Hindawi
Scientific Programming
Volume 2021, Article ID 4702669, 15 pages
https://doi.org/10.1155/2021/4702669

Research Article

The Method of Dynamic Identification of the Maximum Speed Limit of Expressway Based on Electronic Toll Collection Data

Fumin Zou,[1] Feng Guo,[1] Junshan Tian,[2] Sijie Luo ,[2] Xiang Yu,[2] Qing Gu,[3] and Lyuchao Liao[4]

[1]*College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, Fujian, China*
[2]*Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China*
[3]*Fujian Provincial Expressway Information Technology Co., Ltd., Fuzhou 350011, Fujian, China*
[4]*Fujian Provincial Big Data Research Institute of Intelligent Transportation, Fujian University of Technology, Fuzhou 350118, Fujian, China*

Correspondence should be addressed to Sijie Luo; sjluo@fjut.edu.cn

Received 5 August 2021; Revised 16 October 2021; Accepted 2 November 2021; Published 18 November 2021

Academic Editor: Zhu Xiao

To overcome the drawbacks of the maximum speed limit information of expressways (i.e., long update cycle and great complexity of information recognition), in this work, an Electronic Toll Collection (ETC) gantry data-based method for dynamically identifying the maximum speed limit information of expressways is proposed. Firstly, the characteristics of the ETC gantry data are analyzed, and then data are cleaned and reconstructed, after which an algorithm is proposed for constructing a vehicle travel speed data set. Secondly, the speed feature vector model of the road section is established by taking the relationship among the speed distribution feature, time domain feature, and the maximum speed limit of the road section into consideration. Then, a data supplement algorithm is constructed to solve the problem of the imbalance of data samples. Finally, the combined GC-XGBoost classification algorithm is used to train and learn the potential speed limit features, and it is verified through the Fujian Provincial Expressway ETC data and the speed limit information provided by the Fujian Traffic Police. The result shows that the accuracy of the method in the recognition of the maximum limited speed information of the expressway is 97.5%. Compared with the traditional limited speed information recognition and extraction methods, the proposed approach can identify the maximum limited speed information of each section of the expressway more efficiently. It can also accurately identify the dynamic change of the maximum limited speed information, which is able to provide data support for intelligent expressway management systems and map providers.

1. Introduction

In recent years, China's expressway ETC system technology has been developed rapidly. More and more vehicles have installed ETC equipment. These vehicles interact with ETC gantries during driving, resulting in massive ETC data. At present, the cumulative users of ETC have exceeded 220 million, and the utilization rate of vehicle owners is 78% [1]. Moreover, the ETC gantry can also interact with the Manual Toll Collection (MTC) system users. Therefore, the ETC gantry system almost collects the traffic information of all vehicles on the expressway, reflecting the overall traffic situation of the expressway, which can provide strong support for the informatization construction, vehicle infrastructure cooperation, and automatic driving [2] of smart expressway. Obtaining the maximum speed limit information of each section of the expressway is an important part of intelligent management of expressways [3]; it can provide drivers with expressway speed limit information [4,5] to avoid traffic accidents caused by speeding and provide reliable perception and driving speed decision-making for autonomous vehicles. However, the maximum speed limit information is dynamic and changeable. The relevant management departments will adjust the speed limit information of the road section according to road traffic flow, road maintenance conditions, and the number of traffic accidents [6–8]. At present, the method of collecting speed limit identification information is mainly manually collected, then the data is uploaded to the system for updating within a certain period. However, this method has two disadvantages: first, it requires professionals to travel to the expressway and collect speed limit information, which costs immense manpower and material resources. Second, it has a long update cycle, and the driver cannot obtain the latest speed limit information, which leads to safety hazards while driving, and the traffic efficiency of the road is correspondingly reduced. Therefore, the study of how to automatically collect the speed limit information and dynamically identify the maximum speed limit information on the road in real-time has research significance.

Traffic flow prediction and travel time prediction are research hotspots in the field of transportation. Most of their research methods and speed limit recognition are supervised learning based on machine learning algorithms. The difference is that speed limit recognition is a classification problem, and traffic flow prediction and travel time prediction are regression problems. The recognition of road maximum speed limit information mainly relies on image recognition technology [9–12] and floating car trajectory data mining technology. The image recognition technology obtains the speed limit information of each road by recognizing the speed limit information of the traffic signs on the road. Machine learning is widely used in a variety of research fields [13]. Support Vector Machine (SVM) [14], Extreme Learning Machine (ELM) [15], and multitask convolutional neural network (MTCNN) [16] are used to train and learn speed limit signs features to realize the recognition of maximum road speed limit. Although these methods are relatively suitable in terms of recognition effect, they require surveyors to collect pictures of speed limit signs on the road, which consumes a lot of resources. In addition, the collection period is long and cannot achieve real-time and dynamic recognition maximum speed limit information. In terms of floating car trajectory data mining, the floating car is equipped with a global positioning system, which records the time, location, and other information of the vehicle, and the floating car trajectory data mining can obtain the driving speed feature of all floating car on the road [17]. Machine learning algorithm [18] is able to learn the maximum speed limit feature in the vehicle speed information of the road to realize the recognition of the maximum speed limit information. However, the floating car accounts for a small proportion of all cars that cannot fully reflect the speed of the vehicles on the expressway. Therefore, the maximum speed limit recognition based on floating car data still has certain defects.

In view of the high cost of speed limit sign recognition and the shortcomings of trajectory data recognition, this study proposes a method using real-time traffic data collected by an ETC gantry system to identify the maximum speed limit of expressways dynamically, which solves the problems of the high cost of manual information collection and incomplete vehicle data. First, the road section speed set construction algorithm and section driving speed abnormal filtering algorithm are designed to ensure the integrity and reliability of the sample data. Then, the speed feature vector model of the speed limit feature is constructed to mine the speed limit feature of the vehicle speed in different aspects. Finally, taking the road maximum speed limit information of 534 sections of expressways in Fujian Province as the sample set. Then, the multivoting ensemble algorithm is used to perform supervised classification training and cross-validation on the road speed feature. The test results show that this method can well identify the maximum speed limit information and recognize the dynamic changes of the maximum speed limit information on the road.

The contributions of this paper can be summarized as follows. First, an algorithm is proposed for constructing speed sets of road section, which can solve the problem that the speed of road section cannot be calculated due to the lack of transaction records of ETC gantries and obtain the speeds of vehicles on each road section accurately and completely. Second, this proposal extracts the feature of the road section speed from different aspects to construct the road section speed feature vector model and mine the potential correlation features between the speed of the vehicles on the expressway and the road speed limit information. Third, a dynamic recognition method of the maximum speed limit of expressways is proposed to identify the maximum speed limit of the expressway, the validity of the method is verified by the real maximum speed limit information, and the scientificity is verified by comparing a large number of prediction algorithms.

This paper is organized as follows. Section 1 introduces the research methods of road speed limit recognition. Section 2 defines the related concepts in this work. Section 3 describes each part of the dynamic method of expressway maximum speed limit. Section 4 shows the experimental results and analysis. Section 5 draws the conclusion and future work.

2. Relevant Definitions

Definition 1. Each ETC gantry of the expressway is collectively called Node, and two adjacent Nodes on the road constituting an expressway section, which is referred to as $QD = \{\mathrm{Q}, \mathrm{Distance}\}$, $Q = < \mathrm{Node1}, \mathrm{Node2} >$, Node and Q, are shown in Figure 1, where Node1 is the start point of the road section, Node2 is the end point of the road section, and Distance is the actual distance of the road section.

Definition 2. Expressway network, formed by all the expressway sections within this proposal, referred to as $G = \{\mathrm{QD1}, \ldots, \mathrm{QDn},\}$.

