Research Article

Vehicle Emission Detection in Data-Driven Methods

Zheng He ^{1,2}, Gang Ye ^{1,2}, Hui Jiang ³, and Youming Fu ^{1,2}

¹National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China

²Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China

³School of Computer Science, Wuhan University, Wuhan 430072, China

Correspondence should be addressed to Gang Ye; yeg@whu.edu.cn

Received 26 May 2020; Revised 29 July 2020; Accepted 27 September 2020; Published 14 October 2020

Academic Editor: Jun-Jun Jiang

Copyright © 2020 Zheng He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Environmental protection is a fundamental policy in many countries, where the vehicle emission pollution turns to be outstanding as a main component of pollutions in environmental monitoring. Remote sensing technology has been widely used on vehicle emission detection recently and this is mainly due to the fast speed, reality, and large scale of the detection data retrieved from remote sensing methods. In the remote sensing process, the information about the fuel type and registration time of new cars and nonlocal registered vehicles usually cannot be accessed, leading to the failure in assessing vehicle pollution situations directly by analyzing emission pollutants. To handle this problem, this paper adopts data mining methods to analyze the remote sensing data to predict fuel type and registration time. This paper takes full use of decision tree, random forest, AdaBoost, XgBoost, and their fusion models to successfully make precise prediction for these two essential information and further employ them to an essential application: vehicle emission evaluation.

1. Introduction

The popularization of vehicles in our daily life has been continuously enhanced with the expansion of urbanization around the world. Gasoline-engine vehicles are the most popular and widely used type compared with new energy ones, and the pollution gases, such as carbon dioxide, carbon oxide, hydrocarbon, and oxynitride, from vehicles have become the main contaminants in urban atmospheric pollution [1]. Efficient vehicle pollution detection therefore turns to be an emergency task which attracts more and more attention. Exhaust emission detection methods have evolved from periodic detection in the environmental monitoring station to daily road detection with remote sensing technology. This paper studies the vehicle emission detection in cities of China which is one of the largest developing countries.

In the USA, EPA (Environmental Protection Administration) proposed MOVES algorithm [2] to calculate the vehicle emission ratio in some fixed locations and periods of time. The Japanese government enforces the vehicle exhaust

emission monitoring system in their country, and the emission behaviour of each vehicle in Japan can be checked on the official website of Japanese national transportation [3]. In order to rapidly capture the emission detection results, a French transport agency collects the emission pollution-related information from different places and puts them together to realize the sharing network for vehicle emission detection [4]. Related researches and works on this area started a bit later in China. In 2011, Cheng et al. [5] made systematic analysis for the harm caused by vehicle emission, verifying the necessities of exhaust emission controlling. Next year, Wu [6] collected the values of CO₂, HC, CO, and NO exhausted by 1092 vehicles in the Xian Yang city using simplified loaded mode. They established regression equations between the emission value and vehicle information and found that the average emission value was highly related with the vehicle acceptability and the age of the vehicle. Referring to the local standards, they further gave a systematic explanation for the rationalization of the local standard mean emission value based on their research. With the development of remote sensing technology, a large

amount of practical exhaust emission data can be obtained by environmental protection agencies in China. This paper introduces data mining technology to these valuable data to explore efficient information in vehicle exhaust emission detection. This research has a huge potential contribution in promoting the environmental protection department's accurate assessment of unqualified vehicles and providing a theoretical basis for policymakers to learn from.

The first successful vehicle emissions demonstration system was probably an across-road vehicle emissions remote sensing system (VERSS) proposed by Gary Bishop and colleagues in the University of Denver in the late 1980s [7, 8]. A liquid nitrogen cooled nondispersive infrared was the first instrument that can only measure CO and CO₂. In the next two decades, their team continuously refined the system: added hydrocarbon, H₂O, and NO channels to their NDIR system [9, 10], integrated an ultraviolet spectrophotometer and improved it to enhance NO measurement [11, 12], and removed the dependence on the liquid nitrogen cooling [13]. The Denver group designed another commonly used remote sensing device, known as fuel efficiency automobile test, providing some of the inchoate comments on across-road particulate measurement [14]. There are also many other sensing systems typically based on multiple spectrometric approaches proposed for detection of passing vehicle emissions [15–17]. More recently, Hager Environmental and Atmospheric Technologies introduced an infrared laser-based VERSS named Emission Detection and Reporting (EDAR) system, which incorporated several new functions, making it a particularly interesting system for vehicle emission detection.

Important information is buried in the vehicle emission remote sensing data. This paper exploits data mining methods to deal with the data and obtain valuable knowledge from them. There are three main directions in data mining: the improvements of classical data mining algorithms, ensemble learning algorithms, and data mining with deep learning. The improvements on classical algorithms are usually performed and employed in multiple application scenarios taking additional information into consideration. Ensemble learning is actually the integration of multiple learners with a certain structure which completes learning tasks by constructing and combining different learners. Its general structure can be concluded as follows: firstly, generate a set of individual learners and then combine them with some strategies. The combining strategies mainly include average method, voting method, and learning method. Bagging and boosting [18] are the most commonly used ensemble learning algorithms which improve the accuracy and robustness of prediction models. As the rapid development and popularization of deep learning, it plays more and more important roles in data learning with the support of big data and high-performance computing. Many traffic engineering-related researches mainly focus on analyzing relevant data such as traffic diversion [19], traffic safety monitoring [20], engine diagnosis [21], road safety [22] and traffic accident [23], and remote sensing image processing [24–35], extracting useful information and digging out valuable knowledge. A few works are proposed in vehicle

emission evaluation in data mining ways which is the key study subject in this paper. Xu et al. [36] used XgBoost to develop prediction models for CO₂eq and PM_{2.5} emissions at a trip level. In [37], Ferreira et al. applied online analytical processing (OLAP) and knowledge discovery (KD) techniques to deal with the high volume of this dataset and to determine the major factors that influence the average fuel consumption and then classify the drivers involved according to their driving efficiency. Chen et al. [38] proposed a driving-events-based ecodriving behaviour evaluation model and the model was proved to be highly accurate (96.72%).

