in[1,+\infty)$; c_i: the *i*-th cluster center point; x_j: the *j*-th sample point in the sample set.

The objective function is the sum of the distances from all sample points in the sample set to each cluster center multiplied by the membership degrees of each cluster center. The criterion of clustering is to take the minimum value of objective function, which is a constrained optimization problem about independent (*U*, *C*), and the Lagrange multiplier method can be constructed to solve this objective function.

$$\overline{J}(u_{ij},c_i,\lambda_j)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|x_j-c_i\right\|^2 +\lambda_1\left(\sum_{i=1}^{c}u_{i1}-1\right)+\lambda_2\left(\sum_{i=1}^{c}u_{i2}-1\right) +\cdots+\lambda_n\left(\sum_{i=1}^{c}u_{in}-1\right) \tag{3}$$

The first order necessary condition for obtaining the minimum value of the objective function is:

$$\frac{\partial\overline{J}}{\partial u_{ij}}=mu_{ij}^{m-1}\parallel x_j-c_i\parallel^2+\lambda_j=0 \tag{4}$$

$$\frac{\partial\overline{J}}{\partial C_{ij}}=\sum_{j=1}^{n}u_{ij}^{m}\left(x_j-c_i\right)=0 \tag{5}$$

It can be simplified as follows:

- Iterative formula of membership matrix:

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_j-c_i\right\|}{\left\|x_j-c_k\right\|}\right)^{\left(\frac{2}{m-1}\right)}} \tag{6}$$

- Iterative formula of clustering center:

$$c_i=\sum_{j=1}^{n}\frac{u_{ij}^{m}}{\sum_{j=1}^{n}u_{ij}^{m}}x_j \tag{7}$$

If the data set X, the number of clustering categories C and the fuzzy coefficient m are known, the optimal membership matrix and clustering center can be calculated by the iterative formula of membership matrix and clustering center. According to a large number of previous studies, there are three factors that affect the performance of FCM algorithm: the selection of fuzzy coefficient and initial clustering center, as well as the solution method of the objective function.

2.3 Parallel Genetic Fuzzy Clustering Algorithm

In view of the slow searching speed of genetic algorithm and the problem that FCM is easy to fall into local optimum, this paper proposes parallel genetic fuzzy clustering algorithm (Parallel Genetic Fuzzy C-Means, PGFCM). Its basic idea is to interleave the iterative formula of FCM algorithm cluster center into genetic algorithm. The Parallel genetic fuzzy clustering merges the advantages of genetic algorithm in global search and fast iteration of FCM, and the global optimal value can be found by fast convergence with a large probability. The specific flow of the algorithm is as follows:

Step 1: Input data set *A*, set population size *N*, crossover probability P_c, mutation probability P_m, maximum evolution algebra T_{max}, fuzzy coefficient *m*, cluster center number *c*, initialization cluster center $C_i^{(0)}, i=1,2,\cdots,N$, evolution algebra t = 0, enconde $C_i^{(0)}, i=1,2,\cdots,N$ to obtain the initialization population $P^{(0)}$;

Step 2: Code $C_i^{(0)}, i=1,2,\cdots,N$ to obtain the contemporary population $P^{(t)}$ and evaluate its adaptability;

Step 3: Use selection operator to select individuals for crossover and genetic operation generation $P^{(t+1)}$;

Step 4: Decode $P^{(t+1)}$, use Eq. (6) and Eq. (7) to update (*U*, *C*), and get $C_i^{(t+1)}, i=1,2,\cdots,N$;

Step 5: If the maximum number of iterations is reached or the difference of the average fitness of individuals in successive generations of the population is less than a certain threshold value, the algorithm will stop, otherwise, make $t=t+1$ and cycle Steps 2 to 5.

3 MODELING OF TRAFFIC STATUS RECOGNITION BY PARALLEL GENETIC FUZZY CLUSTERING ALGORITHM

The most common parameters to study traffic flow characteristics are traffic flow, speed, occupancy, traffic density, queue length, headway and so on [24, 25]. The classic traffic flow model can be represented by flow, speed and density, but there are some difficulties in obtaining traffic density data. Therefore, by understanding the specific conditions of the parameters used in previous studies, this paper selects three parameters commonly used: speed, flow and occupancy.

Combining the global search ability of genetic algorithm and the fast convergence of fuzzy clustering, the best clustering center is quickly obtained. According to the membership matrix output by FCM objective function, the degree of each sample point belonging to a traffic state is judged. The data set is divided into four corresponding traffic states: smooth, steady, congested and blocked.

Expressing a set of traffic flow parameters (speed, flow, occupancy) on the spatial coordinate axis, selecting several points randomly as the initial clustering center, dividing the data set into four categories and selecting four points randomly are the initial clustering centers of the four categories:

$$\begin{pmatrix} S_1\ F_1\ O_1 \\ S_2\ F_2\ O_2 \\ S_3\ F_3\ O_3 \\ S_4\ F_4\ O_4 \end{pmatrix} \tag{8}$$

where: *S*: speed; *F*: flow; *O*: occupancy. Each row represents a cluster center.

Each row represents a cluster center. Coding the selected cluster center with real value the expression form of chromosome is: $S_1F_1O_1S_2F_2O_2S_3F_3O_3S_4F_4O_4$.

According to the population size, several of the above chromosomes are generated. Evaluating the fitness of each chromosome the fitness function is:

$$f=\frac{1}{1+J},$$
$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|\begin{bmatrix}S_j\\F_j\\O_j\end{bmatrix}-\begin{bmatrix}S_i\\F_i\\O_i\end{bmatrix}\right\|^2, \tag{9}$$
$$\sum_{j=1}^{n}u_{ij}=1,$$
$$j=1,2,\cdots,n$$

The smaller the objective function value of FCM is, the larger the fitness value of the individual will be. The individuals with higher fitness will be retained for crossover and mutation, and the new generation of clustering center matrix will be decoded, and then the third generation of clustering center matrix will be calculated by the clustering center iteration formula of FCM. Cycle the above operations until the difference of average fitness is less than a certain threshold or reaches the maximum number of iterations, and output the optimal clustering center and membership matrix. According to the degree to which each sample belongs to each type of traffic state, the data set is divided into four categories.