Definition 3. A set of ETC gantries by which a vehicle passed while driving on the expressway, forming a sequence of nodes in chronological order called trajectory $\mathrm{Traj} = \{D_0, D_1, Di, \ldots, Dj, D_E\}$, $D_i = (N_i, T_i)$, $0 \le i \le E$, $\forall i \le j$, $T_i \le T_j$. D_i is the trajectory point, including node N_i and time property T_i, N_i is the label of the ith node passed by the vehicle, and T_i is the information interaction time when the vehicle passes through node N_i. D_0 is the start point of the trajectory, and D_E is the end point of the trajectory.

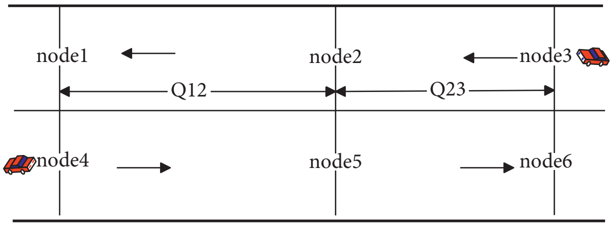

Figure 1: Schematic of the sections.

Definition 4. The average speed of a vehicle passing through a certain road section is called road section speed. The calculation method is shown in the following equation:

$$v = \frac{s}{t_2 - t_1}, \tag{1}$$

where s is the actual length of the road section, t_1 is the time when vehicles pass the start point of the road section, and t_2 is the time when vehicles pass the end point of the road section.

Definition 5. The dispersion of the speed of the road section describes the measures of dispersion of the average speed of vehicles passing through the road section. The section speed of vehicles on the expressway within a certain period of time constitutes the speed set of the section. Sort the value of speed: the speed at 85th percentile is v_1, and the speed at 15th percentile is v_2. The speed dispersion index can be expressed as

$$\Delta v = v_1 - v_2. \tag{2}$$

The larger the value range is, the higher dispersions of the speed information are.

Definition 6. The speed limit includes the minimum speed limit and the maximum speed limit. The speed limit value is generally an integer multiple of 10. In this paper, we only discuss the maximum speed limit.

3. Methods

3.1. ETC Data Preprocessing

3.1.1. ETC Data Cleaning. The ETC gantry system can generate a large amount of transaction data in a short period. Due to system error, information exchange interruption, and severe weather conditions, these factors can lead to abnormal data which can affect the results. In order to reduce interference, the data needs to be preprocessed, mainly including the following aspects.

Data Redundancy: *Duplication between Multiple Data.* The transaction information of each vehicle passing through the ETC gantry should be unique. However, due to problems in data acquisition, transmission, storage process, and other intermediate links, it can cause the repeated data uploading and duplication, resulting in data redundancy. Therefore, these data need to be cleaned.

Data Error. The data record does not conform to the normal driving rules, including two ETC gantries that control different driving directions recorded by the same vehicle at the same time, and different passing records of the same vehicle are recorded at the same time. These data need to be filtered or deleted.

3.1.2. Vehicle Speed Recognition Algorithm in Road Section. In order to calculate the speed distribution of the road section, it is necessary to obtain the transaction data of all vehicles of each gantry. However, gantry transaction data may be missing. Therefore, all traffic data and road network data need to be checked and supplemented to ensure the integrity of the gantry transaction data. After the transaction data of the ETC gantry system is initially cleaned, the trajectory Traj of each vehicle is constructed in chronological order according to the transaction data of each gantry. Traverse each adjacent ETC gantry Node_i , $\text{Node}_{\text{i+1}}$in the Traj one by one. Check whether the road section formed by the two gantries QD_j belong to the expressway road network G. If the road section QD_j belongs to the expressway road network G, the speed v of the vehicle passing through the section QD_j is directly generated. QD_j and the speed v are expressed as follows:

$$\begin{gathered} QD_j = \left\{ \langle \text{Node}_i, \text{Node}_{i+1} \rangle, \text{Distance}_j \right\}, \\ v = \frac{1}{n} \sum_{QD_j, T}^{i \in n} v_i, \end{gathered} \tag{3}$$

where n represents the number of all vehicles within certain time period T of the road section QD_j and v_i represents the average speed of each vehicle on the road section QD_j within certain time period.

If QD_j does not belong to the expressway road network G, it means that the section data of the middle gantries are missing. And path searching algorithm based on Node_i, Node_{i+1} needs to be performed to fill the missing gantry transaction data. As shown in Figure 2, if the road section formed by Node_i and Node_{i+1} cannot be queried in the road network G, use Node_i and Node_{i+1} as the basic node. The feasible path Node_i, Node_a, Node_b, Node_{i+1} can be obtained through path search. Node_a and Node_b are supplementary nodes, and the average speed v between Node_i and Node_{i+1} is taken as speed for$\langle \text{Node}_i, \text{Node}_a \rangle$, $\langle \text{Node}_a, \text{Node}_b \rangle$, $\langle \text{Node}_b, \text{Node}_{i+1} \rangle$.

To ensure the reliability of the average speed v, the minimum speed $v_{\min}$ is set for high-speed driving to 30 km/h and the maximum speed $v_{\max}$ for high-speed driving to 160 km/h [19]. If the average speed value is not in the range $v\varepsilon [v_{\min}, v_{\max}]$, where v is the average speed of all road sections between Node_i and Node_{i+1}, it will be deleted as abnormal data. The specific process of the section speed data construction algorithm is shown in Algorithm 1.

3.1.3. Outlier Information Detection Algorithm for Road Section. To better analyze the road section speed distribution feature of each section, a noise data cleaning model is

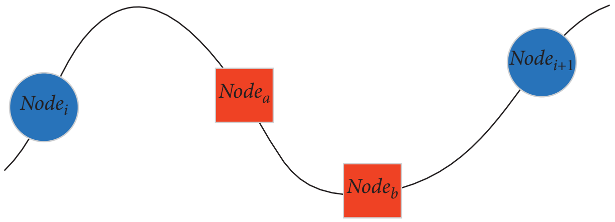

Figure 2: Schematic diagram of driving path.

constructed to detect and eliminate outliers in the data. The basic idea of the model is to use the upper and lower limits of the speed boxplot to detect abnormal points and determine the threshold interval for filtering abnormal speed data. Under the condition of collecting a large amount of expressway ETC transaction data, according to the central limit theorem, the road section speed data set should be a normal distribution. And the upper and lower limits of the speed boxplot that meet the 3σ interval range of the normal distribution can better prove the rationality of realizing outlier detection and filtering through boxplot analysis. As shown in Figure 3, there are 6 element points in the boxplot, among which $q1$ is 1/4 divide point; $q2$ is the median; $q3$ is the 3/4 divide point; and $IQR = q3 - q1$, which is the distance between $q1$ and $q3$. There are also upper limit and lower limit. Here, $q1$ represents the speed value greater than 25% of the traffic flow, $q2$ represents the speed value greater than 50% of the traffic flow, and $q3$ represents the speed value greater than 75% of the traffic flow. Thus, the upper and lower limits of the noise data cleaning threshold model can be obtained, expressed as follows:

$$\begin{aligned}&\text{Upper limit: } q3 + 1.5 * IQR,\\&\text{Lower limit: } q1 - 1.5 * IQR.\end{aligned} \tag{4}$$

Then, the threshold range of velocity filtering is obtained as follows:

$$v_T \in (\text{Lower limit}, \text{Upper limit}). \tag{5}$$

Among which, the speed data of the road section within the range of v_T is retained, and the outlier data is deleted.

3.2. Feature Vector Model of Expressway Speed. Vehicles driving on the expressway have different speeds at different times or on different road sections. Through the statistical analysis of the feature of the traffic speed of the road section, the potential connection between the speed of the vehicle and the road speed limit information can be obtained, after which the road section speed feature vector model is constructed. The feature vector is mainly divided into three categories such that the first is the frequency-speed percentile feature, the second is road section speed evaluation feature, and the third is road section speed time domain feature.