Relevant environmental policies have been introduced to define difficult limitation standards based on the vehicle fuel type and registration time in China. The vehicle license plate number, plate color, speed, acceleration, and VSP (vehicle specific power), etc., will be captured by the surveillance system when vehicles pass by the remote survey stations. The analysis for the smoke plume generated by gas emission is simultaneously conducted by laser gears at the stations, where the exhaust emission value can be calculated. With the fuel type and registration time information learned from vehicle plate numbers, it is able to obtain the gas emission standard value to judge whether the vehicle emission is eligible. However, register information of nonlocal vehicles and partial local vehicles is not recorded in the official database due to the limitation of environmental policies, which leads to the failure to provide the fuel type and registration time information for vehicle emission detection. According to the National Telemetry Standard in China, relevant departments will treat the information-missing vehicles as the diesel consumption ones, and this situation keeps the limitation criteria of the emission value of partial vehicles unknown, resulting in the evaluation for these vehicles being unable to carry on. Therefore, the precise information upon fuel types and registration time of vehicles is an essential prerequisite for finding out the pollution-exceeding vehicles. This paper adopts multiple data mining methods to learn the fuel type and registration information of vehicles from remote sensing data and further utilize cascaded classified framework to make accurate prediction on vehicle emission-related information, providing valuable reference standards on evaluation of different vehicles.

2. Data Mining Models for Analysis

In this section, detailed descriptions on the models and dataset used in this study are given [39].

2.1. Data Mining Methods

2.1.1. Decision Tree Model. Decision tree model [40] is a commonly used data mining method based on information theory and a greedy algorithm-like framework, which is proposed for classification or prediction. The model divides the whole dataset into branch-like parts to construct an inverted tree with a root node, internal nodes, and leaf nodes. The nonparametric design enhances the efficiency

and generalization ability of the decision tree in processing a large and complex dataset.

Five core components made up the tree decision, including the following (1) nodes: root node, internal nodes, and leaf nodes which are the three types that represent different choice operations for data distribution. (2) Branches: they represent the splitting process of nodes in the decision tree, and each branch from the root node to a leaf node represents a corresponding decision rule on classification. (3) Splitting: it is a procedure to generate child nodes from parent nodes terminating when the predetermined homogeneity or stopping criteria is met. (4) Stopping: stopping rules are applied to prevent the overfitting and inaccuracy happening. (5) Pruning: it is an alternative way to establish a large tree first and then prune it to an optimal structure by removing useless nodes. This paper uses decision trees to make classification for vehicle fuel type and specifically calculates the information gain of multiple corresponding attributes to generate the rule model for fuel type prediction. The attributes that greatly affect the final results can be shown in a quite intuitive way. Figure 1 illustrates the decision tree model for fuel type prediction.

2.1.2. Random Forest Model. The random forest model is another classic and efficient data mining method that belongs to bagging learning. In 2001, Leo Breiman combined ensemble learning theory [41] with random subspace method [42], proposing the well-known machine learning methodology: random forest. It is a data-driven nonparametric model without a priori knowledge and has good tolerance to noise and abnormal values, as well as excellent extendibility and parallelism abilities for high-dimensional data classification. The ensemble learning structure enables random forest in some extent overcoming the performance bottleneck and overfitting of single classifiers such as SVM.

Given dataset $D = X_i, Y_i$, $X_i \in R^K, Y_i \in \{1, 2, \dots, C\}$, random forest is essentially a series of combined classifiers made up by M decision trees $\{g(D, \theta_m), m = 1, 2, \dots, M\}$. The classification result is decided by voting of every decision tree and is highly related with two vital randomizations: sample bagging and feature random subspace. The sample bagging process randomly picks M training dataset with return which shares the same size with the oral dataset, constructing a corresponding decision tree. When a node in the decision tree is split, the model will randomly select a feature space from the whole K features (usually use $\log_2 K$) from which an optimal splitting feature is selected to construct trees. These features consist of the feature random subspace and contain more discriminative feature combination for classification. Since in the construction of each decision tree, the process of randomly picking training data and feature subspace is independent and the constructions are procedurally identical; $\{\theta_m, m = 1, 2, \dots, M\}$ is a sequence of random variables with independently distribution. This character makes it applicable and efficient to be realized in a parallel computing way and simultaneously ensures the high extendibility of random forest. The structure and construction of random forest are shown in Figure 2.

FIGURE 1: Decision tree model for vehicle fuel type prediction.

