ISSN 2616-5872 Vol.3, Issue 4: 50-55, DOI: 10.25236/AJEE.2021.030410

Prediction of CO and NOx Emissions from Automotive Engines Based on Machine Learning Algorithms

Yingping Su[1, *], **Xinyu Li**[2]

[1]School of Business, Southwest University 402460, Chongqing, China
[2]School of The first clinical medicine, Shanxi Medical University, 030001, Taiyuan, China
**Corresponding author*

Abstract: *In this paper, five models including BP neural network, Gaussian process regression, support vector machine regression, extreme learning machine and least squares support vector machine are used to model and predict the emissions of carbon monoxide and nitrogen oxides from automobile engines, and the performance of each model is compared. The input data set of this study includes nine parameters (ambient temperature, ambient pressure, ambient humidity, air filter pressure difference, gas turbine exhaust pressure, turbine inlet temperature, turbine outlet temperature, turbine energy yield, and compressor exhaust pressure), and the output data set includes two parameters (carbon monoxide and nitrogen oxides). The comparison of the research results shows that among the five machine learning algorithms, the Gaussian process regression model has the best fitting effect. The model has the highest prediction accuracy and the smallest error. Therefore, the Gaussian process regression model is used to model and predict the emissions of carbon monoxide and nitrogen oxides from automobile engines, and the optimal parameter values of the minimum emissions are found. Applying these automobile engine parameters to daily life is of great significance to alleviating air pollution.*

Keywords: *Modeling Prediction, Environmental Pollution, Regression Model, Automobile engine exhaust emissions, Machine Learning*

1. Introduction

Environmental pollution and global warming have attracted more and more attention. Carbon monoxide and nitrogen oxides contained in the gases emitted by automobile engines will not only aggravate climate warming, but also cause acid rain to destroy vegetation, which seriously affects people's production and life [1]. In recent years, the state has vigorously promoted the green sustainable development strategy and environmental protection policy. Researchers are paying more and more attention to the hazards of automobile engine emission pollution, and are committed to finding solutions to reduce environmental pollution and improve people's quality of life.

Among the exhaust gas emitted by automobile engines, carbon monoxide is mainly caused by the lack of sufficient oxygen during the combustion process; nitrogen oxides are caused by the chemical reaction between oxygen and nitrogen in the air at high temperatures. Environmental temperature, environmental pressure, environmental humidity, air filter pressure difference, gas turbine exhaust pressure, turbine inlet temperature, turbine outlet temperature, turbine energy yield, and compressor exhaust pressure all affect the generation of harmful gases in the tail gas. Therefore, it is of great significance to study the influence of these factors on the emission of CO and NOx.

In recent years, many scholars have conducted extensive research on automobile exhaust emissions. Shan Guor et al. [2] predicted greenhouse gas emissions based on COPERT model. Chauhan Boski P. et al. [3] used VISSIM model to simulate the driving behavior parameters of a single vehicle category to predict vehicle emissions based on vehicle speed. Zhang Lanyi et al. [4] established a micro emission model for predicting automobile exhaust emission based on the least square regression fitting method.

Many studies have shown that the COPERT model is not suitable for prediction using a large number of data. The least square method is a linear estimation, and the default is a linear relationship. This method has limitations when used, which is mainly reflected in the fact that the regression relationship cannot pass every data point. Therefore, this study uses a variety of machine learning algorithms to model and predict the harmful gas emissions of automobile engines.

The framework of this paper is as follows. The first chapter introduces the research significance and status quo of automobile engine exhaust emissions. In the second chapter, the basic principles of five machine learning algorithms, namely BP neural network, support vector machine regression, Gaussian process regression, extreme learning machine and least squares support vector machine, are introduced.The third chapter introduces the construction process of the model and the evaluation index of the model.The fourth chapter introdeuces the results of comparative analysis of five model performance.In the fifth chapter, the research of this paper is summarized and prospected.