4 DECISION ANALYSIS MODEL OF TRAFFIC STATE IDENTIFICATION

4.1 Support Vector Machine Theory

Support vector machine (SVM) [26] is based on statistical theory, which can solve the problem of classification and regression with a small number of samples quickly and accurately. In essence, SVM is mainly used to solve the problem of two classifications. However, in reality, most cases are multi-classification. In this paper, the "one-to-one" classification method is adopted to apply the concept of two classifications in SVM to other multi-classifications, and six classifiers are constructed for four traffic state categories.

SVM is developed from the optimal classification hyperplane in the case of linear separation. For the linear separable case, the optimal hyperplane is required to solve the optimal combination of (w, b) parameter. The optimization problem constructed is as follows:

$$\min_{w,b}\frac{1}{2}\|w\|^2,$$
$$\text{Subject to}\quad y_i\left(w^T\cdot x_i+b\right)\geq 1, \tag{10}$$
$$i=1,2,\cdots,n,$$

where: w is the normal vector of the hyperplane; b is the constant term of the hyperplane; x_i is the i-th sample of the input pattern; $y_i\in\{-1,+1\}$; $\|w\|$: the modulus of the vector x; n is the number of samples

In this way, the original classification problem is transformed into the problem of solving the quadratic programming.

For the linear inseparable case, Vapnik puts forward the concept of soft interval, introducing the penalty factor C and relaxation variable ξ in dealing with problems. C is mainly to adjust classification error and the weight of classification interval. ξ represents the degree of error classification, so as to achieve the purpose of compromise. Therefore, the optimization problem is modified as follows:

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i,$$
$$\text{Subject to}\quad y_i\left(w^T\bullet x_i+b\right)\geq 1-\xi_i, \tag{11}$$
$$\xi_i\geq 0, i=1,2,\cdots,n,$$

In the nonlinear case, SVM transforms the input sample into high-dimensional space. In order to avoid this situation, kernel function can be introduced into SVM, so that the original data can be transformed from linear indivisible to linear separable.

In the identification of urban traffic state, combined with the actual situation of traffic, define the observation matrix x = [S F O], select the appropriate kernel function, and bring the observation matrix x into the SVM discriminant function to achieve state classification.

The kernel functions commonly used at present are:

- Linear kernel function: $k\left(x,x_i\right)=\left(x\cdot x_i\right)$. This function is a kernelless function parameter, but its application is very narrow.
- Polynomial kernel function: $k\left(x,x_i\right)=\left(s\left(x\cdot x_i\right)+c\right)^d$, where s, c and d are all parameters
- Radial basis kernel function: $k\left(x,x_i\right)=\exp\left\{-\left|x-x_i\right|^2/2g^2\right\}$.

  There is only one parameter g in this function.
- Sigmoid kernel function: $k\left(x,x_i\right)=\tanh\left(s\left(x\cdot x_i\right)+c\right)$.

  It includes two parameters: s and c.

Among them, radial basis kernel function has a wide range of applications, relatively few parameters and its solution is less restricted by constraints, which can make the optimization of parameters simpler [27]. Therefore, this paper introduces radial basis kernel function in support vector machine.

4.2 Modelling of Traffic State Classification Model Based on SVM

Essentially, support vector machines are mainly used to solve binary classification problems. However, most of the real situations are multi-classification situations, and the mathematical principles of binary classification cannot support other types of classification problems [28]. Multi-classification problems can be divided into two types: "one-to-one" and "one-to-many".

(1) one-to-one

When using this classification method, it is necessary to build a classifier among all categories. That is to say, two categories are randomly selected from all categories to construct a classifier, so that $n(n-1)/2$ two-class classifier can be generated [29] where n is the size of training samples. In the process of analysing the samples, the two classifiers where they are located are also classified by the voting method, and finally the votes are counted for each sample. The category with the most votes is the category of the sample.

(2) one-to-many

Using this method, the samples need to be divided into two categories: the first category and the other categories, and then all the remaining samples are classified [30]. Repeat this sorting operation until sorting is complete. From another point of view, the number of support vector machines should be the same as the number of categories. When classifying unknown samples, the category with the largest output value of the decision function is the category to which the sample belongs.

Using the "one-to-one" classification method, this paper needs to construct 6 classifiers, and using the "one-to-many" method it only needs to construct 4 classifiers. Using the "one-to-many" method requires training 1440 samples each time, and the negative class samples for each training are much larger than the positive class samples. The problem of unbalanced samples may lead to inaccurate classification. For discriminating new sample points，all models need to be retrained. The "one-to-one" method only needs to train a few hundred samples at a time. Although there are more classifiers constructed in the "one-to-one" method, the "one-to-one" method considering the overall training time is faster, and there is no phenomenon that some sample points are inseparable like the "one-to-many" method. Combined with the research questions, this paper adopts a "one-to-one" approach to classify and model traffic states.