2.1.3. AdaBoost Model. Similar to the bagging method, the boosting also belongs to the ensemble learning method which enhances the classification/prediction accuracy of base learner continuously with an iterative update process. AdaBoost [43] based on boosting learning was proposed by Professor Freund and Schapire in 1995. The algorithm was widely used in various classification/prediction fields due to its outstanding performance. The central idea of AdaBoost is to continuously update the wrong judgment weights. Weights of mistaken classified samples of the previous basic classifier are set to increase while the correctly classified samples' weight is programmed to decrease, and the correct one will be used to train the next basic classifier again. At the same time, a new weak classifier is added to cascaded classifiers in each iteration, and the final strong classifier stays undetermined until a predetermined sufficiently small error rate or the prespecified maximum number of iterations is reached. The concrete procedures are presented in the following paragraph. Given the training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$, $y_i \in \{-1, 1\}$, $i = 1, \dots, n$, x_i is the i -th training data with its label y_i , and $y_i = 1$ and $y_i = -1$ denote the positive and negative labels, respectively. w_i is the i -th classifier of AdaBoost. Initialize the weights at first with the uniform distribution:

$$D_1(i) = (w_1, \dots, w_n) = \left(\frac{1}{n}, \dots, \frac{1}{n}\right). \quad (1)$$

Then, perform $m = 1, 2, \dots, M$ iteration: the base classifier $G_m(x)$ is trained in the cost of D_m , and the cost accumulation of the wrong classified samples labeled by $G_m(x_i)$ is represented by the classification error rate E_m which is defined as

$$E_m = P(G_m(x_i) \neq y_i) = \sum_{i=1}^n w_i I(G_m(x_i) \neq y_i). \quad (2)$$

The coefficient of the base classifier ($G_m(x)$) α_m is calculated as the following formula:

FIGURE 2: The construction of random forest.

$$\alpha_m = \frac{1}{2} \ln \left[\frac{(1 - E_m)}{E_m} \right]. \quad (3)$$

In the next iteration, the cost distribution of samples D_{m+1} is updated as

$$D_{m+1} = \frac{D_m(i) \exp(-\alpha_m y_i G_m(x_i))}{Z_m}, \quad (4)$$

where Z_m is defined as

$$Z_m = \sum_{i=1}^n D_m(i) \exp(-\alpha_m y_i G_m(x_i)). \quad (5)$$

The final strong classifier $G(x)$ is the combination of the trained base classifiers, denoted as

$$G(x) = \text{sign}(f(x)) = \text{sign} \left(\sum_{m=1}^M \alpha_m G_m(x) \right). \quad (6)$$

2.1.4. XgBoost Model. XgBoost [44] is a modified algorithm based on GBDT (gradient boosting decision tree) [45]. Both of

them share the same idea with boosting methodology; in each iteration, the current decision trees are learned by the previous iteration results and move forward to the residual diminishing direction. When dealing with multiclassification problems, the logarithmic likelihood loss function is defined as

$$L(y, f(x)) = - \sum_{k=1}^K y_k \log p_k(x), \quad (7)$$

where y is the label of the input data x , k denotes the attribute value, and y_k is an indicator function that if the predicted value is k , then $y_k = 1$. The prediction probability $p_k(x)$ is denoted as

$$P_k(x) = \frac{\exp(f_k(x))}{\sum_{l=1}^K \exp(f_l(x))}. \quad (8)$$

At the t iteration, the label of sample data i is l ; the current negative gradient can be calculated as

$$R_{til} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f_k(x) = f_{l,t-1}(x)}. \quad (9)$$

As a classic algorithm, GBDT has advantages such as high accuracy, robustness, and conveniences. Yet, it just uses the first-order partial derivative to calculate the negative gradient which may cause relatively big error. Aiming at overcoming this shortage, XgBoost deduces the first-order derivative by second-order Taylor expansion and the results turn to be more approximate to ground truth when the deduction is introduced to calculate the leaf node weight. In addition, XgBoost firstly ranks the sample data which stored the records in the form of block and the speed of XgBoost is much faster than GBDT with the same training data.

2.2. Model Fusion. Although the single models mentioned in the above subsection have satisfactory performances on data

classification and prediction, the model fusion methods based on the difference between the characters of different single models are able to further improve the accuracy and robustness for the final results. The voting manner is a common method for model fusion. This paper adopts two fusion methods to construct the combined model: hard voting and soft voting [46].

The hard voting classifier follows the simple idea that the minority obeys the majority. Given the classification/prediction results, treat each label result for the same variance as a vote, the most voted value is set to the variance in hard voting way. This process can be defined as

$$H_{\text{vote}}(x) = \max \left\{ \sum_i \text{lab}(x, i, 1), \sum_i \text{lab}(x, i, 2), \dots, \sum_i \text{lab}(x, i, k) \right\}, \quad (10)$$

where x denotes the variance, $H_{\text{vote}}(x)$ is the vote result of hard voting, $\text{lab}_c(x, j, c)$ is an indicator function that shows whether the j -th classifier deeming the label of x is c , and $\text{lab}_c(x, j, c) = 1$ when the probability that x belongs to label c calculated by j -th classifier: $p(x, j, c)$ exceeds some threshold values; otherwise, $\text{lab}_c(x, j, c) = 0$.

The soft voting classifier is another fusion strategy, which treats the average of the probabilities of all classification/prediction samples for a certain label as the standard. The corresponding label with the highest probability is the final result and the voting can be demonstrated as

$$S_{\text{vote}}(x) = \max \left\{ \frac{\sum_i p(x, i, 1)}{n_{cf}}, \frac{\sum_i p(x, i, 2)}{n_{cf}}, \dots, \frac{\sum_i p(x, i, k)}{n_{cf}} \right\}, \quad (11)$$

where n_{cf} is the total number of classifiers and k is the number of labels.

This paper combines different models mentioned in Section 2 in both methods denoted as hard voting model and soft voting model, respectively, to compare the final results.

