Prediction model and method of train body vibration based on bagged regression tree

Lele Peng, Wei Xu, Qianwen Zhong and Shubin Zheng

August 2, 2019

Prediction model and method of train body vibration based on bagged regression tree

Lele Peng [1,a], Wei Xu[1,b], Qianwen Zhong [1,c] , Shubin Zheng [1,d]

[1] College of Urban Railway Transportation, Shanghai University of Engineering Science, Songjiang, Shanghai 201620, China

[a] lele.peng@sues.edu.cn, [b] m15221625756_1@163.com, [c] qianwen.zhong@sues.edu.cn, [d] shubin.zheng@sues.edu.cn

Abstract

The vibration acceleration of train body is a key parameter reflecting the running state of train. It is necessary to obtain the acceleration accurately. But the traditional method has low precision. In this paper, a vibration acceleration prediction model and method of train body based on bagged regression tree is proposed. On the basis of GJ-5 to collect a large number of parameters of Guangzhou works section in Guangzhou-Shenzhen II line, Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient are used to analyze the correlation between train body vibration and other detection parameters. Then, the bagging regression tree algorithm is used to establish the prediction model of train body vibration. Finally, the training results are compared with the outputs of the model with multiple linear regression model, support vector machine and back propagation neural network. According to the evaluation index, the prediction accuracy of the bagged regression tree model is highest compared other three models, which is over 94%.

Keywords

Bagged regression tree; Train body vibration; Correlation analysis; Prediction model

1 Introduction

The high speed and heavy load development of the train lead to intense interaction between the train and the track, resulting in serious train body vibration. In addition, the light weight trend of high-speed trains leads to a decrease in the local stiffness of the train body, which is prone to resonance. Therefore, it is necessary to study the vibration of the train body. Besides, the vibration of train body is one of the important indicators to evaluate the ride comfort of passengers. Kumar et al.[1] calculated the comfort level of human body by establishing the human body biodynamic model with the train body vibration as the input. Stammen et al.[2, 3] established a neural network model with vehicle vibration as the input to calculate the comfort level. On the contrary, ride comfort index can also indirectly reflect the vibration of the train body, which has an early warning effect on the train safety.

At present, many scholars have studied the vehicle body vibration by means of signal processing and establishment of dynamic model. Zhang et al.[4] used the improved fast independent component correlation algorithm to separate the vibration signals of the train. The separation results demonstrated the effectiveness of the proposed method. Li et al.[5] analysed the track irregularity signal and train vibration signal by using Fourier transform and wavelet transform respectively, and adaptive filtering was applied to the train vertical suspension. The results showed the method reduces the vehicle vibration. Jabłońska et al.[6, 7, 8] established asymmetric vehicle-track coupled vibration models by considering many nonlinear factors. The simulation results showed that the vehicle vibration response is consistent with the actual operation. Ji et al.[9, 10] established train body vibration prediction model affected by various external factors through using the finite element method, and verified the accuracy of the model prediction by comparing the prediction results with the recorded results. Qian et al.[11] established a high-speed train vibration acceleration prediction model based on nonlinear autocorrelation neural network and multi-body dynamics model. By comparison, it is found that the output results of the two models are highly consistent. Xie et al.[12] used the deep belief network to extract the depth features of the vibration signal after Fourier transform. The superiority of the method was proved by comparing the actual data set with the simulated data set. The signal processing method is used to analyse the time-frequency domain waveform and statistical characteristics of the train body vibration is one-sided. And it is difficult to obtain accurate mathematical models based on the dynamics model of vehicle body vibration analysis. But the machine learning method has the characteristics of high efficiency and robust plasticity for train vibration analysis and prediction.Machine learning deals mostly with classification and regression problems. Machine learning methods include neural network, support vector machines, decision tree algorithm, integrated learning algorithm, etc.[13, 14, 15, 16, 17, 18]. SVM and decision tree are single classifiers. Compared with SVM, decision tree has the advantage of aggregating various types of data for accurate prediction and interpretation. Compared with neural networks, it does not need prior knowledge and is easier to explain. The proposed decision tree algorithms include ID3 decision tree algorithm, C4.5 decision tree algorithm, CART decision tree algorithm and Tsallis decision tree algorithm[19].The decision tree is divided into classification tree and regression tree. The difference between the two is mainly that the predicted values of the classification tree are discrete, and the predicted values of the regression tree are continuous. The integrated learning algorithm is a method of integrating multiple classifiers to improve prediction performance [18]. Bagged algorithm[20] and boosted algorithm[21] are two common integrated learning methods, and the difference of the two is that the training sets randomly selected by bagged algorithm are independent of each other, and the weights of the training sets of the boosted algorithm will be modified according to the training results of each time, training sets have dependencies. Therefore, bagged algorithm has advantages in training speed.