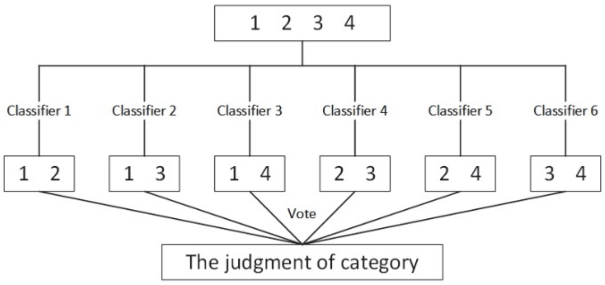

Figure 1 SVM multi-classification process diagram

Four categories of the traffic status are set in this paper. First, a part of the four categories of traffic status data sets obtained based on clustering is selected as categories 1, 2, 3, and 4 for training to obtain the parameters of the support vector machine model. The new samples are tested using the parameters. For example, the newly added sample A is classified by 6 classifiers, and the classification status of each classifier is recorded. If a class belongs to a certain class, one vote is counted for a certain class, and then the traffic status of the newly added sample point A is judged by $\mathrm{Max}_{\mathrm{vote}}\{1\ 2\ 3\ 4\}$. The process is shown in Fig. 1.

4.3 Traffic State Discrimination Model Based on Improved PGFCM and SVM

FCM algorithm needs a lot of historical data as the basis, so if only this algorithm is used, the results will be lack of timeliness. SVM model necessitates data and corresponding state labels as the basis to ensure the accuracy of model classification, but only using SVM algorithm will lack data support. Therefore, the above two algorithms should be applied synthetically in practical application. Based on the multi-classification model of SVM and the improved fuzzy c-means clustering algorithm, the real-time traffic state discrimination model of urban expressway is constructed. The flow chart is shown in Fig. 2.

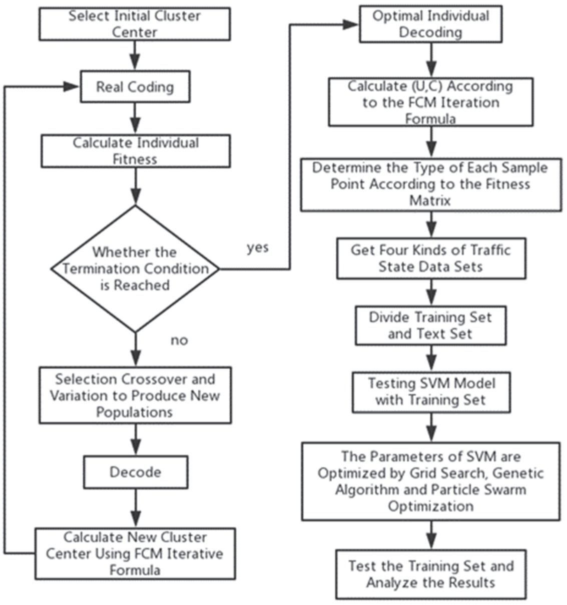
Select Initial Cluster Center
Real Coding
Calculate Individual Fitness
Whether the Termination Condition is Reached
yes
no
Selection Crossover and Variation to Produce New Populations
Decode
Calculate New Cluster Center Using FCM Iterative Formula
Optimal Individual Decoding
Calculate (U,C) According to the FCM Iteration Formula
Determine the Type of Each Sample Point According to the Fitness Matrix
Get Four Kinds of Traffic State Data Sets
Divide Training Set and Text Set
Testing SVM Model with Training Set
The Parameters of SVM are Optimized by Grid Search, Genetic Algorithm and Particle Swarm Optimization
Test the Training Set and Analyze the Results

Figure 2 Flow chart of traffic state discrimination model based on PGFCM and SVM

The accuracy of SVM algorithm is mainly affected by algorithm parameters, so it is necessary to select reasonable parameters to give full play to the role of SVM.

In this paper, the value of penalty factor C, parameter g, number of support vectors L and paranoid coefficient b should be defined. Among them, C and g can only be obtained by optimization algorithm. Due to the development of intelligent algorithm, the method of parameter optimization based on intelligent algorithm has also been applied in support vector machine. The frequently-used optimized algorithms mainly include particle swarm optimization, network search, genetic algorithm and so on. However, each optimization method has obvious advantages and disadvantages. In this paper, the above methods are applied respectively in parameter optimization, and the combination of parameters with the highest classification accuracy after optimization is selected and applied to the model.

5 CASE STUDY

5.1 Data Preprocessing

The data used in this study are 1440 sets of traffic parameters (flow, speed, occupancy) for 24-hour on August 19, 2017 provided by the expressway section detector of a city in Shanghai, and the collection interval is one minute (see Tab. 1).

In order to ensure the reliability of clustering results, it is necessary to guarantee the quality of the obtained data and preprocess the traffic flow parameters. The main processing content is to detect outliers and normalize data

[20], that is, to detect whether there are outliers (missing values) in all the data firstly, normalize the data obtained, convert the three indicators detected into constants within the range of [0, 1], and improve the accuracy of SVM classification and fuzzy c-means clustering algorithm (see Tab. 2).

Table 1 Original sample table of traffic flow data

Test Time	Speed / mph	Flow / veh/min	Occupancy / %
2017/08/19 00:01	72.09	21.00	1.83
2017/08/19 00:02	69.47	17.00	1.66
2017/08/19 00:03	64.99	14.33	1.50
2017/08/19 00:04	67.87	24.00	2.92
2017/08/19 00:05	63.81	11.00	1.25
2017/08/19 00:06	69.11	26.00	2.58
2017/08/19 00:07	64.22	9.00	0.83
2017/08/19 00:08	65.43	23.00	2.41
2017/08/19 00:09	64.92	13.00	1.25
2017/08/19 00:10	67.63	22.00	2.33
2017/08/19 00:11	64.06	13.66	1.38
2017/08/19 00:12	66.60	10.00	1.08

Table 2 Data normalization results

Test Time	Speed / mph	Flow / veh/min	Occupancy / %
2017/08/19 00:01	0.1818	0.8086	0.0324
2017/08/19 00:02	0.1455	0.7759	0.0293
2017/08/19 00:03	0.1212	0.7200	0.0262
2017/08/19 00:04	0.2091	0.7559	0.0524
2017/08/19 00:05	0.0909	0.7053	0.0216
2017/08/19 00:06	0.2273	0.7714	0.0462
2017/08/19 00:07	0.0727	0.7104	0.0139
2017/08/19 00:08	0.2000	0.7255	0.0431
2017/08/19 00:09	0.1091	0.7191	0.0216
2017/08/19 00:10	0.1909	0.7529	0.0416
2017/08/19 00:11	0.1152	0.7084	0.0241
2017/08/19 00:12	0.0818	0.7400	0.0185

After data normalization, the relationship between the three parameters (flow, speed and occupancy) is shown in Fig. 3. After smoothing, a three-parameter smoothing graph is obtained, as shown in Fig. 4.