3. Data Description

The data source is collected by means of remote sensing. The vehicle remote sensing system is a complicated synthetic system made up of some subsystems, including tail assay, environmental information monitoring, traffic condition monitoring, and vehicle identification. When vehicles pass by the remote sensing devices, the equipment performs detection of smoke plume produced by the diffusion of vehicle exhaust. The specific process can be summarized as follows: the probe light from the remote sensing device passes through the air mass and then returns to the detection unit through the right-angle displacement unit, finishing the detection for carbon dioxide, carbon monoxide, hydrocarbons, nitrogen oxides, etc. The intensity of opaque smoke

exhausted by diesel engine vehicles can also be monitored by the gas-diesel integration design. As the parameters of vehicles' running status, the instantaneous velocity and acceleration of vehicles are obtained synchronously by remote sensing system. The exhaustion and running data are used for the final remote sensing results.

Vehicle exhaust includes water vapor, oxygen, hydrogen, nitrogen, carbon dioxide, carbon monoxide, hydrocarbons, nitrogen oxides, sulfur dioxide, and particulate matter. There are two main methods for remote sensing detection on road for vehicle exhaust analysis, namely, nondispersive infrared red analyzer (NDIR) to measure carbon monoxide, carbon dioxide, and hydrocarbons and the dispersion of the ultraviolet (DUV) method to measure nitric oxide, smoke factors, and particulate matter (including opacity). The exhaust diffuses and dilutes in the air immediately after being discharged, and the variation of the dilution concentration is affected by factors such as air disturbance, wind direction, and wind speed. Direct measurement of the concentration of each pollutant in the exhaust plume may not reflect the vehicle emissions accurately and efficiently. Therefore, carbon dioxide is adopted as the reference gas to measure various exhaust pollutants in vehicle remote sensing technology. The same exhaust remote sensing optical path (including horizontal or vertical erection) is incapable of remotely measuring the exhaust of multiple vehicles at the same time. It must pass one by one and the time slot between passing vehicles should be greater than one second so that enough time could be set aside for the remote sensing device to measure the exhaust of the preceding vehicle. This also allows the exhaust of the front vehicle to spread out in time without affecting the remote sensing of vehicles behind.

The remote sensing data used in experiments come from the real detected data in the database of the Environmental Protection Bureau of a certain city, which is derived from the remote sensing database and consists of three parts of

information: the first part is the vehicle basic information $\vec{V}$, it contains the license plate number m , vehicle license plate color c , and the passing time t ; $\vec{V}$ is denoted as $\vec{V} = (m, c, t)$. The second part is the vehicle condition information $\vec{S}$ which includes the vehicle bodywork length cl , the speed v , and the acceleration a , denoted as $\vec{S} = (cl, v, a)$. The third part is the remote sensing result $\vec{C}$. It includes the detection value of carbon dioxide C_{co_2} , carbon monoxide C_{co} , hydrocarbon C_{hc} , nitric oxide C_{no} , smoke intensity C_{op} , and the environmental detected values: wind speed C_{ws} , wind direction C_{wd} , temperature C_{ot} , humidity C_h , and atmospheric pressure C_{ps} , the remote sensing data thus is recorded as $\vec{C} = (C_{co_2}, C_{co}, C_{hc}, C_{no}, C_{op}, C_{ws}, C_{wd}, C_{ot}, C_h, C_{ps})$. Each record used in the research is composed of the above three parts of information and the range of vehicle fuel type and the registration time are the predicted targets.

4. Experiments

4.1. Data Preprocessing

4.1.1. Fuel Type Prediction. The data of fuel type prediction come from two tables: vehicle information table and remote sensing record table. The ID and fuel type fields are extracted from the former table, and the latter contains (a) vehicle running conditions (speed, acceleration, passing time, etc.). (b) Environmental meteorological conditions (lane, wind direction, wind speed, temperature, humidity, and atmospheric pressure). (c) Remote sensing results (detection value of carbon dioxide, carbon monoxide, hydrocarbons, nitrogen oxides, and the intensity of smoke). This paper relates these two tables by vehicle ID of which the specific process is described as follows:

(1). *Data analysis.* The preliminary statistical analysis on the number of vehicles with different fuel types is made and the result (shown in Figure 3) demonstrates that vehicles with multiple fuel types, such as mixed oil and natural gas type, other than gasoline and diesel account for a very low proportion. The ratio between gasoline and diesel cars in the data is about 4.6 : 1, and the unbalanced ratio means this is an unbalanced dataset.

(2). *Feature layering.* In the case of unbalanced distribution dataset, all the data at the end of the horizontal axis in the distribution graph are concentrated to one level so that the data becomes more concentrated, and the number of levels of data can be reduced. A temporary feature will be generated for hierarchical sampling, dividing the dataset into different sections. This paper divides the vehicle passing time into three time periods: morning, afternoon, and evening.

(3). *Data cleaning.* Data processing for character data of vehicles and missing values are the main components of data cleaning procedure. Vehicle character features include license plate number, license plate color, and the test results, and they are usually converted to one-hot codes in machine

FIGURE 3: The unbalance data distribution of different fuel type vehicles.

learning. There are three common methods to deal with missing values: ignoring the record of missing value, ignoring the missing features, and median/average padding. If the license plate number, license plate color, or detection result is missing, the record will be directly ignored, and continuous value parameters are filled with the mean value in other situations.