According to the parameters of the GJ-5 rail inspection vehicle, this paper selects the factors that have close correlation with the three-degree-of-freedom vibration of the vehicle body as the model sample set by correlation analysis, and the prediction model of the train body vibration is established by combining the regression tree algorithm and the bagged integrated algorithm. The proposed method overcomes the complex coupling of traditional dynamic models and the one-sidedness of signal processing method, and the research has great potential in preventing train body vibration and ensuring traffic safety.

2 Model and method

The construction of the model is mainly divided into two parts. The first part is the preprocessing of the data to get the sample set, and linear correlation algorithm and nonlinear correlation algorithm are used. The other part is to build a good regression tree prediction model through adjusting the parameters. The flow chart of the entire modeling process is shown in Figure 1.

2.1. Linear Correlation Algorithm

Pearson correlation coefficient is mainly used to calculate the correlation between continuous variables, and the population of variables is required to be normally distributed[22, 23]. Its calculation formula is:

$$r_{AB} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} . \tag{2}$$

Where $\bar{a}$ and $\bar{b}$ are the mean values of variables A and B, respectively. a_i and b_i are the observed values of variables A and B,respectively. n is the number of samples. The correlation coefficient r_{AB} has a value range of [-1,1]. The closer r_{AB} is to 1, the higher the linear correlation between A and B.

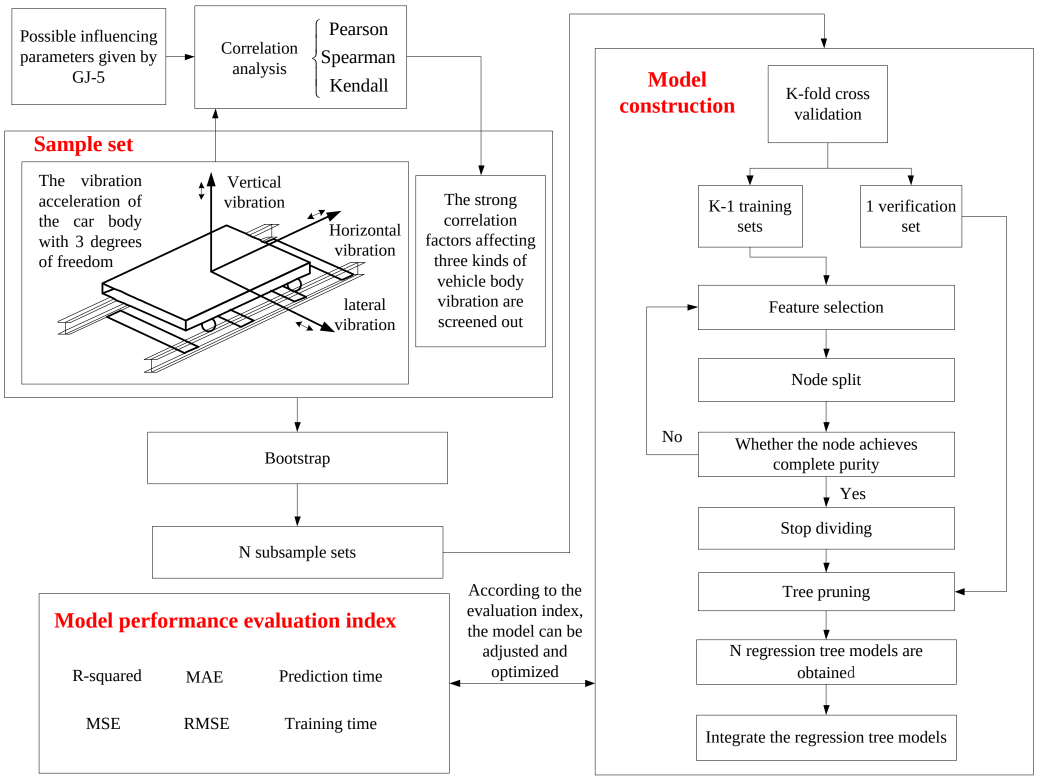

Figure 1. Modeling flow chart.