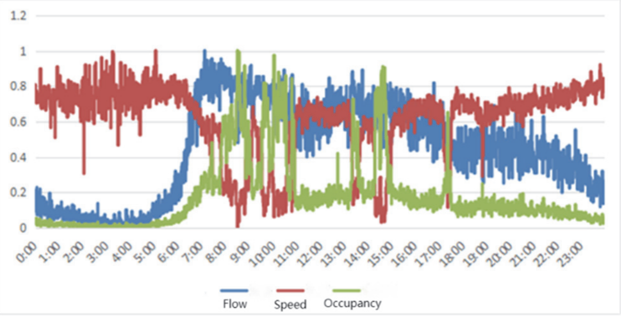

Figure 3 Normalized graph

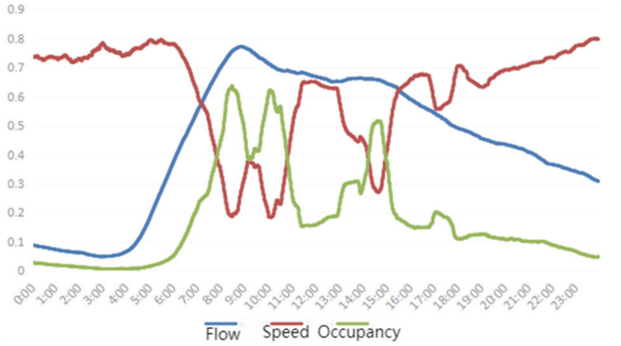

Figure 4 Smooth curve

Fig. 3 and Fig. 4 intuitively show the changing trend of flow, speed and occupancy. The change of flow during complex traffic operation periods does not clearly show the impact on speed and occupancy, while the relationship between speed and occupancy is obviously very different. The analysis of the change of traffic flow with time provides a basis to study the change of traffic state, and also confirms the feasibility of studying the problem from the perspective of flow, speed and occupancy.

5.2 Comparison of Clustering Results Between FCM and PGFCM

5.2.1 FCM Clustering

The specific operation process is as follows:

- Initialization parameters:
  $c = 4, m = 2, \varepsilon = 1 \times 10^{-5}, T_{\max} = 30, v^{[0]}$;
- Update clustering center and membership matrix;
- Termination condition: $t = 30$ or the difference between the centers of two generations is less than ε;

The clustering centers of four traffic states are obtained:

$$V^f = [v_1 v_2 v_3 v_4]^T = \begin{bmatrix} 12.782 & 64.337 & 2.065 \\ 47.267 & 63.672 & 7.223 \\ 71.443 & 52.653 & 10.142 \\ 83.641 & 24.713 & 29.937 \end{bmatrix}$$

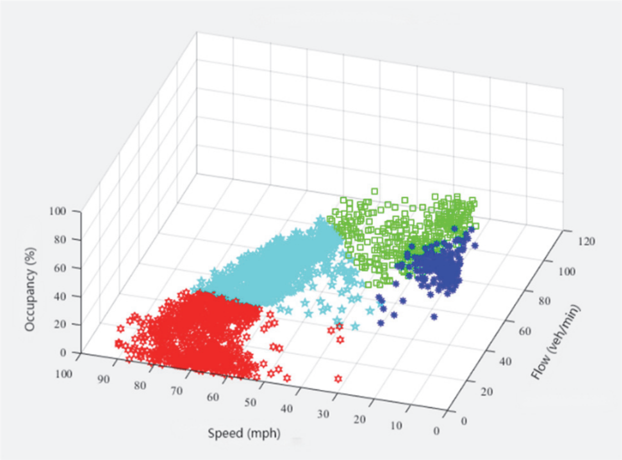

Figure 5 Spatial distribution of traffic state based on FCM

Each row of the matrix represents the clustering center of smooth, steady, congested and blocked states, the elements of each column in the matrix are traffic flow (veh/min), speed (mph) and occupancy (%). The specific distribution of these samples in state space is shown in Fig. 5.

5.2.2 PGFCM Clustering

The specific operation process is as follows:

- Parameters of genetic algorithm: population number N = 50, evolution algebra $T_{\max}$ = 30, crossover probability $p_c = 0.6$, mutation probability $p_m = 0.1$; the number of clusters $c = 4$; the fuzzy coefficient $m = 1.6$.
- Code the initial value with real value.

- Population initialization p_t; Determine the upper and lower bounds of the three parameters, and generate three random numbers in the bounds of the three parameters respectively as an initial cluster centre. In this paper, the clustering number is four, so four cluster centers are generated four times. Four randomly generated initial cluster centers are coded by real numbers to form a chromosome, and fifty chromosomes are randomly generated.
- Fitness function: $f = \dfrac{1}{J_m\left(U,V\right)}$;
- Design genetic operators: Two methods of fitness ratio algorithm and elite preservation are applied synthetically; The arithmetic crossover operator based on shortest distance gene matching is selected; Mutation operator adopts basic bit mutation;
- The second generation cluster center matrix is obtained by decoding, and the new generation cluster center matrix is solved by FCM iterative formula;
- When the number of iterations reaches the maximum or the fitness changes little or no, the operation of the algorithm is over. Otherwise, the coding, fitness evaluation and genetic operation will continue.