3.2.1. Road Section Frequency-Speed Percentile Feature. Road section frequency-speed percentile feature reflects the distribution of the section speed at different times, including the speed values of the 50th percentile, upper and lower 25th percentile, and the upper and lower 15th percentile of the speed set of the road section, and then converts it into multidimensional feature vector α. It can be expressed as follows:

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_6)^T, \tag{6}$$

where $\alpha_1 \sim \alpha_6$ are, respectively, the 15th, 25th, 50th, 75th, 85th, and 95th percentile of the total section speed distribution, which can describe the overall distribution of the speed in road section.

3.2.2. Road Section Speed Evaluation Feature. Road section speed feature are described by the relevant evaluation indexes in frequency domain, including average speed, speed standard deviation, and speed dispersion, which can transform into multidimensional feature vectors β. It is expressed as follows:

$$\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T, \tag{7}$$

where β_1 is the majority number of section speed, representing the general level of vehicle speed statistical law; β_2 and β_3 are the overall average interval speed of the road section μ and standard deviation σ, respectively; and β_4 attributes the speed dispersion indices, which reflects the changing range and dispersion range of speed data.

3.2.3. Road Section Speed Time Domain Feature. Road section speed time domain feature reflects the speed evolution regularity of the traffic flow on different road sections under different limited speed conditions. If the section speed data was analyzed by day without considering the feature of different periods, it was easily affected by road congestion and other factors in individual periods, and it cannot reflect the speed evolution feature of the road. Therefore, it is necessary to fully integrate the speed feature information of roads in different periods. The whole day is divided into 24 time periods, denoted as 0, 1, ..., 23, respectively. Then, mining and counting the speed information of each road section in each period is carried out to find the speed change law of each road section. As shown in Figure 3, the multidimensional velocity time domain feature vector is constructed. It is expressed as follows:

$$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^T, \tag{8}$$

where $\gamma_1 \sim \gamma_n$ is the average road section speed of each period in the data sample; that is, the average road section speed of 24 time periods in the whole day, in order from large to small, takes the first n values. Here, we take the first 6 values to avoid the disturbance caused by the relatively low road section speed caused by traffic congestion or road maintenance in some periods.

3.3. Sample Imbalance Processing. The road speed limit classification values constructed in this paper conform to the 80 km/h, 100 km/h, 110 km/h, and 120 km/h specified in the

```
Input: trajectory data of a car $D$, expressway road network data G
Output: speed data of the road section
fuction Sections($D$)//The vehicle trajectory data is divided into the data of each section of the vehicle
$D = \{D_1, D_{2,} \ldots, D_E\}$, $D_i = \{N_i, T_i\}$
for $i = 0$ to E-1 do
    Node$_i$, Node$_{i+1}$←$D_{i.N}$, $D_{i+1.N}$//Extracting the node information of two adjacent data points
    Time$_i$, Time$_{i+1}$←$D_{i.T}$, $D_{i+1.T}$//Extracting the time information of two adjacent data points
    delta = Time$_{i+1}$-Time$_i$//Calculating the time difference between two adjacent data points
    $R_{i.Q}$←(Node$_i$, Node$_{i+1}$)//Reconstitute the front and back node information of the vehicle passing section
    $R_{i.T}$←(Time$_i$, Time$_{i+1}$)//The front and back time information of the vehicle passing section
    $R_i$←(road$_{i.Q}$, road$_{i.T}$, delta)
        Sec←{ $R_0$, $R_1$ ,. . ., R$_{E-1}$}//Encapsulating into Sec
    end for
return Sec
end fuction
Sec = { $R_1$, $R_{2,}$ . . ., $R_m$}←Sections (D)
for each $R_j$ in Sec($j$ = 0,1,2, . . ., m) do//Extracting road information from the data, which $R_j$ = ($Q_j$, $T_j$, delta$_j$)
    if $Q_j$ in $G$ then
        Distance$_k$←$G_{k\cdot\text{Distance}}$//Getting road section distance from expressway network, which $k = Q_j$
        $t$ = Sec$_{j.\text{delta}}$//Extracting the time required for vehicles to pass through the road section
        $v_j$ = Distance$_k$/$t$//Speed of vehicle passing through road section
        $R_{j.V}$ ←$v_j$//Adding speed attribute
    if $Q_j$ not in $G$ then//The road information cannot be found in the expressway network, and there is uncollected node information between two nodes of the road section
    {N$_1$, N$_2$, . . ., $N_Z$}←shortest_path(G, $N_j$)//Searching the shortest path between two nodes, getting the path node data set, which Q$_j$=(N$_1$,$N_Z$)
    $A$ = {$A_1$, $A_2$,. . .,A$_{Z-1}$}←{N$_1$,N$_2$,. . ., $N_Z$}//Converting path node data set to road section data set
    path = { }
    for $A_i$ in A then
    path = {path$_1$, path$_2$, . . ., path$_{Z-1}$}←$G_{k\cdot\text{Distance}}$//Getting road section distance from expressway network, and add to path, which $k = A_i$
    end for
    A.$_{\text{Distance}}$←Sum(path)
    $V_A$←A.$_{\text{Distance}}$/$R_{j\cdot\text{delta}}$
    if $V_A \geq V_{\min}$ and $V_A \leq V_{\max}$ then
        for $A_i$ in A then
        $t_i$← ($\sum_1^i$ path$_i$)/$V_A$//Calculating time difference
        $t_1$, $t_2$←$R_{j\cdot T}$//Extracting the time passing through the two nodes separately
        if $i = 1$ then
            $A_{i.tq}$←$t_1$//The time when the vehicle enters the entrance $A_i$
            $A_{i.th}$←$t_1$+$t_i$//The time when the vehicle leaves the entrance $A_i$
            $A_{i.\text{delta}}$←$t_i$//Time difference
        else
            $A_{i.tq}$←$t_1$+t$_{i-1}$//The time when the vehicle enters the entrance $A_i$
            $A_{i.th}$←$t_1$+$t_i$//The time when the vehicle leaves the entrance $A_i$
            $A_{i.\text{delta}}$←t$_i$-t$_{i-1}$//Time of passing through the road section
            $A_{i.T}$←($A_{i.tq}$, $A_{i.th}$)
            $A_{i.V}$←$V_A$
        end for
        A←{Q,T,delta,V}//Getting the corrected section information, including road section node, time and road section speed attributes
    $R_j$←A//A replaces the original $R_j$, and to generate a new $R_j$
        end if
    end if
end for
speed_data←{$R_0$, $R_1$, . . ., $R_c$}//Generating speed data of road section
```

Algorithm 1: Algorithm of speed data construction in road section.

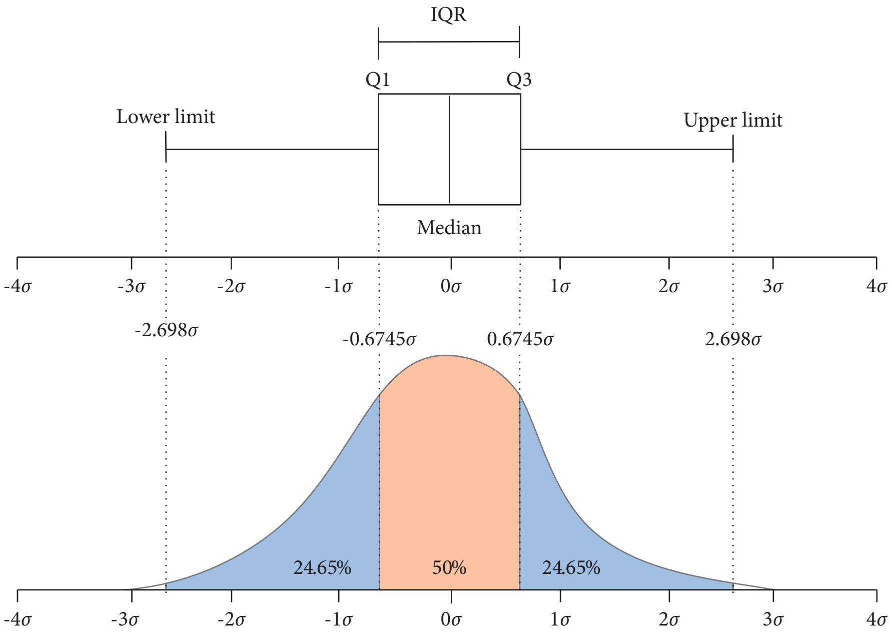

Figure 3: Schematic diagram of noise data cleaning threshold model based on boxplot analysis.