2.2. Nonlinear correlation algorithm

Spearman correlation coefficient is mainly used to measure the correlation between non-normal distribution or sequential variables. Its calculation is independent of the number and distribution of the original samples, only related to the ordering of the samples, and it is less sensitive to abnormal data values[24, 25]. The correlation coefficient is calculated as follows:

$$r_{AB}=\frac{\sum_{i=1}^{n}(m_i-\overline{m})(s_i-\overline{s})}{\sqrt{\sum_{i=1}^{n}(m_i-\overline{m})^2\sum_{i=1}^{n}(s_i-\overline{s})^2}}. \tag{3}$$

Where n is the number of samples of the variable, m_i and s_i are the ranks of a_i and b_i, respectively. The value of r_{AB} ranges from [-1,1].The core idea of the Kendall correlation coefficient is to judge the correlation based on the consistency of the ordered pairs between the two variables[26].When $a_i > a_j$ and $b_i > b_j$, or $a_i < a_j$ and $b_i < b_j$, the ordered pair is consistent;

when $a_i > a_j$ and $b_i < b_j$, or $a_i < a_j$ and $b_i > b_j$, the order pair is inconsistent; When $a_i = a_j$ or $b_i = b_j$, the correlation between the two cannot be effectively judged. The Kendall correlation coefficient is defined as follows:

$$r_{AB} = \frac{4P}{n(n-1)} - 1. \tag{4}$$

Where P is the number of consistent ordered pairs, n is the number of samples of the variable. The value range of the Kendall correlation coefficient is [-1,1].

2.3 Prediction model algorithm

The construction of the regression tree is mainly divided into two parts: the growth of the tree and the pruning of the tree[14].The principle of regression tree construction is to start from the root node, by comparing the attributes of the sample set with the feature nodes of the tree, and dividing the sample set into different child nodes according to the comparison result. Then the sample subset in each child node has a range of values for the attribute. If the child node can also perform attribute testing, the next comparison branch is selected until the leaf node is the final decision result, and the average value of all leaf nodes of the regression tree is the prediction result.

Suppose X and Y are input and output variables, respectively, and Y is a continuous variable, given the training data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ .A regression tree corresponds to a division of the input space and the output values on the divided cells. By selecting the j th variable $x(j)$ and its value s as the segmentation variable and the segmentation point, and defining two regions:

$$\alpha_1(j,s) = \{x | x^{(j)} \le s\}, \alpha_2(j,s) = \{x | x^{(j)} > s\} \tag{5}$$

The given data set is divided into L units $\alpha_1, \alpha_2, \ldots, \alpha_L$ and there is a fixed output value C_l on each unit α_l, so the regression tree model can be expressed as:

$$f(x) = \sum_{l=1}^{L} C_l I(x \in \alpha_l) \tag{6}$$

The loss function of the model is represented by the squared error $\sum_{x^i \in \alpha_l} (y_i - f(x_i))^2$. The optimal output value on each unit is solved using the criterion of square error minimization. According to the principle of least squares, it can be known that when C_l is the average of all actual values, the square error can be optimized:

$$C_l = ave(y_i | x_i \in \alpha_l) . \tag{7}$$

The core of regression tree growth is to determine the branching criterion of regression tree. By searching for the variable j, scanning the segmentation point s of the fixed segmentation variable j, selecting the minimum value pair (j,s). According to this step, the input space can be divided into two regions, and then repeat the above steps for each region until the node is pure. The specific steps are as follows:

1）Find the optimal segmentation variable j and the optimal segmentation point s make:

$$\min_{j,s}[\min_{c_1} \sum_{x_i \in \alpha_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in \alpha_1(j,s)} (y_i - c_2)^2] \tag{8}$$

2）Use the selected (j,s) to divide the area and determine the corresponding output value:

$$\alpha_1(j,s) = \{x | x^{(j)} \leq s\}, \alpha_2(j,s) = \{x | x^{(j)} > s\} \tag{9}$$

$$\hat{c}_l = \frac{1}{N_l} \sum_{x_i \in \alpha_l(j,s)} y_i , x \in \alpha_l , l = 1,2 . \tag{10}$$

By repeating the above steps for two subareas until the stop condition is met. The original regression tree is prone to overfitting. To avoid this phenomenon, the tree can be pruned using the validation data set. The specific principle is that when the training sample produces a new bifurcation, the test sample is tested to see if the bifurcation rule can be reproduced. If not, it is considered to be over-fitting and trimmed. K-fold cross validation is used to avoid overfitting in this paper. The K-fold cross validation method is an effective data modelling and inspection analysis technique[27].The basic idea is to divide the data set into K groups, where K-1 is used as the training set and the remaining one is used as the verification set to train the model. Repeat the above process K times to obtain K models, and finally use the average of the prediction accuracy of the k models as the final estimate of the model prediction accuracy. This allows all samples in the data set to participate in both the training and the testing.