2. Principles of Machine Learning Methods

2.1. BP neural network (BPNN)

BP neural network only learns some rules through its own training, so that it can get the result that is closest to the expected output value under the given input value [5]. BP neural network is a kind of multi-layer feed-forward network trained by error back propagation. Its basic idea is gradient descent method, and gradient search technology is used to minimize the error mean square deviation of the actual output value and the expected output value of the network [6].

The core of BP neural network is the inverse feedback formula. The weights in the neural network can be updated by formula 1, which can make the output results more accurate.

$$\delta_j^l = (\sum_{i=1...n} \delta_i^{l+1} \bullet w_{ij}^{l+1}) \bullet f\,'(z_i^l) \tag{1}$$

2.2. Support vector machine regression model (SVR)

Support vector machine (SVM) maps feature vectors of instances to some points in space.The SVR model can be optimized by minimizing the width of the interval band and the total loss.The SVR kernel maps x of the input space to a higher dimensional feature space by mapping functionφ(x), so the linear model formula in the feature space is shown in formulas 2:

$$f(x) = \sum_{i=1}^{m} (a_i^* - a_i) k(x, x_i) + b \tag{2}$$

2.3. Gaussian Process Regression (GPR)

Gaussian process regression is a nonparametric model, which uses the prior knowledge of Gaussian process to regression analyze the data [8] .GPR can be derived from Bayesian linear regression of normal hypothesis. Bayesian linear regression is a multivariate linear regression model shown in formulas 3:

$$f(X) = X^T w, y = f(X) + \xi \tag{3}$$

In the formula, w is the weight coefficient and ξ is the residual or noise.

Bayesian linear regression assumes that residuals follow an independent and identically distributed 0-mean normal distribution: $p(\xi) = N(\xi \mid \delta_n^2)$. The likelihood of Bayesian linear regression is shown in formula 4:

$$p(y \mid X, w, \delta_n^2) = N(y \mid X^T w + 0, \delta_n^2) = 1/\sqrt{2\pi}\delta_n \exp(-(| y - X^T w|^2 / 2\delta_n^2) \tag{4}$$

2.4. Extreme Learning Machine (ELM)

When solving the weights and thresholds, the extreme learning machine is simpler than the traditional gradient descent method of feedforward neural network based on BP algorithm.For the input layer weights and hidden layer thresholds, the extreme learning machine uses a random given value, and does not need repeated iterations to adjust the weights and thresholds. Determining the number of hidden layer nodes is the key to the fitting effect of the extreme learning machine, and the weight of the output is solved by the least square method [9]. When the values of W and b are fixed, the output weight β can be obtained, and then the estimated value of β is obtained by least squares.

2.5. Least Squares Support Vector Machine model (LSSVM)

The least squares support vector machine model transforms the inequality constraint conditions in the original standard support vector machine into the equality constraint conditions, and uses the error square sum loss function as the empirical loss of the training process to reflect the least squares [10]. The quadratic programming problem is transformed into the solution of linear equations [11].

The Gaussian function is often selected as the kernel function in LSSVM model. The kernel function formula is shown in Formulas 5.

$$K(x_i, x_j) = \exp((-\| x_i - x_j \|^2) / 2g^2) \tag{5}$$

3. Model Construction

3.1. Model Dataset

The data set of this study is obtained from the UCI data set. Each data set includes ambient temperature AT, ambient pressure AP, ambient humidity AH, air filter pressure difference AFDP, gas turbine exhaust pressure GTEP, turbine inlet temperature TIT, turbine rear temperature TAT, turbine energy yield TEY, compressor exhaust pressure CDP, carbon monoxide CO and nitrogen oxide NOx. Because AT, AP, AH, AFDP and other factors have a significant impact on vehicle CO and NOx emissions, these parameters are selected as model inputs and CO and NOx emissions as outputs.

