IOP Conf. Series: Materials Science and Engineering **758** (2020) 012031 doi:10.1088/1757-899X/758/1/012031

Surface Temperature Prediction of Asphalt Pavement Based on GBDT

X Qiu[1], W Y Xu[1], Z H Zhang[2], N N Li[1] and H J Hong[1]

1 Road and Traffic Engineering Research Center, Zhejiang Normal University, Jinhua, 321004, China
2 Zhejiang Zhengda Testing Technology Co., Ltd., Jinhua, 321000, China
E-mail: zjnuhhj@outlook.com

Abstract. Asphalt is a temperature sensitive material, distribution characteristics and vary rules of asphalt pavement temperature have an important impact on the bearing capacity and performance of pavement, which is a concern of domestic and foreign researchers. The objective of this study was to explore the correlation between pavement temperature of asphalt pavements and meteorological factors and implement an accurate trend prediction of the asphalt pavement temperature. First, errors and missing data in the meteorological dataset were cleaned. Then, the three kinds of temperature prediction models of asphalt pavements in winter were established by Gradient Boosting Decision Tree (GBDT), Random Forest (RF) and Linear Regression (LR). The results indicate that GBDT would perform an excellent ability on prediction. The mean-square-error of the GBDT predicting results has a lower value of 1.5 when compared with the Random Forest and Linear Regression owing to the high robustness and the good generalization ability, which reflects the GBDT model has a good applicability in the field of prediction. The research would serve as a technical support for the machine learning algorithms applied in the field of the application of prediction problems.

1. Introduction

In terms of the properties of asphalt, it's a typical temperature-sensitive material, and asphalt pavement temperature is one of the important factors affecting road traffic safety. Changes in road surface temperature and state will have a direct impact on the safety and efficiency of road traffic, while meteorological conditions are the most precise causes affecting the change of pavement temperature. Therefore, how to effectively predict the change law of pavement temperature has important theoretical and practical significance for road traffic safety.

Now, domestic and foreign researchers have conducted extensive research on the temperature change of asphalt pavement, while their research direction is primarily focussed on the prediction model of pavement temperature field. In 1975, Barder [1] first used the heat conduction equation to establish a model for the highest temperature prediction of the road surface and performed statistical analysis. Subsequently, Huber [2] conducted a regression analysis of local temperature and road surface temperature. Then a simple regression prediction model is established for predicting the maximum temperature of the road surface. Then, Krsmanc [3] proposed a purely statistical approach for forecasting road surface temperature is based on stepwise linear regression analysis with appropriate selection of the input parameters. In summary, the current research method of asphalt pavement temperature still stays in statistical analysis and linear regression analysis. These methods are relatively simple with a high accuracy in most studies. However, as the amount of data increases, the process of model establishment becomes complicated, also the precision of the prediction model is hard to be guaranteed. Therefore, how to establish a high accuracy asphalt pavement temperature prediction model is the research emphasis.

Gradient Boosting Decision Tree (GBDT) belongs to decision tree-based models in machine learning which has been introduced by Professor Firedman [4] for the first time. The algorithm trains each of the decision trees in a sequential manner, and each iteration fits a decision on the residuals left by the previous one and then the prediction is accomplished by combining the trees [5-6]. In recent years, more and more scholars have begun to choose to improve the accuracy of predictive models in an iterative way, making GBDT become the new direction of forecasting [7]. For example, GBDT is applied to short-term subway ridership prediction in Ding's study, and the findings suggest that the GBDT model has considerable advantages in improving the short-term subway ridership forecast [8]. Similarly, Yang also proposed a short-term traffic prediction method based on gradient boosting machine, concluded that the prediction errors of GBDT are smaller than SVM and BPNN [9]. It's prediction effect is excellent. Then, Gradient boosting decision trees, an ensemble learning method, is proposed to make asphalt pavement temperature prediction based on the meteorological data in this study. And compared with the temperature prediction model built by random forest (RF) and linear regression (LR) to verify the applicability and performance of GBDT in the field of asphalt pavement temperature prediction.

This paper, in the light of related research, selects the winter meteorological observation data from December 2015 to January 2018 in Jinhua area of Zhejiang Province of China. The study analyzes the relationship between meteorological factors and road surface temperature and provides technical support for the accurate prediction of winter road surface temperature.

2. Data Preprocessing

2.1. Data Cleaning

The amount of meteorological data is huge, and it is inevitable that data is missing and wrong during the process of collection and storage. Data cleaning is intended to handle anomalous data in data sets for data mining. The study reads meteorological data through the Pandas function in the Python platform and uses the dropna function to delete the data set where the missing data are located. Finally, 3261 sets of valid data are generated.

2.2. Key Factor Determination

In order to determine the key factors affecting the pavement temperature, the relationship between meteorological factors and asphalt pavement temperature was linearly fitted to calculate the Pearson coefficient, and the calculation formula is as shown in the Equation (1). The meteorological data used for association analysis mainly include: air temperature, dew point temperature, air pressure, relative humidity, wind speed, precipitation, visibility and asphalt pavement temperature.

$$Pearson(x, y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \tag{1}$$

Where, x and y are two independent variables; $\bar{x}$ and $\bar{y}$ are the mean of two independent variables; n is the number of variables; i is the variable number.

Table 1 shows the results of Pearson coefficient, it can be known that the dew point temperature, air pressure, relative humidity and air temperature have a great influence on the temperature change of the asphalt pavement.