For a single regression tree, a slight change in the training set is likely to completely change the structure of the tree, resulting in a low prediction accuracy. Compared with a single unstable learner, the integrated regression tree has a better prediction effect. The main idea is to learn a number of base learners by certain means, and then combine these base learners in some way for prediction. The bagged algorithm is an integrated learning algorithm. The main idea of bagged is to construct a training set with the same size but containing different training individuals by bootstrap through randomly sampling with the return, and each bootstrap sample contains about 63.2% of the original training set[20].In this way, the diversity between the base classifiers is increased, and the generalization ability of the learning system is improved. Bagged randomly extracts the training set, and the training sets are independent of each other, each round of training can be carried out in parallel. Therefore, the bagged model has obvious advantages in training speed. The general flow of the bagged regression algorithm is shown in Figure 2.

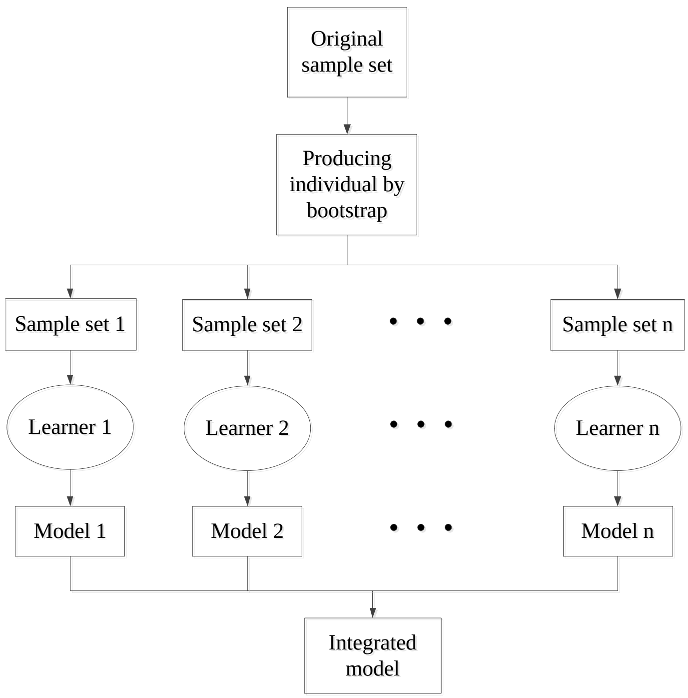

Figure 2. Flowchart of the bagged regression algorithm.

In the construction process of the bagged regression tree model, each time a training set is used to establish a regression tree, each tree has a predicted value. The result of the final model output is the average of all regression tree predictions. The mathematical model of the bagged regression tree is expressed as follows:

$$h_A(x) = sign(\sum h_i(x)) . \tag{11}$$

Where $h_i(x)$ is the predicted value of a single regression tree model.

3 Results and Discussion

In order to confirm the validity of the proposed method, experimental results of the proposed algorithm are provided to validate the performance. The data in this paper was collected on August 13, 2018 by the GJ-5 rail inspection vehicle. The collection section is Guangzhou Works Section of Guangzhou-Shenzhen II line. There are a total of 132 test items, each item contains 200,000 pieces of data, the detection mileage is 145km, and the maximum speed is 148km/h. According to prior knowledge, the vibration acceleration of train body with three degrees of freedom (horizontal, vertical and lateral) and 41 possible influencing factors were selected for analysis. The Pearson, Spearman and Kendall correlation algorithms are used to select the close correlation factors affecting the three kinds of the train body vibration. The train is relatively stable during the operation, so the factor with the correlation coefficient above 0.05 is selected as the close correlation parameter. The result is shown in Figure 3.