"Road Speed Limit Sign Design Specification" (JTG/T 3381-02-2020) and the "Expressway Engineering Technical Standard" (JTG B01-2003). Because most of the data we collect is 100 km/h, this means the data size of 100 km/h is far more than the other three types of sample data, 80 km/h, 110 km/h, and 120 km/h. This creates an imbalance among sample categories. Therefore, to tackle the problem of unbalanced data samples, there are two processing methods, including oversampling and undersampling [20]. Oversampling is to copy the minority samples multiple times to expand the data volume of the minority samples. This oversampling method will duplicate the preexisting sample data, which will lead to a certain degree of overfitting during the model training process. Undersampling is to randomly remove part of the data from the majority samples or select a part of the sample in this category according to a certain proportion as the sample data. This method will cause the model to only learn a part of the rules of the sample data; thus, it cannot effectively reflect the complete pattern of the sample in this category. In order to alleviate these problems, an improved random oversampling method SOMTE [21] is utilized, which analyzes the minority samples, by using their similarity in feature space to add the simulated new samples to the data set. The number of minority samples in the original data set is expanded, and the dispersion between categories is reduced; therefore, the imbalance problem is solved. The process of the SOMTE can be divided into the following steps:

Step 1. Select the speed feature vector set of minority sample categories with speed limit values of 80, 110, and 120 km/h

Step 2. For each category of sample set, Euclidean distance is used as the metric in the feature space, and then the distance between each sample in the sample set is iteratively calculated to determine the k-nearest neighbor sample points

Step 3. Perform random linear interpolation on the connection line between sample points and the selected s neighboring sample points to generate new samples

Step 4. Repeat Step 2 and Step 3 until the various categories of the expressway speed feature vector data set reach a balance

3.4. Maximum Speed Limit Recognition Classification Model.

The acquisition of speed limit information on expressways is an important factor that affects the driving safety. Different road sections correspond to different speed limit information, and the differences of speed limit information directly affect the state of the vehicles, which makes the relevant data show a certain pattern. Using strong learning machine to perform in-depth learning and training on related data can achieve high-precision recognition results. XGBoost is a method of integrated learning based on a boosting algorithm [22]. Its learning machine usually takes the decision tree model and learns the true value and the residuals of the current prediction values of all trees through the continuous iterative generation of new trees. Then, the results of all trees are accumulated as the final result to obtain a better classification accuracy [23–25]. By using the XGBoost algorithm as a classifier for identifying the maximum speed limit information on expressways, the maximum speed limit information can be determined accurately.

A sample data set is constructed by extracting 16-dimensional speed feature vectors from the expressway section data with the known speed limit information. Suppose the data set is $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. $x_i\,(i = 1, 2, \ldots, M)$ is the feature vector of the ith sample, also known as the input value, that is, the constructed 16-dimensional expressway speed feature vector. $y_i\,(i = 1, 2, \ldots, M)$ is the output value of the ith sample, that is, the road speed limit classification labeled value corresponding to x_i. Assuming that the XGBoost integrated learning model integrates a total of K regression trees, the prediction result of the XGBoost algorithm can be expressed as in the following equation:

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{9}$$

where K is the number of trees, f_k corresponds to the kth regression tree with structure q_k and leaf weight w_k, F is an integrated classifier composed of all regression trees, and$f_k(x_i)$ corresponds to the predicted score of the kth regression tree on the sample x_i.

The objective function of XGBoost consists of a loss function and a regular term, expressed as follows:

$$\mathrm{Obj} = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{10}$$

where l is the error function and $\Omega(f_k)$ is the regularization term. The regular term can be expressed as follows:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda\|w_k\|, \tag{11}$$

where γ represents the penalty coefficient of the model, and the value range is [0,1]. T_k represents the number of leaves of the kth tree; γ is the regular term coefficient.

The XGBoost algorithm adopts an additive step-by-step integration strategy in the training process. First, optimize the first tree, and then optimize the second tree until the kth tree is optimized, and the loss function is continuously reduced during the optimization process. By adding an incremental function f_t in the iterative process to optimize the objective function, the prediction accuracy can be improved, and the calculation method can be expressed as in the following equation:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + c, \tag{12}$$

where c is a constant term and $\widehat{y}_i^{(t-1)}$ represents the predicted value in the $(t-1)$th iteration on the ith sample. Then, carry out the expansion of the second-order Taylor equation and discard the constant term in order to reduce the running time of the model, expressed as follows:

$$\begin{aligned}\mathrm{Obj}^{(t)} &= \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda w_j^{\,2}\right)\right] + \gamma T,\end{aligned} \tag{13}$$

where $I_j = \{i|q(x_i) = j\}$ represents the sample set of leaf j and g_i and h_i are the first derivative and the second derivative of the loss function, respectively.

The objective function is converted into a quadratic function $\mathrm{Obj}^{(t)}$ about w_j to find the minimum value, and then the optimal prediction score of each leaf node and the optimal value of the objective function are obtained as follows:

$$\begin{aligned} w_j^{\,*} &= -\frac{G_j}{H_j + \lambda}, \\ \left(\mathrm{Obj}^{(t)}\right)^* &= -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^{\,2}}{H_j + \lambda} + \gamma T,\end{aligned} \tag{14}$$

where $G_j = \sum_{i\in I_j} g_i$, $H_j = \sum_{i\in I_j} h_i$.

After that, the optimization of XGBoost parameters mainly include the following 4 steps:

Step 1. Choose a higher learning rate, set a reasonable initial value of the booster parameters, and use K-fold cross-validation in each iteration to get the ideal number of decision trees

Step 2. According to Step 1, the learning rate and the number of decision trees are determined, and the $K-$fold cross-validation method and grid search method are used to optimize the parameters of each boosting machine

Step 3. The method is the same as Step 2; based on the given data, adjust the regularization parameters to reduce overfitting

Step 4. Appropriately reduce the learning rate to determine the final ideal parameter combination of the model

3.5. Maximum Speed Limit Recognition Model. The problem of identifying the maximum speed limit information on expressways is a classification problem. The framework of identification model is shown in Figure 4. Dynamic