Thus, the clustering centers of the four traffic states obtained are as follows:

$$V^f = \left[v_1 v_2 v_3 v_4\right]^{\mathrm{T}} = \begin{bmatrix} 9.653 & 67.802 & 1.053 \\ 48.003 & 63.361 & 5.918 \\ 75.853 & 55.832 & 10.639 \\ 85.105 & 23.967 & 31.339 \end{bmatrix}$$

The specific distribution of the four categories of samples in state space is shown in Fig. 6.

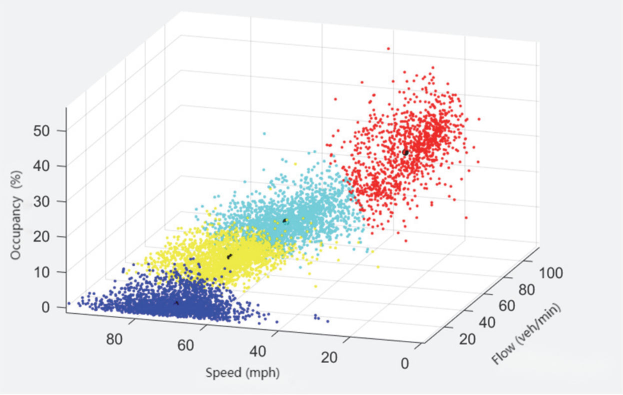

Figure 6 Spatial distribution of traffic state based on PGFCM

According to the clustering center matrix, the difference between the classes obtained by the algorithm is obvious, which means the clustering effect is excellent.

5.3 Comparative Analysis of Convergence Ability and Misjudged Rate Between PGFCM and FCM

5.3.1 Convergence Analysis

It is found in Fig. 7 and Fig. 8 that the improved PGFCM algorithm will gradually approach the optimal value after 5 iterations, and the maximum value is 50911.263.

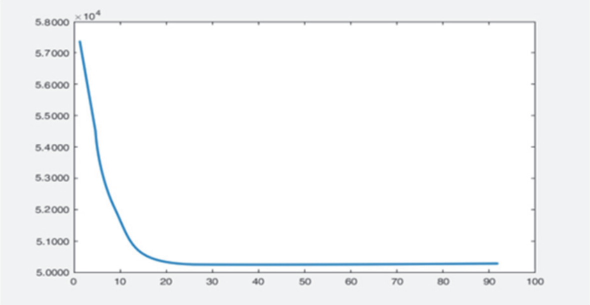

Figure 7 Convergence curve of FCM algorithm

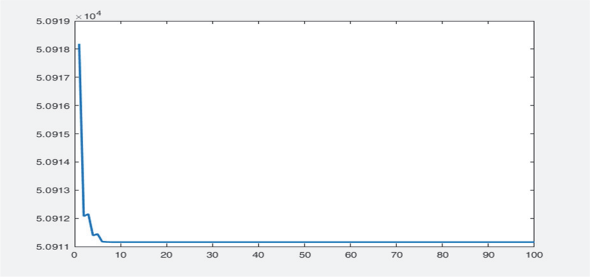

Figure 8 Convergence curve of PGFCM algorithm

However, if you adopt the initial algorithm, you need to iterate up to 20 times to approach the target value 50912.649 slowly. The convergence speed is slow, and the objective function has converged before reaching the minimum value. However, the minimum value obtained by the parallel genetic fuzzy clustering algorithm is much smaller than that obtained by FCM alone, and it does not fall into the local minimum. It can be seen that parallel genetic fuzzy clustering has obvious effect. Compared with FCM algorithm, PGFCM algorithm has obvious advantages in convergence speed and optimization ability.

5.3.2 Analysis of Misjudged Rate

The cross-estimation method of misjudged rate is used to compare and analyse the misjudged rate of PGFCM and FCM in the article. Set the sample size as N, use PGFCM and FCM to divide the data into four categories, and record the total number of samples of each category. The steps mainly include:

- Select a sample from all samples and remove it, then use the above two methods to cluster the remaining samples and record the results respectively. The cluster center and membership degree are judged by the eliminated samples to determine which category they belong to;
- Repeat the first step to remove each sample. Compare the result of clustering after elimination with the original result. If the result is different, the sample will be regarded as a misjudged sample and the sample size will be recorded. Calculate the error rate with the following formula:

$$\partial = \frac{n_1^* + n_2^* + n_3^* + n_4^*}{n_1 + n_2 + n_3 + n_3} \times 100\%, \tag{12}$$

where: n_1, n_2, n_3, n_4 are the sample sizes of the four types of original samples; $n_1^*, n_2^*, n_3^*, n_4^*$ are the sample sizes of the misjudged samples.

The results of the two clustering methods are shown in Tab. 3.

Table 3 Comparison of misjudged cross estimates

Traffic Status	Algorithm	Smooth	Steady	Congested	Blocked
Number of original samples	FCM	407	399	371	263
	PGFCM	424	378	367	271
Number of misjudged samples	FCM	20	51	79	11
	PGFCM	9	22	37	8

According to Eq. (12), the error rate of FCM is 11.2, while the error rate of PGFCM is 5.3%, which is about twice as low as that of FCM. It shows that the improved algorithm provides a good data base for the classification model.

5.4 Verification of Classification Model Based on SVM

5.4.1 Data Set Partition

60% of the data sets were randomly divided into training set and the other 40% into test set (see Tab. 4).