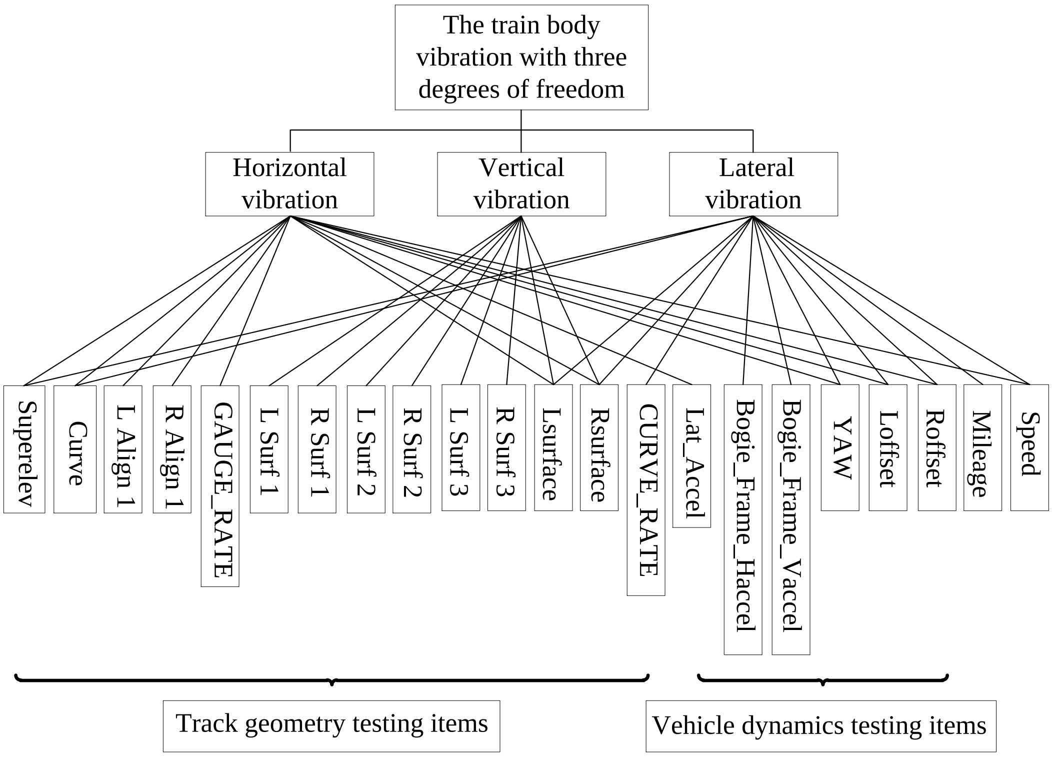

Figure 3. Strong correlation factors affecting train body vibration.

Table 1. Evaluation index of prediction model of train body horizontal vibration.

Model type	R-Squared	MAE	MSE	RMSE	prediction time (obs/sec)	Training time (s)
Multiple linear regression model	0.70	0.0874	0.0128	0.1133	92000	178.42
SVM	0.76	0.0698	0.0103	0.1017	190	70609
BP neural network	0.73	0.0765	0.0116	0.1077	350000	86
Bagged regression tree	0.86	0.0559	0.0061	0.0782	18000	1075.2

Table 2. Evaluation index of prediction model of train body vertical vibration.

Model type	R-Squared	MAE	MSE	RMSE	prediction time (obs/sec)	Training time (s)
Multiple linear	0.14	0.0925	0.0150	0.1224	350000	87.255

regression model						
SVM	0.16	0.0914	0.0146	0.1210	180	25628
BP neural network	0.17	0.0911	0.0145	0.1204	1505000	20
Bagged regression tree	0.36	0.0780	0.0111	0.1052	19000	784.69

Table 3. Evaluation index of prediction model of train body lateral vibration.

Model type	R-Squared	MAE	MSE	RMSE	prediction time (obs/sec)	Training time (s)
Multiple linear regression model	0.30	0.0563	0.0064	0.0801	230000	101.4
SVM	0.60	0.0410	0.0037	0.0604	180	46503
BP neural network	0.69	0.0444	0.0040	0.0630	17000	791
Bagged regression tree	0.80	0.0319	0.0018	0.0423	19000	949.73

Table 1 shows the performance indexes of the models for the horizontal vibration prediction of the train body. It can be seen from the table that the R-Squared of the bagged regression tree model is closest to 1 compared with the other three models, indicating that the degree of fit and the accuracy of the model prediction are higher; while the MAE, MSE and RMSE[28] of regression tree are lower than the other three models, indicating that the error between the prediction value and the actual value of the vehicle body horizontal vibration is small. As can be seen from the prediction speed and training time, the BP neural network has the fastest prediction speed and the shortest training time. But SVM has the slowest prediction speed and the longest training time.The performance index of the models for the vertical vibration prediction of the train body is shown in Table 2. It is apparent from the table that the value of R-Squared of the bagged regression tree model is 0.36.Compared with the vibration prediction of the other two degrees of freedom, this model has a larger deviation to predict the vertical vibration of the train body. From the perspective of correlation, the reason is that the factors affecting the vertical vibration of the train body are the surface conditions of the track. The characteristics of each parameter are similar, while the regression tree is sensitive to accurately extract the features of the data types with differences during the splitting process. Therefore, the prediction performance of the regression tree is insensitive for the factor of small difference of data feature. Compared with the other three models, MAE, MSE and RMSE of the bagged regression tree model is smallest, the prediction effect is better than the other three models, but the prediction speed is relatively slow and the training time is relatively long.Table 3 is the performance index of the models for predicting the lateral vibration of the train body. According to the parameters in the table, the results are similar to that of horizontal vibration of train body. The R-Squared of the bagged regression tree model is 0.8, which is better than the other three models. The MAE, MSE and RMSE indicators are smaller than the multiple linear regression model, SVM and the BP neural network, indicating that the regression tree has good prediction performance. However, the prediction speed and training time of regression tree are general, the speed of the multiple linear regression model is the best, and the speed of the support vector machine is the slowest.