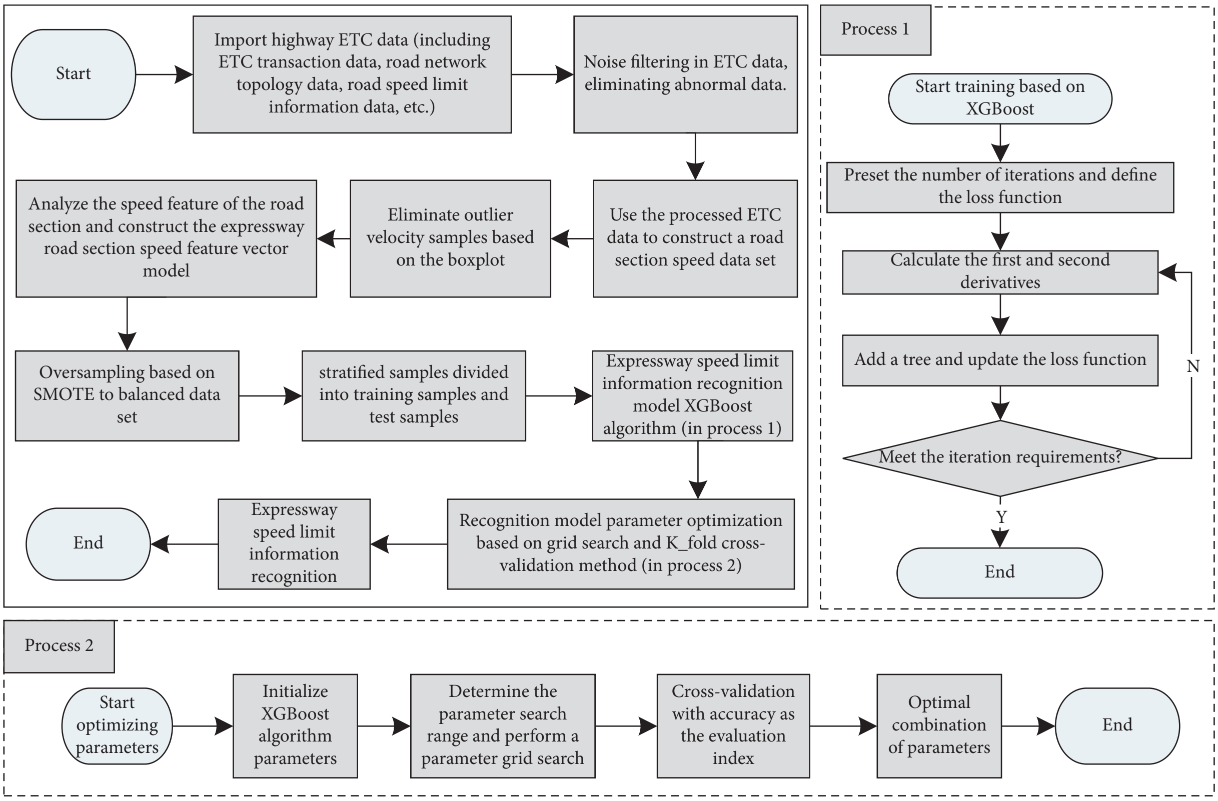
Start
Import highway ETC data (including ETC transaction data, road network topology data, road speed limit information data, etc.)
Noise filtering in ETC data, eliminating abnormal data.
Use the processed ETC data to construct a road section speed data set
Eliminate outlier velocity samples based on the boxplot
Analyze the speed feature of the road section and construct the expressway road section speed feature vector model
Oversampling based on SMOTE to balanced data set
stratified samples divided into training samples and test samples
Expressway speed limit information recognition model XGBoost algorithm (in process 1)
Recognition model parameter optimization based on grid search and K_fold cross-validation method (in process 2)
Expressway speed limit information recognition
End
Process 1
Start training based on XGBoost
Preset the number of iterations and define the loss function
Calculate the first and second derivatives
Add a tree and update the loss function
Meet the iteration requirements?
N
Y
End
Process 2
Start optimizing parameters
Initialize XGBoost algorithm parameters
Determine the parameter search range and perform a parameter grid search
Cross-validation with accuracy as the evaluation index
Optimal combination of parameters
End

Figure 4: The flowchart of expressway speed limit information recognition model.

identification of highway speed limit information is realized based on the following steps. First, the data cleaning is adopted on ETC gantries transaction data, removing duplicated data and error data. Taking vehicle speed recognition, the algorithm is used to find the missing records in the ETC gantries transaction data and to accurately reduce of gantry distribution on expressways. The speed of the road section can be obtained by calculating the speed of the vehicle between the gantries. However, there are some very large or small outliers in the speed of the road section so that boxplot is utilized to remove speed outliers. Next, the speed of each driving section is analyzed, and the models of frequency-speed percentile feature, interval speed evaluation feature, and interval speed time domain feature are constructed. Since the velocity distributions of various types in the data are quite different, the oversampling algorithm is used to expand the minority samples to obtain the balanced data. Finally, data are divided into training data and test data. The training data are inputted into XGBoost algorithm for training and learning; the training process is shown in process 1 in Figure 4. At the same time, the grid search and cross-validation are used to find the optimal parameters of each boosting machine in XGBoost; the optimization process is shown in process 2 in Figure 4.

4. Experiments and Results

4.1. Introduction of Experimental Data. ETC gantry system is one of the main components of the Expressway ETC System, which is used for real-time vehicle driving information supervision and record, vehicle path identification, toll data fitting, and other functions [14]. The experimental data mainly includes three categories. One is the ETC transaction data collected by the ETC gantry on various sections of the expressway in Fujian Province for 9 days from September 3 to September 11, 2020; it contains 50 expressways including Fuyin Expressway, Xiazhang Expressway, and Longchang Expressway, which contains 534 sections, about 100 million pieces of data. The average distance between each section is 8.9 km, 85% of the section distance are less than 16 km, and the maximum distance is 30 km; its distribution is shown in Figure 5. These data are sourced from Fujian Provincial Expressway Information Technology Co., Ltd. The main attributes of the data are shown in Table 1. The second category is the road speed limit information data, including the name of the road section and the maximum speed limit value of the road section, which is derived from the online announcement of the Fujian traffic police. It is used for model learning, training, and testing; the third category is the distance of each section of the expressway from the Amap, including the node pair of the gantry of each section and the actual road section distance.

4.2. Experimental Results and Analysis

4.2.1. ETC Data Preprocessing. Matching the initially cleaned ETC data with the road network topology data, the road section speed of each vehicle is calculated, and then the expressway road section speed data set is constructed. Table 2 shows the

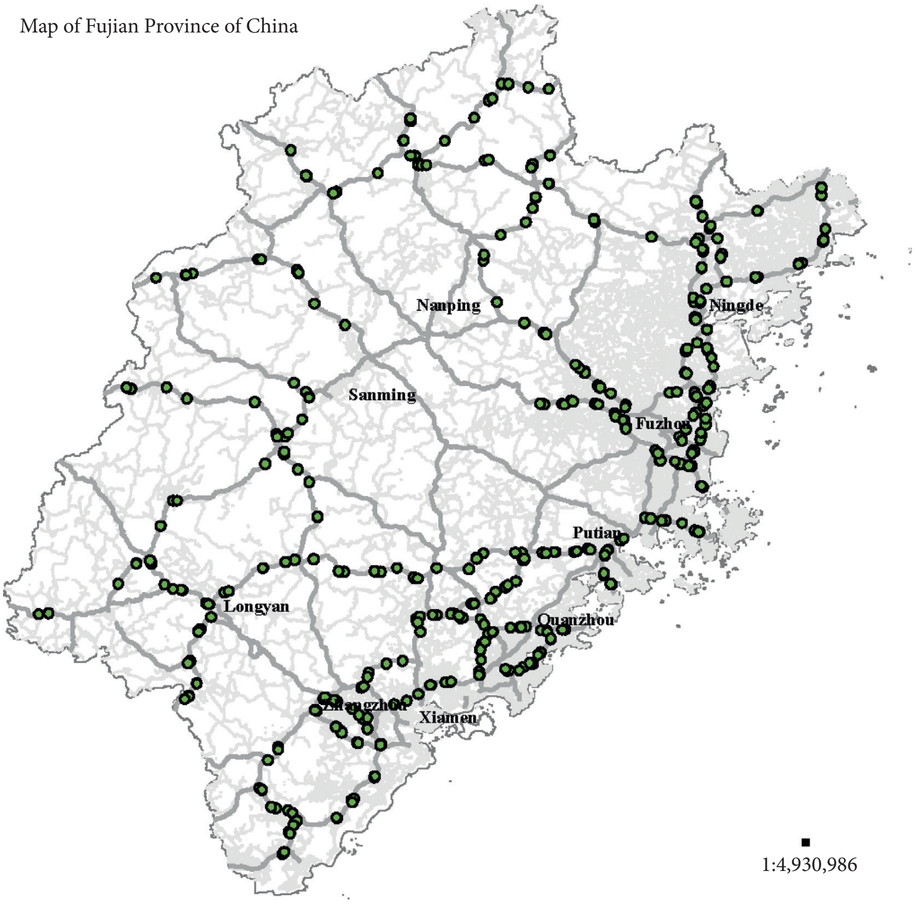

Figure 5: Distribution of expressway gantries in Fujian province.

Table 1: ETC shelf system transaction data attribute table.