Table 4 Training set and test set of support vector machine model

Training Set				Test Set			
Speed	Flow	Occupancy	Label Value	Speed	Flow	Occupancy	Label Value
0.5182	0.0674	0.8367	3	0.4000	0.6843	0.1002	1
0.5424	0.0904	0.7535	3	0.3455	0.6778	0.0955	2
0.4545	0.0349	0.9045	3	0.4818	0.6537	0.1279	1
0.6455	0.4376	0.2265	1	0.5545	0.0578	0.8861	3
0.6273	0.3826	0.2666	4	0.6273	0.0839	0.8197	4
0.8182	0.3808	0.3513	4	0.6455	0.0496	0.9106	4
0.2727	0.4065	0.1387	2	0.4818	0.7462	0.1279	1
0.2545	0.6284	0.0663	2	0.4091	0.6783	0.1017	1
0.4182	0.6303	0.1094	1	0.3364	0.7921	0.0801	2

The numbers 1, 2, 3 and 4 represent states one to four, respectively, which are called labels.

5.4.2 Comparative Analysis of Training and Testing of Two Clustering Results

Based on the clustering results of the two methods obtained in the previous section, the original SVM was used to classify the two sets of clustering results, and the test and training accuracy and program running time were recorded. Support vector machine setting: using radial basis kernel function, $C = 5$, $g = 0.1$, the average results of 10 trials are shown in Tab. 5 and Tab. 6:

Table 5 PGFCM-SVM test results

Training Accuracy	Test Accuracy	Training Time	Test Time
82.64%	80.21%	7.24 s	2.47 s

Table 6 FCM-SVM test results

Training Accuracy	Test Accuracy	Training Time	Test Time
80.79%	78.13%	8.75 s	3.53 s

From Tab. 5 and Tab. 6 it can be seen that the improved PGFCM clustering results have more clear distinction between classes, and it is easier to get the classification boundary when using SVM for classification, so it has advantages in test time and test accuracy.

5.4.3 Parameter Optimization of SVM Model

(1) Grid search method

It can be seen from Fig. 9 that the result of optimization by grid search method is as follows: when $C = 5.2780000$, $g = 0.035897$, the optimal classification accuracy is 98.2638%. Because the grid search method is aimed at all parameters in the search range, the computation is huge and affected by the search step. When the search step size is small, the high precision can be obtained. If the search step size is increased, the optimal parameter combination will be skipped, and the sub-optimal result will be obtained, so the classification accuracy will be reduced. It is not particularly suitable for the problem of urban expressway traffic state discrimination, which has a large amount of data and needs to get results quickly.

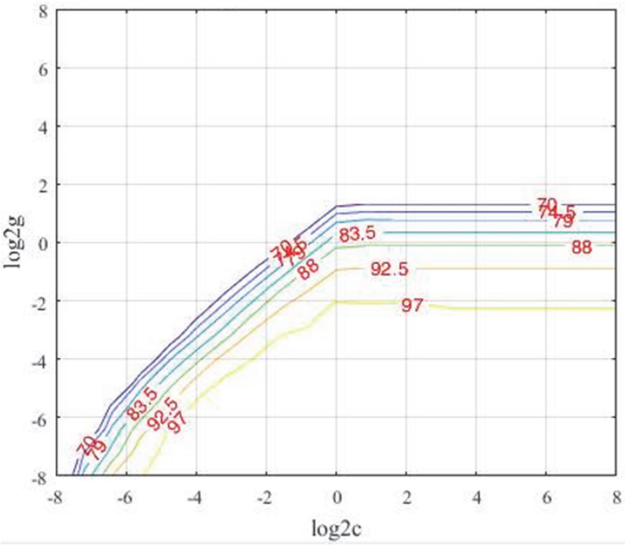

Figure 9 *C*, *g* parameter contour diagram

(2) Particle swarm optimization

Fig. 10 shows the results of particle swarm optimization. When $C = 0.5172000$, $g = 0.0100000$, the optimal classification accuracy is 97.7431%. When particle swarm optimization is used to optimize the parameters, the optimal fitness value can be achieved quickly.

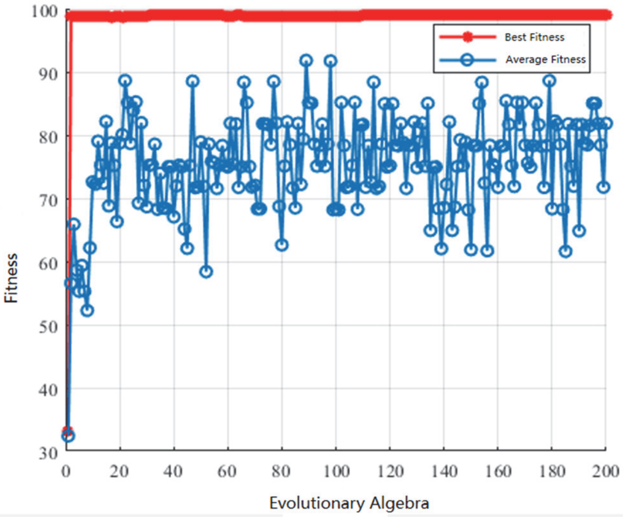

Figure 10 Iterative fitness curve of particle swarm optimization

The convergence speed is very fast in the early stage of evolution, while the convergence speed is slow in the late stage of evolution. At the same time, the convergence accuracy of the algorithm is relatively low. However, the algorithm is suitable for the problems studied in this paper.

(3) Genetic algorithm

Fig. 11 shows the optimized results of genetic algorithm. When C = 0.962920, g = 0.0038147, the best classification accuracy is 98.7269%. Due to the global search characteristics of genetic algorithm, the fitness of the initial evolution has declined. The classification accuracy of this method is the highest, and the running time of the algorithm is longer than that of PSO, which is suitable for the problems studied in this paper.