In order to more intuitively highlight the prediction effect of the bagged regression tree model, the actual value of the train body vibration is compared with the predicted values of the bagged regression tree, the linear regression model, SVM, and the BP neural network. The x-axis represents the amount of data. The y-axis represent the vibration response of the train body. In order to more clearly compare the measured value with the predicted value, 500 samples are selected for partial enlargement. The vibration response of the train body with three degree of freedom is shown in following figures.

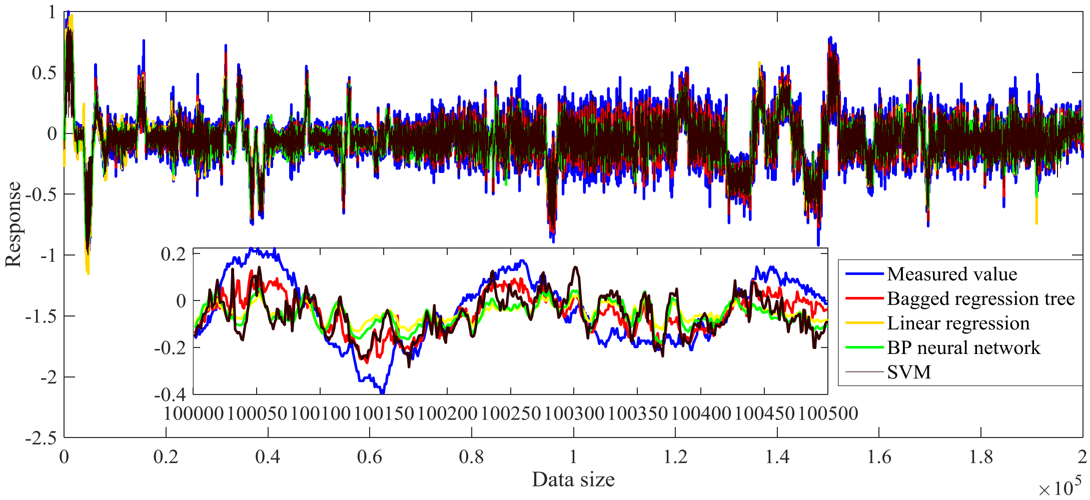

Figure 4. Horizontal vibration response diagram of train body.

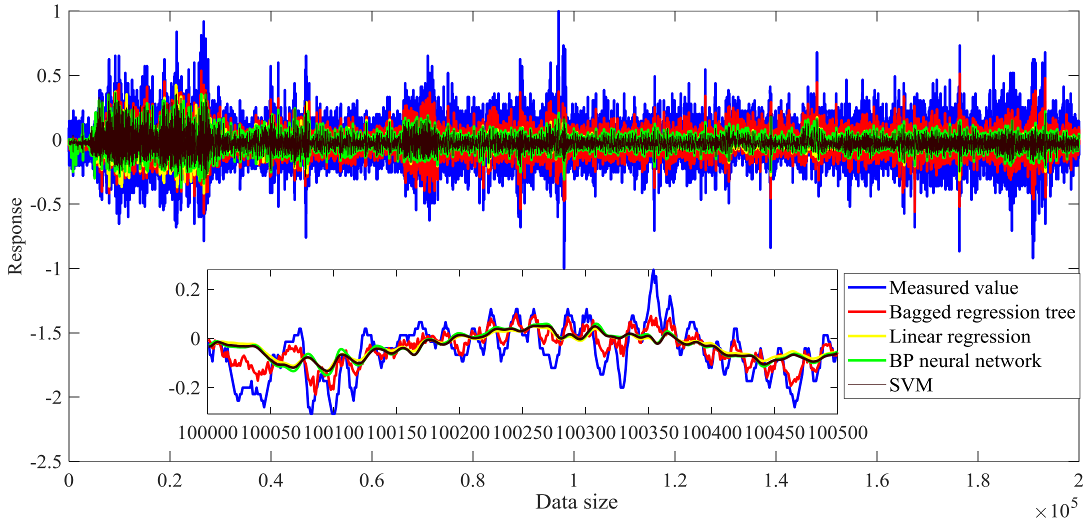

Figure 5. Vertical vibration response diagram of train body.