(4). *Feature selection.* The main work is to find correlations and feature combinations and generate a correlation matrix for the original data. Check the features in the original data that are most relevant to the fuel type, and use related techniques to find the features that are positively correlated, such as license plate color and nitric oxide. Features that show obvious negative correlation include validity, transit time, and carbon dioxide. Then, check the features that are most relevant to the date of registration of diesel vehicles in the original data, and through technical means, find that the features that show more obvious positive correlation include license plate color, etc., and the attributes that show more negative correlation include validity, detection line, and nitric oxide. In terms of feature combination, because the concentration of plume is affected by the wind speed and the vehicle's own speed during the remote sensing process, the CO_2 gas concentration is used as a reference when recording the exhaust pollutant concentration. The new feature combined is the ratio of other pollution items to CO_2 . The amount of data in this study is trillions; after filtering, the number of features is about 30. The decision tree algorithm is used to calculate feature importance of fuel type prediction. The calculation results are shown in Figure 4. Using random forest to calculate the feature importance of gasoline vehicle registration time period is shown in Figure 5, and calculating the feature importance of diesel vehicle registration time period is shown in Figure 6.

(5). *Feature scaling.* This experiment discovered that the difference in the value range between different features is very large which brings huge obstacles to the decision part

FIGURE 4: Predicting the importance of fuel type attributes.

FIGURE 5: Importance of features of gasoline vehicle registration time period.

FIGURE 6: Importance of features of diesel vehicle registration time period.

during the learning process of the model, and larger values usually cause greater changes in the model. Commonly used scaling methods such as normalization and standardization are introduced in experiments. The normalization formula is $(x - \text{mean})/\text{num}(v)$, x denotes the value of variance, mean is the average value of all the variances, and $\text{num}(v)$ is the total number of variance. The normalization formula can be denoted as $x - \min/\max - \min$ where min and max are the maximum value and minimum value among variances, respectively. With the normalization process, the entire feature values are mapped into the period of [0, 1].

(6). *Dataset division.* This paper randomly divides the source dataset into two parts: 85% data as training dataset and 15% data as validation dataset.

(7). *Model training.* The processed data is put into the three models for training and verification. The comparison among different results from these models is performed in the experiment part.

4.1.2. *Registration Time Prediction.* According to the national and local standards of motor vehicles, the following subdivisions are made for the vehicle registration time. Relevant departments stipulate that for gasoline-powered vehicles, it is divided into two categories: before 2001-10-1 and after 2001-10-1, recorded as

$$GY = \begin{cases} 00, & \text{if before 2001-10-1,} \\ 01, & \text{if after 2001-10-1.} \end{cases} \quad (12)$$

The registration time of diesel-powered cars is divided into three periods, i.e., before 2008-7-1, between 2008-7-1 and 2013-7-1, and after 2013-7-1, denoted as

$$FY = \begin{cases} 10, & \text{if before 2008-7-1,} \\ 11, & \text{if between 2008-7-1 and 2013-7-1,} \\ 01, & \text{if after 2013-7-1.} \end{cases} \quad (13)$$

When the period division is done, the prediction of registration time can be treated as a multiclassification problem.

This remote sensing record table contains three parts information: (a) remote sensing data of vehicle running conditions, including speed, acceleration, and passing time; (b) environmental meteorological conditions, including lane condition, wind direction, wind speed, temperature, humidity, and atmospheric pressure; (c) exhaust remote sensing results that contain detection value of carbon dioxide, carbon monoxide, hydrocarbons, nitrogen oxides, and the intensity of smoke. In the classification of the national landmarks and local landmarks mentioned above, there exists a large error that it is impossible for the neighboring vehicles around certain segment points to change a lot. Therefore, this paper chooses to discard the record data of three months before and after the segmentation point.

4.2. Parameter Optimization

4.2.1. *Decision Tree Tuning.* There are mainly three parameters to adjust in decision tree algorithm:

D_max_depth represents the max depth of a decision tree. According to the principle of decision tree, it can be seen that with deeper layers a decision tree has more power to thoroughly divide attributes and mine the deep relationship between data. This paper experimentally sets the max_depth range from 1 to 32. The F1 curve is shown in Figure 7.

FIGURE 7: Effect of the maximum depth of a decision tree in target prediction.

$D_{min_samples_split}$ denotes the minimum number of samples required to split the internal nodes. At least one sample is required for each node to perform splitting. When the number of internal node samples increases, more samples will participate in the split of the tree. The decision tree will suffer more constraints with more reference data utilized in node splitting, which could also affect the speed of model execution. This paper sets the minimum number of samples for internal nodes ranging from 10 to 500. The F1 curve is shown in Figure 8.

$D_{min_samples_leaf}$ is the number of samples required for leaf node splitting, which is called the minimum number of samples. The leaf node will be pruned if the number of leaf node samples is less than the minimum one.

In experiments, the $D_{min_samples_leaf}$ is set to range from 1 to 100 and the F1 curve is plotted in Figure 9.

According to Figures 7–9, the optimal parameter selection of the decision tree algorithm obtained by the experiment in this paper is as follows: D_{max_depth} is 12, $D_{min_samples_split}$ is 150, and $D_{min_samples_leaf}$ is 10.

4.2.2. Random Forest Tuning. There are four main parameters to adjust in the random forest model:

$n_estimators$ is the number of decision trees in random forest that plays a significant role in the performance of the model. Small value of $n_estimators$ means fewer base classifiers participate in the decision process, leading to the decrease in prediction accuracy, while large number of decision trees will bring out computational burden to the system and take more running time. This paper sets $n_estimators$ from 10 to 100 and plots the F1 curve, as shown in Figure 10.