Attribute name	Examples	Attribute name	Examples
Trade ID	340***98	OBU plate	Blue Fujian A12345
Trade time	2020/9/5 21 : 29 : 26	Vehicle class	1
Flag ID	35**15	Enter time	2020/9/5 21 : 29 : 26
Flag type	0	Enter station	25*7
Flag index	1	OBU ID	12B***E7

main characteristics of the data. Due to the influences of some random factors, there may be a certain amount of outlier data; these outlier values of each road section can be detected through the noise data filtering model. After the noise data is eliminated, the road section velocity data after preprocessing is obtained. As shown in Figure 6, the road section speed data of the road section from September 3, 2020, to September 11, 2020, is used. Among them, the abscissa denotes the date of each day, and the ordinate represents the magnitude of the road section speed. In addition, each box represents the overall distribution of the road section speed of the road section on that day, and the black origin represents the part need to be deleted. The original speed data of the road section are around 1.229 million, the abnormal data are about 1.19 million, accounting for 9.68%, and the preprocessed section speed data is approximately 11.1 million.

4.2.2. Road Section Velocity Feature Vector. After obtaining the preprocessed speed data set of the road section, the road section speed feature vector model is constructed based on the statistical analysis of the expressway road section speed feature by day. Thus, the expressway road section data set contains 3 types, including 16-dimensional feature vector, and its sample classification mark value is obtained. The attributes shown in Tables 3–5 are the feature vectors, and output of the model after the speed data feature is extracted. Among them, $Q\ D$ is a road section; for example, $QD_{340507-351C03}$ represents the road

Table 2: Expressway road section speed data attribute table.

Attribute name	Examples	Attribute name	Examples
OBUPLATE	Blue Fujian A12345	Time delta	439.0 s
Before trade time	2020-09-07 11 : 07 : 35	Speed	87.15 km/h
Before flag ID	34**05	Enter time	2020-09-06 21 : 24 : 20
After trade time	2020-09-07 11 : 14 : 54	Enter station	330**11
After flag ID	34**07	Road distance	10628 m

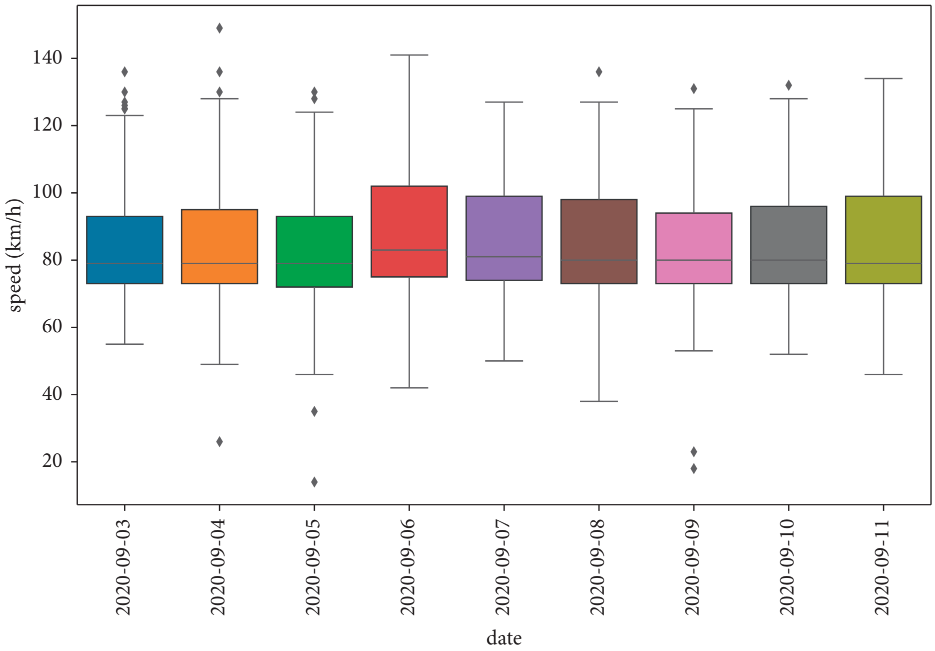

Figure 6: Velocity information distribution boxplot.

Table 3: Frequency-speed percentile feature (unit: km/h).

$Q\ D$	Date	α_1	α_2	α_3	α_4	α_5	α_6	l
$QD_{340507-351C03}$	2020-9-3	70	73	79	92	103	109	110
$QD_{340507-351C03}$	2020-9-4	70	73	79	95	102	110	110
$QD_{34012B-34012\ D}$	2020-9-8	51	59	80	92	95	103	100
$QD_{34012B-34012\ D}$	2020-9-9	51	58	82	94	94	104	100
$QD_{350703-350701}$	2020-9-7	67	71	78	83	87	92	80
$QD_{350703-350701}$	2020-9-8	70	75	82	89	92	96	80
$QD_{341801-341801}$	2020-9-3	88	95	106	114	117	123	120
$QD_{341801-341801}$	2020-9-4	91	97	107	114	117	123	120

Table 4: Road section speed evaluation feature (unit:km/h).

$Q\ D$	Date	β_1	β_2	β_3	β_4	l
$QD_{340507-351C03}$	2020-9-3	76	83	13	32	110
$QD_{340507-351C03}$	2020-9-4	76	83	14	32	110
$QD_{34012B-34012\ D}$	2020-9-8	95	75	19	44	100
$QD_{34012B-34012\ D}$	2020-9-9	95	76	20	46	100
$QD_{350703-350701}$	2020-9-7	82	77	9	20	80
$QD_{350703-350701}$	2020-9-8	82	81	10	22	80
$QD_{341801-341801}$	2020-9-3	114	103	13	29	120
$QD_{341801-341801}$	2020-9-4	111	104	12	26	120

Table 5: Road section speed time domain feature (unit:km/h).

$Q D$	Date	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	l
$QD_{340507-351C03}$	2020-9-3	88	87	87	87 8	6	86	110
$QD_{340507-351C03}$	2020-9-4	91	89	89	89	88	86	110
$QD_{34012B-34012\ D}$	2020-9-8	81	80	80	8	80	80	100
$QD_{34012B-34012\ D}$	2020-9-9	85	83	82	82	81	81	100
$QD_{350703-350701}$	2020-9-7	82	81	79	79	79	78	80
$QD_{350703-350701}$	2020-9-8	84	84	83	83	82	82	80
$QD_{341801-341801}$	2020-9-3	106	106	105	105	105	105	120
$QD_{341801-341801}$	2020-9-4	107	106	105	105	105	105	120

Table 6: Optimal combination of model parameters.

Parameter	Search scope	Step length	Optimal value
n_estimators	[100,1000]	100	700
learn_rate	[0,0.5]	0.01	0.07
max_depth	[1,15]	1	8
min_child_weight	[1,9]	1	1

section between ETC gantry 340507 to ETC gantry 351C03. Date represents the date when the traffic condition occurred, and $\alpha_1 - \alpha_6$ represent that each section is between 15% and 95% of driving speed, where $\beta_1 - \beta_4$ represents the mode, average, standard deviation, and dispersion of vehicle speed, $\gamma_1 - \gamma_6$ represent the first 6 values after sorting the average road speed in 24 time periods of the day, and$l$ represents the maximum speed limit value.

4.2.3. Balance Analysis of Sample Data. There are 5,081 samples in road section speed feature vector data set, among which the number of samples with 80 km/h, 100 km/h, 110 km/h, and 120 km/h speed limits accounts for 5.31%, 87.24%, 9.39%, and 2.83%, respectively, which are seriously unbalanced among different categories and have adverse effects on the efficiency of model identification. Therefore, the SMOTE is used to oversample the sample data with speed limits of 80, 100, and 120 km/h, which makes it possible to achieve relative balance among all kinds of samples. In the experiment, the new data obtained by the SMOTE algorithm is used as the input of the algorithm model. The sample data consists of training sample data and testing sample data.