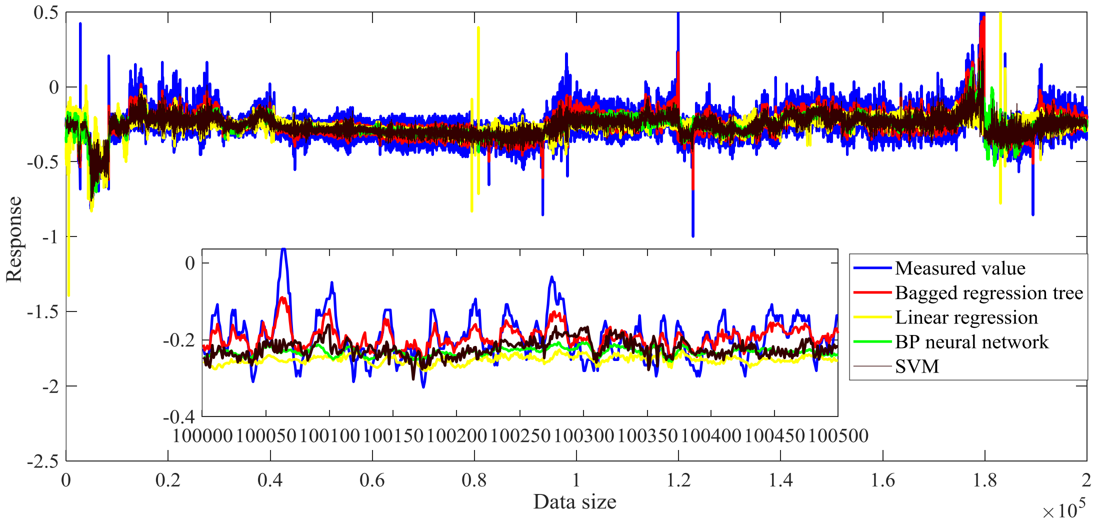

Figure 6. Lateral vibration response diagram of train body.

It can be seen from the overall graph in Figure 4 that the predicted curves are roughly the same as the measured curve. According to the partial enlargement, it can be seen that the prediction of the horizontal vibration of the bagged regression tree is obviously consistent with the fluctuation of the actual vibration. It can better reflect the actual horizontal vibration trend of the train body. The curves of the fluctuations of the linear regression model and the BP neural network are relatively flat, and the deviation is large compared the measured value. The predicted curve of SVM is steep, which does not reflect the actual vibration condition well.The overall graph in Figure 5 shows all models can generally reflect the vertical vibration of the train body, but the fluctuation of the vibration prediction curve of the four models is not obvious compared with the actual vibration curve. According to the partial enlargement, only the curve of regression tree has higher consistency with the actual vibration curve, while the curves of other three models are basically similar and have little fluctuation compared actual curve, which cannot reflect the actual vibration situation well. According to the overall diagram in Figure 6, there is a significant difference between the partial predicted and measured values of the linear regression model, which illustrates the prediction accuracy of this model is low. And from the obvious vibration fluctuation, the fluctuation of the bagged regression tree curve is consistent with the curve of measured value. It can be seen from the partial enlargement, the three curves of the linear regression model, BP neural network and SVM are very flat, and the deviation is larger compared with the measured value, but the curve of bagged regression tree can better reflect the actual trend of the train lateral vibration.

5 Conclusions

In this paper, the correlation analysis method is used to select the correlative factors from the testing items given by GJ-5, then the sample set of the prediction model is constructed. The bagged integration method is used to combine multiple sets of regression tree models, and the K-fold cross validation method is used to optimize the model, the prediction model of train body vibration is constructed. The experimental results are compared with the prediction results of linear regression model, BP neural network and SVM. It is shown that the prediction accuracy of train body vibration based on bagged regression tree is the highest than other three models, but the speed of model training and prediction needs to be further optimized.

Acknowledgement

This research was funded by National Natural Science Foundation of China [Grant No.: 51478258], and the Shanghai Committee of Science and Technology [Grant No.: 18030501300].

References

[1] Kumar V, Rastogi V, Pathak PM. Simulation for whole-body vibration to assess ride comfort of a low-medium speed railway vehicle. Simulation. 2017;93(3):225-236.

[2] Albers A, Lerspalungsanti S, Off S, et al. Generating virtual vehicle start-up using simulation models of the drive-train and evaluation of subjective comfort by means of Artificial Neural Networks (ANNs). Human Vibration: Effects on Health - Performance - Comfort; 2007 Oct 08-09. p. 229-245, Dresden, Germany.