$max_features$ represents the maximum number of features that can be used when the splitting of decision trees happens. Each node selects all features in the

FIGURE 8: Effect of the minimum number of samples required to split the internal nodes for a decision tree in target prediction.

FIGURE 9: Effect of the number of samples required for leaf node splitting for a decision tree in target prediction.

FIGURE 10: Effect of the number of decision trees in random forest in target prediction.

splitting process when $\text{max_features} = \text{None/Auto}$ and selects no more than $\log N$ and $\sqrt{N}$ when $\text{max_features} = \log$ and $\text{max_features} = \text{sqrt}$, respectively. N is the total number of features. There are few sample attributes in the experiment, and this paper sets the value of max_feature to Auto

$R_{\text{max_depth}}$ is the maximum depth of decision trees in random forest. This paper sets $R_{\text{max_depth}}$ ranging from 1 to 100 and F1 curve is shown in Figure 11.

$R_{\text{min_samples_leaf}}$ is the minimum number of samples in leaf node for splitting. This paper sets it ranging from 1 to 100 and F1 curve is shown in Figure 12.

According to the result of Figures 10–12, this paper defines that $n_{\text{estimators}}$ is 100, max_features is 5, $R_{\text{max_depth}}$ is 24, and $R_{\text{min_samples_leaf}}$ is 2.

4.2.3. AdaBoost Parameters Setting. The default base classifier of AdaBoost is decision tree, which is a classic ensemble learning algorithm with a boosting structure. The base classifier parameters refer to the optimal parameters of decision tree model above, where $\text{max_depth} = 12$, $\text{min_samples_split} = 150$, and $\text{min_samples_leaf} = 10$. Two significant parameters of AdaBoost are the number of tuning for base classifiers N_{tune} and learning rate L_{rate} . The model is easy to overfit if the N_{tune} is too large, and underfitting in reverse. This paper sets $N_{\text{tune}} = 50$ and L_{rate} to be 1 by default.

4.3. Experimental Results and Analysis

4.3.1. Fuel Type Prediction Model. This section uses decision tree, random forest, and AdaBoost algorithm to make fuel type prediction. Vehicle fuel type prediction is a typical two-class classification. Five classification models are used in the experiment including decision tree, random forest, AdaBoost algorithm, hard voting fusion model, and soft voting fusion model. After parameters optimization of all kinds of models, the single classifier models are compared to fusion ones to evaluate these models. The whole technological process is illustrated in Figure 13. More details are given in the following subsections.

In Section 4.2, this paper optimizes the parameters of a single model. Although the performance of the single model is already very good, it is based on the differences between the single models, and the model fusion of the single models becomes very meaningful.

It can be seen from Table 1 that random forest performs best in a single model scene. The fusion model obtained by the voting method also performed very well. Compared with most single models, fusion models have higher prediction accuracy. In the process of predicting the vehicle's fuel type, this paper has obtained a good prediction effect. The random forest algorithm and the fusion model have the best prediction results. The F1 value of the random forest is 90.41%, and the F1 value of the soft fusion model is 90.3%. Because the random forest prediction speed is faster and the

FIGURE 11: Effect of maximum depth of decision trees in random forest in target prediction.

FIGURE 12: Effect of the minimum number of samples in leaf node for splitting for target prediction using random forest.

prediction model is better to explain, the random forest prediction model is selected as the final model for fuel type prediction.

4.3.2. Registration Time Prediction Model: Mixed Fuel Type. Vehicle registration time predictions are divided into diesel vehicles registration time prediction and gasoline vehicles registration time prediction. From the statistical analysis of the data, it can be known that diesel vehicles are mainly divided into vehicles registered between 2009 and 2013 and after 2013. The proportion of registrations before 2009 was very low. Gasoline vehicles are mainly vehicles after 2008, accounting for about 90%. The purpose of predicting the registration date and fuel type is to find the limit value of unknown vehicles to judge whether the emission is qualified. The earlier the registration, the higher the limit value standard. Predicting a car before 2008 to a car after 2008 is equivalent to selecting a low limit value, and it is easy to

FIGURE 13: Architecture diagram of the classification model.

TABLE 1: The performance of multiple models used in the experiment for fuel type prediction.

Model	Accuracy rate (%)	Precision rate (%)	Recall (%)	F1 (%)
Decision tree	97.34	96.83	84.01	89.96
Random forest	97.46	97.32	84.41	90.41
AdaBoost	96.78	93.41	83.21	88.02
Hard voting	97.39	97.74	83.58	90.11
Soft voting	97.43	97.42	84.16	90.30

judge a vehicle with a qualified emission as unqualified, which causes a waste of resources for car owners and environmental inspection workstations. Based on the above analysis, random forest and XgBoost are used in this section to create a prediction model of registration time.

The classification periods are as follows: gasoline + after 2001-10-1, gasoline + before 2001-10-1, diesel + during 2008-7-1 and 2013-7-1, diesel + during before 2008-7-1, and diesel + after 2013-7-1. Since it is a multiclassification

problem, this paper uses random forest and XgBoost to perform prediction and the results are shown in Table 2. In order to reduce the randomness of the learning algorithm, the results show the mean and variance of the results of 10 independent runs.

This paper finds that when the data is divided into five categories, the verification results are unbalanced. The number of gasoline-powered vehicles after 2001 is higher and the verification accuracy is much lower than the training

TABLE 2: Temperature and wildlife count in the five areas covered by the study.