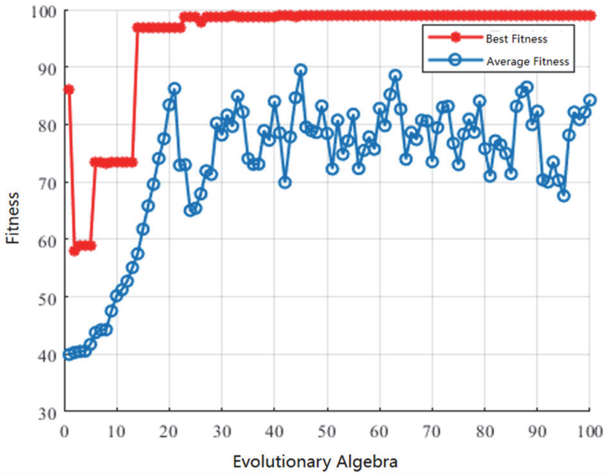

Figure 11 Iterative fitness curve of genetic algorithm

(4) Comparison of accuracy and test time of three optimization methods

Table 7 Comparison of operation time and classification accuracy of the three algorithms

Search Method Comparison Item	Grid Search Method	Genetic Algorithm	Particle Swarm Optimization
Test time	15 s	35 s	29 s
Penalty factor C	5.2780000	0.962920	0.5172000
Kernel function parameters g	0.0358970	0.0038147	0.0100000
Training set classification accuracy	98.2638%	98.7269%	97.7431%

It can be seen from Tab. 7 that the optimization effect of genetic algorithm is the best and the time is relatively short, so the parameters C = 0.962920, g = 0.0038147.

5.5 Training and Testing of Optimized SVM Model

Set the number of categories n = 4, C = 0.962920, g = 0.0038147, and use SVM to read training data and training label. The test set data is brought into support vector machine for training, and the corresponding traffic state prediction label value is obtained. Tab. 8 shows the partial test and forecast tag values.

Only one test label deviates from the prediction label according to the result of classification, and the classification accuracy reaches 98.61%. The experimental results show that the model based on parallel genetic fuzzy clustering and SVM has high accuracy. By inputting the traffic flow parameter matrix into SVM model, the real-time discrimination can be completed, so that the real-time situation of traffic state can be understood. Therefore, the real-time traffic state discrimination method of urban expressway established in this paper is feasible.

Table 8 Prediction label and test label

Data Sheet Number	Test Label	Forecast Label
1	4	4
2	1	1
3	1	1
4	1	1
5	1	1
6	2	2
7	2	2
8	2	2
9	2	2
10	2	2
11	2	3
12	1	1

6 CONCLUSIONS

Based on the measured data, a parallel genetic fuzzy clustering algorithm is established to divide the traffic state in this paper, and then the SVM model is used to distinguish the traffic flow data, which can obtain the real-time traffic situation on a certain section, so as to provide a reference for people's travel. On the one hand, the PGFCM-SVM model uses the genetic algorithm for the optimization of FCM, and obtains a clustering method that has fast convergence speed and can search for the best value with high probability. On the other hand, using grid search method, genetic algorithm and particle swarm algorithm for SVM optimization can get more accurate traffic state discrimination results. However, there are still many deficiencies in the research process. Firstly, in the fuzzy C-means clustering algorithm, the value of the fuzzy coefficient m has a great influence on the clustering effect. The basis of this paper is only based on the previous research experience. When m is set to [1.5, 2.5], the clustering effect is optimal. The value range of m needs to be further explored. Secondly, the weight of the three parameters of traffic flow in clustering is not considered. Lastly, this paper studies the traffic status on a single road section. In the follow-up research, the traffic status analysis of multiple road sections can be carried out to explore the traffic status relationship between each adjacent road section.

Funding Statement

This research was funded by Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX20_1117, SJCX21_1420, KYCX21_2999), Construction System Project of Jiangsu Province (2020ZD14, 2018ZD258), Suzhou Social Science Fund (Y2020LX017, Y2020LX025), General Project of Jiangsu Natural Fund (BK20151201, BK20160357) and Philosophy and Social Science Projects of Universities in Jiangsu Province (2018SJA1348).

Acknowledgements

The authors would like to express their sincere thanks to the anonymous reviewers for their constructive comments on an earlier version of this manuscript. They would also like to acknowledge the support of the participating organization and its personnel who provided various effective assistance, such as the data collection, the case study project and so on. Lastly, they sincerely thank the financial support provided by the above-mentioned mechanism.

Conflicts of Interest

The authors declare no conflicts of interest.

7 REFERENCES

[1] Dudek, C. L., &Messer, G. M. (1974). Incident Detection on Urban Freeways. *TRB, National Research Council*, (495), 12-24.

[2] Cook, A. R. & Cleveland, D. E. (1974). Detection of Freeway Capacity Reducing Incidents by Traffic Stream Identification. *TRB, National Research Council*, (495), 1-11. https://doi.org/10.1139/l74-013

[3] Martin, P., Perrin, H., & Hansen, B. (2000). *Incident detection algorithm evaluation*. Technical Report, Utah Traffic Laboratory.

[4] Yang, K., Zhang, X., & Wang, X. (2019). Traffic Flow State Recognition Based on Multi-source Data Fusion. *Information & Communications*, *7*, 14-15.

[5] Lin, H., Li, L., & Wang, H. (2020). Survey on Research and Application of Support Vector Machines in Intelligent Transportation System. *Journal of Frontiers of Computer Science and Technology*, *14*(6), 901-917.

[6] You, J., Fang, S., Tang, T. et al. (2018). A Support Vector Machine Approach on Real-time Hazardous Traffic State Detection. *Journal of Transportation Systems Engineering and Information Technology*, *18*(4), 83-87, 95.