The minimum, maximum, average and unit values of the factors are listed in Table 1:

Table 1: Dataset Information

Variable (Abbr)	Unit	MinimumValue	MaximumValue	MeanValue
AT	C	6.23	37.1	17.71
AP	mbar	985.85	1036.56	1013.07
AH	(%)	24.08	100.2	77.87
AFDP	mbar	2.09	7.61	3.93
GTEP	millibar	17.70	40.72	25.56
TIT	C	1000.85	1100.89	1081.43
TAT	C	511.04	550.61	546.16
CDP	millibar	9.85	15.16	12.06
TEY	MWH	100.02	179.50	133.51
CO	mg/m3	0.00	44.10	2.37
NOx	mg/m3	25.90	119.91	65.29

In this study, MATLAB software version 2020b is used for coding on the computer of windows 10. In this paper, five machine learning modeling algorithms, BPNN, SVM, GPR, ELM and LSSVM, are used to fit the emissions of carbon monoxide and nitrogen oxides from automobile engines through 29575 data sets. In the obtained data set, the data set is divided into two parts by random division, one for training model, and the other for testing model.

3.2. Evaluation Indicators

Combined with the machine learning model used in this study, the commonly used evaluation indexes for judging the fitting effect of the model are collected, including seven statistical parameters [12]: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), mean square error (MSE), standard mean square error (NMSE), determination coefficient (R2) and correction determination coefficient (Ra). The lower the values of MSE, MAPE, RMSE, MAE and NMSE, the greater the values of R2 and Ra, indicating that the fitting effect of the model is good [13]. The specific calculation formulas of the above seven regression model evaluation indicators are shown in Table 2:

Table 2: Evaluation Index Calculation Formula

EvaluatingIndicator	ComputingFormula	EvaluatingIndicator	ComputingFormula
MAE	$MAE = (1/\mathrm{N})\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	MSE	$MSE = (1/N)\sum_{i=1}^{N}((\hat{y}_t - y_t)^2 / \hat{y}_t \bullet y_t)$
RMSE	$RMSE = \sqrt{(1/N)\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	NMSE	$NMSE{=}101\log_{10}(\sum_{n=1}^{N}\lvert z(n) - \hat{z}(n)\rvert^2 / \sum_{n=1}^{N}\lvert z(n)\rvert^2)$
R2	$\mathrm{R}2 = \sum_{i=1}^{n}(f_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$	Ra	$Ra = 1 - (1 - R2)(n-1)/(n-k)$
MAPE	$\mathrm{MAPE} = 1/N\sum_{i=1}^{N}\lvert(y_i - \hat{y}_i)/y_i\rvert \times 100\%$		

4. Results and Discussion

4.1. Model Evaluation

The models used in this study are evaluated by MAE, MSE, RMSE, NMSE, MAPE, R2 and Ra, and the results are listed in Table 3. It is well known that the determination coefficient R2 of GPR model is closest to 1, indicating that the model has the best fitting effect, that is, the model has the strongest ability to explain the dependent variable, and the correction coefficient of GPR model is the largest, indicating that the model is the optimal regression model. Therefore, it can be seen from the model evaluation indexes in table 3 that GPR model is the most suitable model for predicting carbon monoxide and nitrogen oxide emissions of automobile engine. Among the evaluation indexes of LSSVM model, R2 value and Ra value are larger, and MAE, MSE, RMSE, NMSE and MAPE values are smaller than the other three models. Therefore, LSSVM model is inferior to GPR model, and SVR model and ELM model have the worst fitting effect.

Table 3: Comparison of Evaluation Indexes of Regression Model

Evaluating Indicator	BP	ELM	SVR	GPR	LSSVM
MAE	1.5383	2.1345	4.6943	1.1804	1.3299
MSE	7.7808	13.3161	48.7356	5.8031	6.7364
RMSE	2.7894	3.6491	6.9811	2.409	2.5955
NMSE	8.78E-07	1.50E-06	5.08E-06	6.53E-07	7.59E-07
MAPE	0.1364	0.1545	0.4621	0.1138	0.1214
R2	0.8838	0.7439	0.4617	0.9134	0.8994
Ra	0.8837	0.7437	0.4612	0.9133	0.8993

In order to more intuitively see the advantages and disadvantages of the models, the evaluation index data are visualized. Figure 11 shows the comparison of the evaluation results of each model on MAE, MSE, RMSE, NMSE, MAPE, R2 and Ra. For the convenience of image observation, the original MSE data are reduced by 10 times, and the original NMSE data are expanded by 10^7, thus weakening the large difference between data caused by different dimensions. The smaller the MAE, MSE, RMSE, NMSE and MAPE, the better the R2 and Ra.