4.2.4. The Result of the Model's Performance. The parameter setting of XGBoost algorithm is an important factor that affects the performance of the model. In order to improve the accuracy of the model, a set of sensitivity experiments is conducted to optimize the performance of the model. First, four boosting machine parameters are identified that have a significant impact on the model, including n_estimators, learn_rate, max_depth, and min_child_weight. Second, a combination of grid search and K-fold cross-validation (GK) are used to obtain the optimal parameters, in which $K = 5$ for cross-validation. Follow the method of Section 3.4 for parameter optimization. The search range, step length, and postexperiment parameter optimizations for each parameter are shown in Table 6.

The model can be established through the above processing, using test data to verify the effectiveness of the model, and the results of the confusion matrix are shown in Table 7. In 3295 test samples, 3212 were identified correctly, with an accuracy rate of 97.5%. The recognition accuracy of 80 km/h data is 100%. This is because the data with a speed limit of 80 km/h is quite different from other categories and can be better distinguished. However, the gap between the category data with100 km/h and110 km/h is very small, and it is easy to cause mistakes in identification. Among them, there are 824 sample data with a speed limit of 100 km/h, 759 correctly identified, and 47 with a speed limit of 110 km/h, which makes the accuracy rate decrease to some extent. For the same reason, the accuracy rate of the 110 km/h limit is also lower position compared with the other three categories.

4.2.5. Comparison and Analysis

(1) Impact Analysis of Data Equalization. In order to verify the influence of oversampling model on SMOTE algorithm, the original data set and the data set processed by SMOTE algorithm are used for training and learning. The other steps of the model are consistent, and two model classifiers are obtained. The comparison of classification results is shown in Table 8. The first category is the model result corresponding to the data set processed by the SMOTE algorithm, and the second category is the model result corresponding to the original data set. The following can be seen from Table 8:

(1) After the SMOTE algorithm oversampled the data, the accuracy, recall rate, and $F1$-score of all categories were greatly improved.

(2) The data with the speed limit value of 100 km/has the most samples. Without data expansion in the oversampling process, the evaluation indexes of this class are still improved, indicating that the SMOTE algorithm can not only greatly improve the

Table 7: Confusion matrix.

	Real class				
Speed-limiting class (km/h)	80	100	110	120	Accuracy rate (%)
80	824	0	0	0	100
100	6	759	47	12	92.1
110	0	12	807	5	97.9
120	0	1	0	822	99.9
Accuracy rate	99.3%	98.3%	94.5%	98.0%	97.5
	Forecast result				

Table 8: Effect comparison before and after data oversampling.

Category	Speed limit category (km/h)	Precision	Recall	$F1$-score
After oversampling	80	1.00	1.00	1.00
Before oversampling	80	1.00	0.51	0.67
After oversampling	100	0.98	0.92	0.95
Before oversampling	100	0.91	0.98	0.94
After oversampling	110	0.94	0.98	0.96
Before oversampling	110	0.73	0.46	0.57
After oversampling	120	0.98	1.00	0.99
Before oversampling	120	0.80	0.32	0.46
After oversampling	Avg/total	0.98	0.97	0.97
Before oversampling	Avg/total	0.89	0.90	0.89

recognition accuracy of minority speed limit information, but also effectively improve the recognition accuracy of majority speed limit information.

(3) The SMOTE algorithm improves the prediction accuracy of data with a speed limit of 110 km/h and 120 km/h, and the recall rate and $F1$-score are also greatly improved. It has little effect on the prediction accuracy of class data with a speed limit of 80 km/h but has a great influence on the recall rate and $F1$-score.

(2) Comparison and Analysis of Feature Vector Model. By only adjusting input features, the other steps remain the same; the effectiveness of different types of features in expressway section speed feature vector model can be verified. Seven sets of experiments are set up to verify the influence of a single-feature and multiple-feature combinations on the model. Model A_α indicates that only frequency-velocity percentile feature is considered. Model A_β only considers the road section velocity evaluation feature. Model A_γ only considers time domain feature of road section velocity. Model $A_{\alpha,\beta}$ indicates that frequency-velocity percentile feature and road section velocity evaluation feature are considered. Model $A_{\alpha,\gamma}$ takes into account the frequency-velocity percentile feature and road section velocity time domain feature. Model $A_{\beta,\gamma}$ takes into account the road section velocity evaluation feature and road section velocity time domain feature. Model $A_{\alpha,\beta,\gamma}$ takes into account the frequency-velocity percentile feature, road section velocity evaluation feature, and road section velocity time domain feature. All the features are taken into account, and the experimental results are compared. The experimental results are shown in Figure 7, where A1–A7 represent models A_α, A_β, A_γ, $A_{\alpha,\beta}$, $A_{\alpha,\gamma}$, $A_{\beta,\gamma}$, and $A_{\alpha,\beta,\gamma}$, respectively. The following can be seen:

(1) When only a single feature is added, a better model prediction effect can be obtained by adding frequency-velocity percentile feature, followed by interval velocity evaluation feature model and interval velocity time domain feature model.

(2) When two features are added, the prediction effect is improved compared to a single feature. When all the features are added, the prediction effect is the best.

(3) The contribution of each feature in the speed feature vector model of the expressway section to the prediction model is arranged from large to small, which is the road section speed-frequency percentile feature, road section speed time domain feature, and road section speed evaluation feature; the contribution of the feature vector in each feature is shown in Figure 8.

(3) Comparison of Classification Models. To further illustrate the advantages of the model, we compare the performance of GBDT, KNN, SVM, AdaBoost, and Logistic Regression (LR) with our method. The experimental results are shown in Table 9. From the comparison of six different classification methods in Table 7, SVM, AdaBoost, and LR

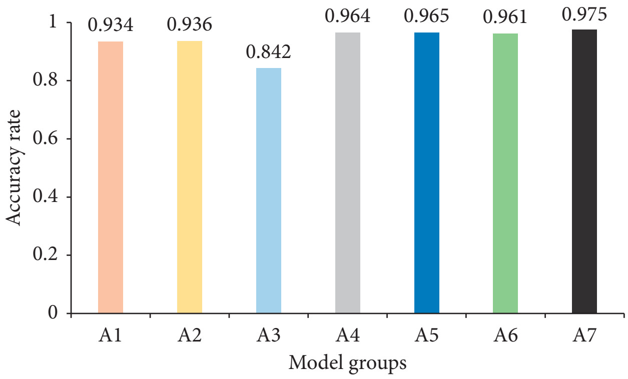

Figure 7: Model accuracy comparison.

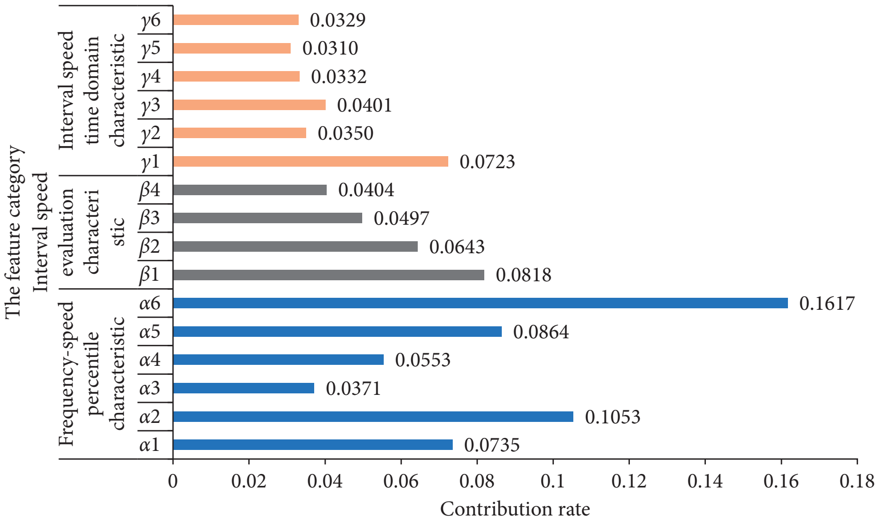

Figure 8: Feature contribution.

Table 9: Model comparison results.