[3] Stammen K, Meywerk M, Vdi. Usability of artificial neural networks to evaluate vehicle ride comfort - A study in the context of virtual vehicle development. Human Vibration: Effects on Health - Performance - Comfort; 2007 Oct 08-09. p. 197-214, Dresden, Germany.

[4] Zhang XY, Tu JL, Dong DC. Separation of the train vibration signal based on improved fast ICA. Proceeding of 4th International Conference on Intelligent Computation Technology and Automation(ICICTA); 2011 Mar 28-29; Shenzhen, Guangdong, China; 2011.p.912-914.

[5] Li GJ, Zhang J, Li C. Signal process technology of train vibration. Proceeding of 35th Chinese Control Conference(CCC); 2016 Jul 27-29; Chengdu, China; 2016.p.10187-10192.

[6] He ZX, Yang XW. Dynamic response analysis of an asymmetric coupled vehicle-track system generated by voided elastic two-block sleeper. Shock and Vibration. 2016; 2016:1-11.

[7] Hou K, Kalousek J, Dong R. A dynamic model for an asymmetrical vehicle/track system. Journal of Sound and Vibration.2003; 267(3):591 -604.

[8] Jabłońska J, Choromański W. Modelling, simulation and diagnostic on non-symmetric railway vehicle. Proceeding of 8th Mini Conference on Vehicle System Dynamics, Identification and Anomalies; 2002 Nov 11-13.p. 255-260, Budapest, Hungary.

[9] Galvín P, Domínguez J. Experimental and numerical analyses of vibrations induced by high-speed trains on the Córdoba-Málaga line. Soil Dynamics and Earthquake Engineering.2009; 29(4): 641-657.

[10] Ji ZL, Yang GW, Liu YB, et al. Analysis of vertical vibration characteristics of the vehicle-flexible track coupling system under wind load and track irregularity. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit. 2018; 232(10):2444-2455.

[11] Qian K, Liang J, Gao YH. The prediction of vibration and noise for the high-speed train based on neural network and boundary element method. Journal of Vibroengineering.2015; 17(8):4445-4457.

[12] Xie JP, Yang Y, Li TR, et al. Learning features from high speed train vibration signals with deep belief networks. Proceedings of the 2014 International Joint Conference on Neural Networks. IEEE International Joint Conference on Neural Networks (IJCNN); 2014 Jul 06-11. p. 2205-2210, Beijing, China.

[13] Cortes C, Vapnik V. Support-vector networks. Machine Learning.1995; 20(3):273-297.

[14] Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1(1):14-23.

[15] Malsburg CVD. Frank Rosenblatt: principles of neurodynamics: perceptrons and the theory of brain mechanisms. 1986.

[16] Quinlan JR. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publisher; 1992.

[17] Quinlan JR. Induction of decision trees. Machine learning.1986; 1(1):81-106.

[18] Zhou ZH. Ensemble methods: foundations and algorithms. Boca Raton: Taylor & Francis; 2012.

[19] Wang YS, Xia ST. Unifying attribute splitting criteria of decision trees by Tsallis entropy. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2017 Mar 05-09.p.2507-2511, New Orleans, LA.

[20] Breiman LJML. Bagging predictors. Machine learning. 1996;24(2):123-140.

[21] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, Boosting, and Variants. Machine Learning.1998; 1999 (1):1-38.

[22] Di LP, Margara L. Optimal global alignment of signals by maximization of Pearson correlation. Information Processing Letters. 2010; 110(16):679-686.

[23] Wiedermann W, Hagmann M. Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals. Commun Stat Theory Methods.2016; 45(21):6263-6283.

[24] Best DJ, Roberts DE. Algorithm AS 89: the upper tail probabilities of Spearman's rho. Journal of the Royal Statistical Society.1975; 24(3):377-379.

[25] Xiao C, Ye J, Esteves RM, et al. Using Spearman's correlation coefficients for exploratory data analysis on big dataset.Concurrency & Computation Practice & Experience. 2016; 28(14):3866-3878.

[26] Kendall MG. A new measure of rank correlation. Biometrika. 1938; 30(1-2):81-93.

[27] Wong TT. Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets. Pattern Recognition. 2017;65(11):97-107.

[28] Lele Peng,Shubin Zheng, Xiaodong Chai,Liming Li.A novel tangent error maximum power point tracking algorithm for photovoltaic system under fast multi-changing solar irradiances. Applied Energy.2018;210(2018):303-316.