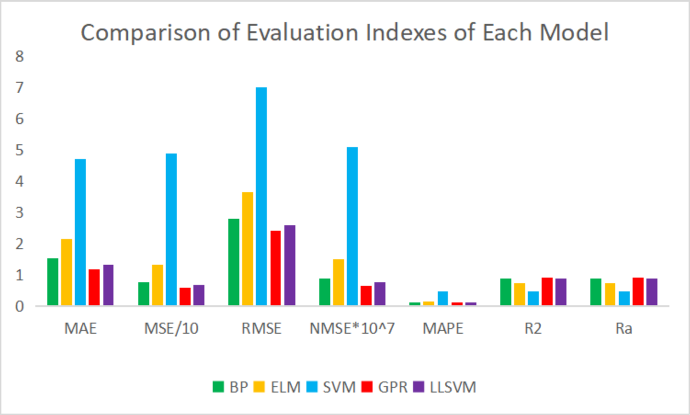


Figure 1: Comparison of Evaluation Indicators for Each Model

According to the comparison chart of each model evaluation index drawn in Figure 1, it can be seen that the GPR model has the best fitting effect, and the SVR model has the worst fitting effect. The results obtained are consistent with the results obtained from the data table and the fitting effect diagram of each model. GPR model and LSSVM model have better fitting effect than other models, and SVR model and ELM model have the worst fitting effect.

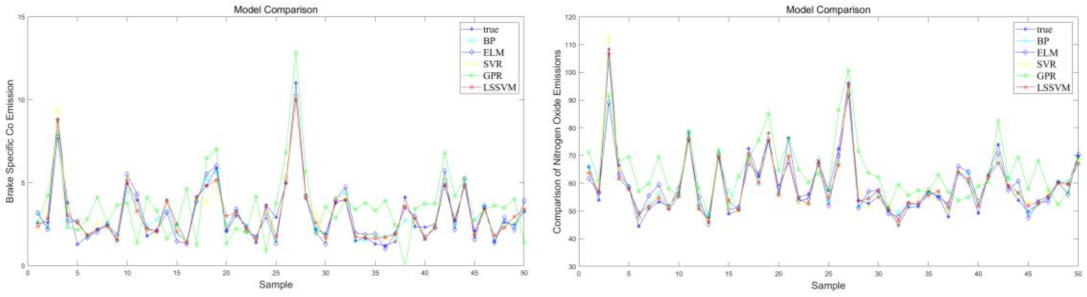
4.2. Model Comparison

It can be found by observing the prediction results of each model that BPNN, ELM and SVR model are general, and there are a big difference between the predicted value and the real value .Besides,the error between the true value and the predicted value of multiple samples is large.It can be seen from the result diagram of the model that the error between the predicted value and the real value of GPR and LSSVM is small.

In order to make the model prediction results more realistic, this study selected a lot of data. All data sets are used to establish the model, so that the fitting effect of the model is closer to the actual situation. In order to clearly show the fitting effect of each model, this paper selects 50 samples from the 2384 test data set to draw a comparison chart of the prediction results.

The fitting effects of the predicted values and real values of the five models are plotted in the same graph to observe the prediction effects of carbon monoxide and nitrogen oxide emissions. It can be seen from Figure 2 that the GPR model has the best fitting effect, followed by the LSSVM model. The fitting effect of the SVR model and the ELM model is poor, which is the same as the result obtained from the perspective of the evaluation index.

Therefore, in the fitting prediction of carbon monoxide and nitrogen oxide emissions from automobile engines, GPR model and LSSVM model are better.