Model	The ratio of five classes	Training accuracy	Test accuracy
Random forest	1776631 : 3326 : 223333 : 157129 : 6376	0.990 $\pm$ 0.005	0.917 $\pm$ 0.006
XgBoost	1776631 : 3326 : 223333 : 157129 : 6376	0.983 $\pm$ 0.004	0.903 $\pm$ 0.005

TABLE 3: Temperature and wildlife count in the three areas covered by the study.

Model	The ratio of three classes	Training accuracy	Test accuracy
Random forest	223333 : 157129 : 6376	0.985 $\pm$ 0.009	0.679 $\pm$ 0.008
XgBoost	223333 : 157129 : 6376	0.987 $\pm$ 0.010	0.901 $\pm$ 0.006

accuracy. The random forest model is superior to the XgBoost model in terms of training accuracy and verification accuracy. Its training accuracy reaches 99.0%, and its verification accuracy is about 91.7%, which indicates that overfitting occasionally happens, leading to the decrease in verification accuracy.

4.3.3. Registration Time Prediction Model: Gasoline Vehicle. The gasoline cars are classified as follows: gasoline + after 2001-10-1 and gasoline + before 2001-10-1. The results are shown in Table 3.

The prediction accuracy and verification are more than 99% using the random forest model. This is mainly caused by the severely unbalanced data distribution and the insufficient data amount.

5. Conclusions

Environmental protection has been a hot topic in academic and industrial communities. This paper focuses on predicting the missing basic information of vehicles from telemetry data to monitor the vehicle emission. A variety of data mining methods are adopted to perform predictions based on the vehicle telemetry data provided by an environmental protection agency in a certain city and successfully made precise inferences on fuel type and gasoline-powered vehicle registration time. In the prediction for the registration time of diesel vehicles, the prediction accuracy rate just reaches about 70% due to the fact that the division of registration time is artificially controlled and the status of different vehicles varies a lot for different users. Further work will be carried out on the basis of more related data and improved algorithms to make more precise prediction on the vehicle emission-related information.

Data Availability

The SQL data used to support the findings of this study have not been made available because the data provider is the Municipal Bureau of Ecology and Environment of a certain city and the data is not public.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61772386 and 61862015 and in part by the State Grid Hubei Electric Power Co., Ltd., under Grant SGHBDK00DWJS1800134.

References

- [1] S. A. Yashnik, S. P. Denisov, N. M. Danchenko, and Z. R. Ismagilov, "Synergetic effect of Pd addition on catalytic behavior of monolithic platinum-manganese-alumina catalysts for diesel vehicle emission control," *Applied Catalysis B: Environmental*, vol. 185, no. 15, pp. 322–336, 2016.
- [2] E. M. Fujita, D. E. Campbell, B. Zielinska et al., "Comparison of the MOVES2010a, MOBILE6.2, and EMFAC2007 mobile source emission models with on-road traffic tunnel and remote sensing measurements," *Journal of the Air & Waste Management Association*, vol. 62, no. 10, pp. 1134–1149, 2012.
- [3] J. S. Fu, X. Dong, Y. Gao, D. C. Wong, and Y. F. Lam, "Sensitivity and linearity analysis of ozone in East Asia: the effects of domestic emission and intercontinental transport," *Journal of the Air & Waste Management Association*, vol. 62, no. 9, pp. 1102–1114, 2012.
- [4] A. Kfoury, F. Ledoux, C. Roche, G. Delmaire, G. Roussel, and D. Courcot, "PM2.5 source apportionment in a French urban coastal site under steelworks emission influences using constrained non-negative matrix factorization receptor model," *Journal of Environmental Sciences*, vol. 40, pp. 114–128, 2016.
- [5] Y. Cheng, *Comparative Study on Sino-US Control Systems of Motor Vehicle Exhaust Pollution*, China University of Geosciences, Beijing, China, 2011.
- [6] X. Wu, *Detection and Standard Revision of Exhaust Pollutants by Acceleration Simulation Mode for In-Use Gasoline Vehicle*, Chang'an University, Xi'an, China, 2012.
- [7] G. A. Bishop, J. R. Starkey, A. Ihlenfeldt, W. J. Williams, and D. H. Stedman, "IR long-path photometry: a remote sensing tool for automobile emissions," *Analytical Chemistry*, vol. 61, no. 10, pp. 671A–677A, 1989.
- [8] R. D. Stephens and S. H. Cadle, "Remote sensing measurements of carbon monoxide emissions from on-road vehicles," *Journal of the Air & Waste Management Association*, vol. 41, no. 1, pp. 39–46, 1991.
- [9] D. H. Stedman, G. Bishop, and S. McLaren, 1995, Alexandria, VA, USA, US Patent and Trademark Office, US Patent No. 5,401,967.