Model	Testing samples	Prediction of correct samples	Accuracy (%)	Precision	Recall rate	$F1$-score
GC-XGBoost	3295	3212	97.5	0.98	0.97	0.97
GBDT	3295	2908	88.3	0.88	0.88	0.88
KNN	3295	3079	93.4	0.94	0.93	0.93
SVM	3295	2374	72.0	0.70	0.72	0.70
AdaBoost	3295	1911	58.0	0.61	0.58	0.51
LR	3295	1684	51.1	0.48	0.51	0.49

classifiers perform poorly in terms of the accuracy, recall rate, and $F1$-score. GC-XGBoost, GBDT, and KNN can get an ideal result on the expressway maximum speed limit information recognition, and the recognition accuracy is high. In particular, GC-XGBoost outperforms GBDT and KNN in terms of the quality of results, with the highest accuracy rate of 97.5%.

5. Conclusion

This paper proposes a method of identifying expressway speed limit information based on ETC data mining analysis. First, the abnormal data of ETC gantry is processed, and a road section speed data set construction algorithm is proposed. The speed data of the road section is constructed, and the outlier samples in each road section are eliminated by the boxplot analysis to ensure the accuracy of the ETC data expression. Then, the SMOTE algorithm is used to oversample the samples of the minority speed limit categories to achieve the balance between the various types of road section speed limit information. Finally, the oversampled training samples are input into the proposed GC-XGBoost (grid search + cross-validation + XGBoost) algorithm for training and learning; then it is compared and analyzed with multiple similar algorithms. The experimental results show the following:

(1) The contribution of each feature in the speed feature vector model of expressway section to the

prediction model is arranged from large to small, followed by the speed-frequency percentage feature, time domain feature, and speed evaluation feature. Three categories of features have an improvement effect on the prediction model, and the frequency-speed percentile feature has the best improvement effect.

(2) In the test sample data, the speed limits of 80 km/h, 100 km/h, 110 km/h, and 120 km/h classification data recognition accuracy are 100%, 92.1%, 97.9%, and 99.9%; the overall accuracy is 97.5%. The gap between the category data with 100 km/h and 110 km/h is very small, so the recognition accuracy is relatively low.

(3) The speed limit recognition accuracy of GC-XGBoost is 97.5%, precision is 0.98, recall is 0.97, and $F1$-score is 0.97. The experimental results are significantly better than those of the other five algorithms, which can accurately identify the maximum speed limit information of expressway.

This paper considers the speed feature of hybrid vehicles, which is suitable for the identification of the maximum speed limit information of expressway. However, this work still has some limitations:

(1) The speed limit recognition of 100 km/h and 110 km/h is less effective. More speed limit features can be considered to explore the differences between the two to improve their speed limit recognition effect.

(2) In this study, we do not consider the speed limit values of different lanes on the same road. In the future, they can be considered to analyze the speed limit information on different lanes of the same road through vehicle classification and road lane number and construct a more complete expressway speed limit information recognition model.

Data Availability

The data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (41971340), the Special Funds for the Central Government to Guide Local Scientific and Technological Development (2020L3014), the 2020 Fujian Province "the Belt and Road" Technology Innovation Platform (2020D002), and the Provincial Candidates for the Hundred, Thousand and Ten Thousand Talent of Fujian (GY-Z19113).

References

[1] "Ministry of Transport of the People's Republic of China. Today, announce the change process of our country's expressways[EB/OL].(2021-03-22)," https://mp.weixin.qq.com/s?__biz=MzI3MDQwMDQ5NQ==&mid=2247537632&idx=1&sn=8c806399c88108c7bae2c3dd00f56e30&scene=0.

[2] Z. Yao, H. Jiang, Y. Cheng, Y. Jiang, and B. Ran, "Integrated schedule and trajectory optimization for connected automated vehicles in a conflict zone," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2020.

[3] M. H. Hosseinlou, S. A. Kheyrabadi, and A. Zolfaghari, "Determining optimal speed limits in traffic networks," *IATSS Research*, vol. 39, no. 1, pp. 36–41, 2015.

[4] L. Aarts and I. Van Schagen, "Driving speed and the risk of road crashes: a review," *Accident Analysis & Prevention*, vol. 38, no. 2, pp. 215–224, 2006.

[5] G. Sugiyanto and S. Malkhamah, "Determining the maximum speed limit in urban road to increase traffic safety," *Jurnal Teknologi*, vol. 80, no. 5, 2018.

[6] B. Khondaker and L. Kattan, "Variable speed limit: an overview," *Transportation Letters*, vol. 7, no. 5, pp. 264–278, 2015.

[7] A. Van Benthem, "What is the optimal speed limit on freeways?" *Journal of Public Economics*, vol. 124, pp. 44–62, 2015.

[8] Y. Zhang and P. A. Ioannou, "Combined variable speed limit and lane change control for highway traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1812–1823, 2016.

[9] J. Cao, C. Song, and S. Peng, "Improved traffic sign detection and recognition algorithm for intelligent vehicles," *Sensors*, vol. 19, no. 18, p. 4021, 2019.

[10] S. K. Berkaya, H. Gunduz, O. Ozsen, and G Serkan, "On circular traffic sign detection and recognition," *Expert Systems with Applications*, vol. 48, pp. 67–75, 2016.

[11] D. Tabernik and D. Skočaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1427–1440, 2019.

[12] M. Liang, X. Cui, and Q. Song, "Traffic sign recognition method based on HOG-Gabor feature fusion and Softmax classifier," *Journal of Traffic and Transportation Engineering*, vol. 17, no. 03, pp. 151–158, 2017.

[13] C. Jiang and X. Xue, "A uniform compact genetic algorithm for matching bibliographic ontologies," *Applied Intelligence*, vol. 51, pp. 7517–7532, 2021.

[14] F. Zaklouta and B. Stanciulescu, "Real-time traffic sign recognition in three stages," *Robotics and Autonomous Systems*, vol. 62, no. 1, pp. 16–24, 2014.

[15] S. Aziz and F. Youssef, "Traffic sign recognition based on multi-feature fusion and ELM classifier," *Procedia Computer Science*, vol. 127, pp. 146–153, 2018.

[16] H. Luo, Y. Yang, B. Tong, W. Fuchao, and F. Bin, "Traffic sign recognition using a multi-task convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1100–1111, 2017.

[17] A. Pascale, F. Deflorio, M. Nicoli, B. Dalla Chiara, and M. Pedroli, "Motorway speed pattern identification from floating vehicle data for freight applications," *Transportation Research Part C: Emerging Technologies*, vol. 51, pp. 104–119, 2015.

[18] L. Liao, X. Jiang, M. Lin, and F. M Zou, “Recognition method of road speed limit information based on data mining of traffic trajectory,” *Journal of Traffic and Transportation Engineering*, vol. 15, no. 5, pp. 118–126, 2015.

[19] J. Yang, J. Xu, C. Gao, B. Guohua, X. Linfang, and L. Menghui, “Modeling of the relationship between speed limit and characteristic speed of expressway traffic flow,” *Sustainability*, vol. 11, no. 17, p. 4621, 2019.

[20] R. C. Prati, G. E. Batista, and M. C. Monard, “A study with class imbalance and random sampling for a decision tree learning system,” in *Proceedings of the IFIP International Conference on Artificial Intelligence in Theory and Practice*, pp. 131–140, Springer, Boston, MA, July-2008.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] T. Chen and C. Guestrin, “Xgboost: a scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, University Press, New York, NY, USA, August 2016.

[23] T. Chen, T. He, M. Benesty et al., “Xgboost: extreme gradient boosting,” *XGBoost contributors [cph] (base XGBoost implementation*, vol. 1, no. 4, 2015.

[24] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, “Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis,” *Accident Analysis & Prevention*, vol. 136, Article ID 105405, 2020.

[25] X. Shi, Y. D. Wong, M. Z. F. Li, C. Palanisamy, and C. Chai, “A feature learning approach based on XGBoost for driving assessment and risk prediction,” *Accident Analysis & Prevention*, vol. 129, pp. 170–179, 2019.