(a) (b)

Figure 2: Comparison of prediction results of various regression models

5. Conclusions and Outlook

In this study, BP neural network, support vector machine, Gaussian regression, extreme learning machine and least square support vector machine models were used to perform regression prediction on engine carbon monoxide and nitrogen oxide emissions. The performance of each model is evaluated by 7 evaluation indicators.The determination coefficient R2 of GPR model is closest to 1, indicating that the model has the best fitting effect. At the same time, the MAE, MSER, MSE, NMSE and MAPE indexes are small, and the Ra value is the largest.Therefore, the GPR model is more accurate and effective in fitting the carbon monoxide and nitrogen oxide emissions of automobile engines. The model can be used to find the parameter value to minimize the pollutant emission of automobile engine, so as to reduce environmental pollution.

In this study, through the prediction comparison of five models, the most suitable model for predicting carbon monoxide and nitrogen oxide emissions is obtained, but some parameters of the model can still be optimized by optimization algorithm to improve the accuracy of the model.In the next step, we will use optimization algorithms to optimize the parameters of the GPR model to further improve the accuracy of the prediction model for carbon monoxide and nitrogen oxide emissions from automobile engines.

References

[1] An Jie Yang, Jian Ping Bi. Detection and Treatment Technology for Automobile Exhaust Pollution [J]. Advanced Materials Research, 2012, 1793 (518-523):

[2] Shan Guor, Ye Zhang, Guo Qiang Cai. Study on Exhaust Emission Test of Diesel Vehicles Based on PEMS [J]. Procedia Computer Science, 2020, 166:

[3] Boski P. Chauhan, G.J. Joshi, Purnima Parida. Car following model for urban signalised intersection to estimate speed based vehicle exhaust emissions [J]. Urban Climate, 2019, 29:

[4] Lanyi Zhang, Rongzu Qiu. Research progress on vehicle exhaust emission factor model and dispersion model [P]. Proceedings of the 2016 International Conference on Civil, Transportation and Environment, 2016.

[5] Yuan Mei, Li Chunyang. Research on Global Higher Education Quality Based on BP Neural Network and Analytic Hierarchy Process [J]. Journal of Computer and Communications, 2021, 09(06):

[6] Zhigui Guan, Yuanjun Zhao, Guojing Geng. The Risk Early-Warning Model of Financial Operation in Family Farms Based on Back Propagation Neural Network Methods [J]. Computational Economics, 2021 (prepublish):

[7] Fangbin Zhou,Lianhua Zou,Xuejun Liu,Yunfei Zhang,Fanyi Meng,Caichang Xie,Shanshan Zhang. Microlandform classification method for grid DEMs based on support vector machine [J]. Arabian Journal of Geosciences, 2021, 14(13):

[8] Ghasemi Porya, Karbasi Masoud, Zamani Nouri Alireza,Sarai Tabrizi Mahdi,Azamathulla Hazi Mohammad. Application of Gaussian process regression to forecast multi-step ahead SPEI drought index [J]. Alexandria Engineering Journal, 2021, 60(6):

[9] Annaby M.H., Said M.H., Eldeib A.M., Rushdi M.A. EEG-based motor imagery classification using digraph Fourier transforms and extreme learning machines [J]. Biomedical Signal Processing and Control, 2021, 69:

[10] Wang Lin, Zhou Hao, Yang Jie, Xiong Yonghua, She Jinhua, Chen Wei. A decision support system for tobacco cultivation measures based on BPNN and GA [J]. Computers and Electronics in Agriculture, 2021, 181:

[11] Niu Wenjing, Feng Zhongkai, Xu Yinshan, Feng Baofei, Min Yaowu. Improving Prediction Accuracy of Hydrologic Time Series by Least-Squares Support Vector Machine Using Decomposition Reconstruction and Swarm Intelligence [J]. Journal of Hydrologic Engineering, 2021, 26(9):

[12] Zhang Wei, Xie Peng, Li Yuxing, Zhu Jianlu. A machine learning model for predicting the mass transfer performance of rotating packed beds based on a least squares support vector machine approach [J]. Chemical Engineering and Processing - Process Intensification, 2021 (prepublish):

[13] DU Peng, MA Xiaoqi, WANG Zhuanping, MO Yuanfu, PENG Peng. A prediction method of missing vehicle position information based on least square support vector machine [J]. Sustainable Operations and Computers, 2021 (prepublish):