- [10] P. L. Guenther, D. H. Stedman, G. A. Bishop, S. P. Beaton, J. H. Bean, and R. W. Quine, "A hydrocarbon detector for the remote sensing of vehicle exhaust emissions," *Review of Scientific Instruments*, vol. 66, no. 4, pp. 3024–3029, 1995.
- [11] Y. Zhang, D. H. Stedman, G. A. Bishop, P. L. Guenther, and S. P. Beaton, "Worldwide on-road vehicle exhaust emissions study by remote sensing," *Environmental Science & Technology*, vol. 29, no. 9, pp. 2286–2294, 1995.
- [12] P. J. Popp, G. A. Bishop, and D. H. Stedman, "Development of a high-speed ultraviolet spectrometer for remote sensing of mobile source nitric oxide emissions," *Journal of the Air & Waste Management Association*, vol. 49, no. 12, pp. 1463–1468, 1999.
- [13] D. A. Burgard, G. A. Bishop, R. S. Stadtmüller, T. R. Dalton, and D. H. Stedman, "Spectroscopy applied to on-road mobile source emissions," *Applied Spectroscopy*, vol. 60, no. 5, pp. 135A–148A, 2006.
- [14] D. H. Stedman and G. A. Bishop, "Opacity enhancement of the on-road remote sensor for HC, CO and NO," Final Report prepared for CRC-E56-2, University of Denver, Denver, CO, USA, 2002.
- [15] H. Moosmüller, C. Mazzoleni, P. W. Barber, H. D. Kuhns, R. E. Keislar, and J. G. Watson, "On-road measurement of automotive particle emissions by ultraviolet lidar and transmissometer: Instrument," *Environmental Science & Technology*, vol. 37, no. 21, pp. 4971–4978, 2003.
- [16] C. Brekke and A. H. S. Solberg, "Oil spill detection by satellite remote sensing," *Remote Sensing of Environment*, vol. 95, no. 1, pp. 1–13, 2005.
- [17] J. L. Jiménez, M. D. Koplow, D. D. Nelson, M. S. Zahniser, and S. E. Schmidt, "Characterization of on-road vehicle NO emissions by a TILDAS remote sensor," *Journal of the Air & Waste Management Association*, vol. 49, no. 4, pp. 463–470, 1999.
- [18] S. Marsland, *Machine Learning: An Algorithmic Perspective*, CRC Press, Boca Raton, FL, USA, 2005.
- [19] R. Barrett, A. Facey, W. Nxumalo et al., "Dynamic traffic diversion in SDN: testbed vs mininet," in *Proceedings of the 2017 International Conference on Computing, Networking and Communications (ICNC)*, pp. 167–171, IEEE, Santa Clara, CA, USA, January 2017.
- [20] K. Goniewicz, M. Goniewicz, W. Pawłowski, and P. Fiedor, "Road accident rates: strategies and programmes for improving road traffic safety," *European Journal of Trauma and Emergency Surgery*, vol. 42, no. 4, pp. 433–438, 2016.
- [21] M. Yuan, Y. Wu, and L. Lin, "Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network," in *Proceedings of the 2016 IEEE International Conference on Aircraft Utility Systems (AUS)*, pp. 135–140, IEEE, Beijing, China, October 2016.
- [22] P. Cordellieri, F. Baralla, F. Ferlazzo et al., "Gender effects in young road users on road safety attitudes, behaviors and risk perception," *Frontiers in Psychology*, vol. 7, p. 1412, 2016.
- [23] D. A. Wiegmann and S. A. Shappell, *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*, Routledge, Abingdon, UK, 2017.
- [24] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *International Journal of Computer Vision*, vol. 127, no. 5, pp. 512–531, 2019.
- [25] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.
- [26] J. Ma, X. Wang, and J. Jiang, "Image superresolution via dense discriminative network," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 7, pp. 5687–5695, 2020.
- [27] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng, "Nonrigid point set registration with robust transformation learning under manifold regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3584–3597, 2019.
- [28] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sensing of Environment*, vol. 235, p. 111425, 2019.
- [29] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware Convolution neural network for ship detection in surveillance video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 781–794, 2020.
- [30] L. Zhou, Z. Wang, Y. Luo, and Z. Xiong, "Separability and Compactness network for image recognition and super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3275–3286, 2019.
- [31] Z. Wang, P. Yi, K. Jiang et al., "Multi-memory Convolutional neural network for video super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2530–2544, 2019.
- [32] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, p. 1, 2020.
- [33] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, and R. Hu, "SRLSP: a face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 27–40, 2017.
- [34] J. Jiang, X. Ma, C. Chen, T. Lu, Z. Wang, and J. Ma, "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 15–26, 2017.
- [35] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.
- [36] J. Xu, M. Saleh, and M. Hatzopoulou, "A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions," *Atmospheric Environment*, vol. 224, p. 117311, 2020.
- [37] J. C. Ferreira, J. de Almeida, and A. R. da Silva, "The impact of driving styles on fuel consumption: a data-warehouse-and-data-mining-based discovery process," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2653–2662, 2015.
- [38] C. Chen, X. Zhao, Y. Yao et al., "Driver's eco-driving behavior evaluation modeling based on driving events," *Journal of Advanced Transportation*, vol. 2018, Article ID 9530470, 12 pages, 2018.
- [39] J. V. Leme, W. Casaca, M. Colnago, and M. A. Dias, "Towards assessing the electricity demand in Brazil: data-driven analysis and ensemble learning models," *Energies*, vol. 13, no. 6, p. 1407, 2020.
- [40] Y. Y. Song and L. U. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, p. 130, 2015.
- [41] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [42] S. W. Kwok and C. Carter, "Multiple decision trees," *Uncertainty in Artificial Intelligence*, vol. 9, pp. 327–335, 1990.

- [43] R. Kumar and R. Verma, "Classification algorithms for data mining: a survey," *International Journal of Innovations in Engineering and Technology (IJJET)*, vol. 1, no. 2, pp. 7–14, 2012.
- [44] 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, San Francisco, CA, USA, August 2016.
- [45] J. H. Friedman, "machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [46] C. Macdonald and I. Ounis, "Voting for candidates: adapting data fusion techniques for an expert search task," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 387–396, Arlington, VA, USA, November 2006.