[7] Hawas, Y. E. (2007). A fuzzy-based system for incident detection in urban street networks. *Transportation Research Part C: Emerging Technologies*, *15*(2), 69-95. https://doi.org/10.1016/j.trc.2007.02.001

[8] Ritchie, S. G. & Cheu, R. L. (1993). Neural network models for automated detection of non-recurring congestion. *California Partners for Advanced Transit and Highways (PATH)*, 10-16.

[9] Stephanedes, Y. J. & Liu, X. (1995). Artificial Neural Networks for Freeway Incident Detection. *Transportation Research Record*, (1494), 91-97.

[10] Prashant Borkar, L. & Malik, G. (2013). Acoustic Signal based Traffic Density State Estimation using SVM. *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, *5*(8). https://doi.org/10.5815/ijigsp.2013.08.05

[11] Huang, Y., Xu, L., & Kuang, X. (2015). Urban Road Traffic State Identification Based on Fuzzy C-mean Clustering. *Journal of Chongqing Jiaotong University*, *34* (2), 102-107.

[12] Hongzhao, D., Shuai, M., & Mingfei, G. (2012). Spatial and temporal model for urban regional traffic state analysis based on fuzzy C-means clustering. *Application Research of Computers*, *29*(4), 1263-1266.

[13] Li, X. & Xu, J. (2011). Discriminating for Traffic Situation of Highway Based on RBF Neural Network. *Computer Simulation*, *028*(02), 350-353.

[14] Wang, J., Wu, J., Ni, J., Chen, J., & Xi, C. (2018). Relationship Between Urban Road Traffic Characteristics and Road Grade Based on a Time Series Clustering Model: A Case Study in Nanjing, China. *Chinese Geographical Science*, *28*(06), 1048-1060. https://doi.org/10.1007/s11769-018-0982-2

[15] Dong, C., Shao, C. J., & Xiong, Z. (2011). Identification of traffic states with optimized SVM method on urban expressway network. *Journal of Beijing Jiaotong University*, *35*(6),13-16,22.

[16] Yu, R., Wang, G., Zhang, J., & Wang, H.-Y. (2013). Urban Road Traffic Condition Pattern Recognition Based on Support Vector Machine. *Journal of Transportation Systems Engineering and Information Technology*, (01), 134-140.

[17] Li, Q., Gao, D., & Yang, B. (2009). Urban-road traffic status classification based on fuzzy support vector machines. *Journal of Jilin University (Engineering and Technology Edition)*, *39*, 131-134.

[18] Goldberg, D. E. (1989). *Genetic Algorithms in Search. Optimization and Machine Learning*. MA: Addison-Wesley.

[19] Zhang, Y. & Zhou, Y. (2019). Review of clustering algorithms. *Computer application*, *39*(07), 1869-1882.

[20] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, *4*(1), 95-104. https://doi.org/10.1080/01969727408546059

[21] Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithm*. New York: Plenum. https://doi.org/10.1007/978-1-4757-0450-1

[22] Li, X. (2016). Urban Road Traffic State Identification Based on Fuzzy C-Mean Clustering. *Technology and Economy in Areas of Communications*, *18*(4), 32- 36.

[23] Nefti, S., Oussalah, M., & Kaymak, U. (2008). A New Fuzzy Set Merging Technique Using Inclusion-Based Fuzzy Clustering. *IEEE Trans on Control System*, *16*(1), 145-161. https://doi.org/10.1109/TFUZZ.2007.902011

[24] Zhang, L., Ma, J., Ran, B., & Yan, L. (2017). Traffic Multiresolution Modeling and Consistency Analysis of Urban Expressway Based on Asynchronous Integration Strategy. *Modelling and Simulation in Engineering*, 19. https://doi.org/10.1155/2017/3694791

[25] Ma, J., Zhang, L., & Yan, L. (2019). Macroscopic Traffic Flow: A Modeling and Simulation Study Considering Impedance Characteristics. *Advances in Transportation Studies. an international Journal*, *2*, 161-174.

[26] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

[27] Lian, M., Zhong, Z., & Chen, Z. (2015). Research and simulation of kernel function selection for support vector machine. *Computer Engineering & Science*, *37*(06), 1135-1141.

[28] Zhou, T. (2015). *Research on Multi-classification Method Based on Support Vector Machine*. University of Electronic Science and Technology of China.

[29] Su, X. (2015). Pattern Recognition Based on Multi-class Support Vector Machine. *Computer & Digital Engineering*, *43*(07), 1202-1206.

[30] Yu, S., Zhang, J., Zhang, X., & An, Y. (2018). Survey on Multi Class Twin Support Vector Machines. *Journal of Software*, *29*(1), 89-108.

Contact information:

Liyan ZHANG, Senior Experimentalist
School of Civil Engineering,
Suzhou University of Science and Technology,
1701 Binhe Road, New District, Suzhou 215011, China
E-mail: 379067846@qq.com

Jian MA, Associate Professor
(Corresponding author)
1) School of Civil Engineering, Suzhou University of Science and Technology,
1701 Binhe Road, New District, Suzhou 215011, China
2) Graduate School of Environmental Studies,
Nagoya University, 465-0015, Japan
E-mail: 9764634@qq.com

Xiaofeng LIU, Associate Professor
School of Automotive and Transportation,
Tianjin University of Technology and Education,
Tianjin, 300222, China

Min ZHANG, Master
School of Civil Engineering,
Suzhou University of Science and Technology,
1701 Binhe Road, New District, Suzhou 215011, China
E-mail: 1964187878@qq.com

Xiaoke DUAN, Master
School of Civil Engineering,
Suzhou University of Science and Technology,
1701 Binhe Road, New District, Suzhou 215011, China
E-mail: 422285946@qq.com

Zheng WANG, Master
School of Civil Engineering,
Suzhou University of Science and Technology,
Binhe Road, New District, Suzhou 215011, China
E-mail: 993102706@qq.com