Table 1. Pearson coefficient of meteorological factors

Meteorological factors	Air temperature	Dew point temperature	Air pressure	Relative humidity	Wind speed	Precipitation	Visibility
Pearson coefficient	0.79	0.37	-0.38	-0.34	0.07	-0.11	0.06

IOP Conf. Series: Materials Science and Engineering **758** (2020) 012031 doi:10.1088/1757-899X/758/1/012031

3. Asphalt Pavement Temperature Prediction

In this part, key factors obtained by linear fitting are used as input variables and asphalt pavement temperature is used as the output variable to establish the prediction model of asphalt pavement temperature based on gradient boosting decision tree (GBDT). Two algorithms, random forest and linear regression, were introduced for comparative analysis.

3.1. Predictive Model Building Process

Taking the winter meteorological observation data in Jinhua area of central Zhejiang Province as the data sample of the model, using the train_test_split function in the Python platform, 3246 sets of data are randomly selected from 3261 sets of meteorological data to set up the asphalt pavement temperature prediction model, and the remaining 15 sets of data are used as test data to verify the prediction effect of the model.

Gradient Boosting Decision Tree is an integrated machine learning algorithm, which consists of Classification and Regression Tree (CART), Gradient Boosting and Learning Rate. During the establishment of the model, there are 5 main parameters that need to be tuned, such as learning rate, loss function, the number of trees, max-depth and max-features. The default parameter of the model learning rate is 0.1, in order to slow down the learning rate of the model and improve the accuracy of the model, the learning rate in this paper is set to 0.01; The loss function is used to calculate the residual of the prediction result. As shown in Equation (2), here the mean square loss function is used in study; Since there are only 4 features in the training sample, and all of them are related parameters, so the max-features is set to None which indicating that all features will be used for model building. On this basis, the max-depth of GBDT is first set as the default value, and the value of the number of trees can be set by multiple trial calculations. As is shown in Figure 1, when the number of trees reaches 1000, the mean absolute error of the GBDT prediction results tends to be stable, so the number of trees of the model is set to 1000. Similarly, the max-depth of GBDT is calculated. It can be observed in from Figure 2 that when the depth of the tree is 4, the average error of prediction result is the smallest, so the max-depth is set to 4.

$$L(y, f(x)) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{2}$$

Where, x are the input variables, y are the corresponding labels, $f(x)$ are the prediction results, and i is the iteration number of the prediction model.

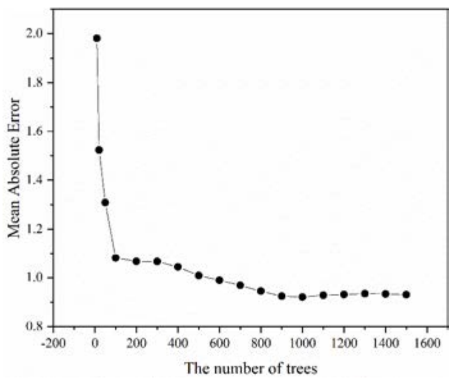

Figure 1. Prediction errors varying with the number of trees.

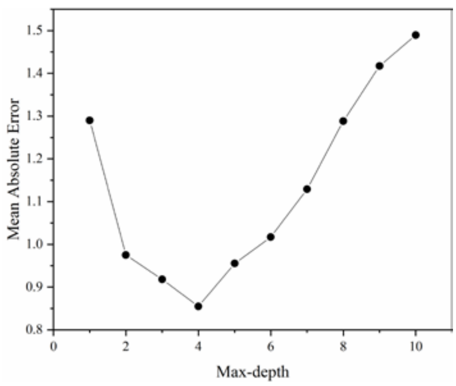

Figure 2. Prediction errors varying with max-depth.

Random Forest is an integrated machine learning algorithm, which consists of decision trees and bagging. During the establishment of the model, there are 3 main parameters that need to be tuned, such as max-depth, max-features and the number of trees. Since each tree in the random forest is independent of each other and the prediction results do not affect each other, so max-depth can be set to None which indicating that the depth of CART is not limited. The default setting parameter for max-features is auto, same as GBDT. Here the max-features is set to None. On this basis, the value of the number of trees can be set by multiple trial calculations. As is shown in Figure 3, when the number of trees reaches 12, the mean absolute error of the random forest prediction results tends to be stable, so the number of trees of the model is set to 12.

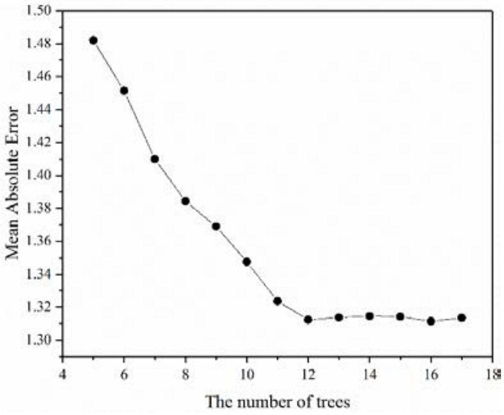

Figure 3. Random Forest predictive model building process.

Linear regression is a method to establish the relationship between dependent variable and multiple independent variables by using a linear model. The linear regression model finds the corresponding regression coefficient according to the input parameters and multiplies the regression coefficient by the corresponding independent variable to accumulate the dependent variable.

3.2. Model Prediction Result Analysis

The prediction results of each model are shown in Figure 4. In the figure, Test represents the test data, GBDT represents the gradient boosting decision tree prediction result, RF represents the random forest prediction result, and LR represents the linear regression prediction result.

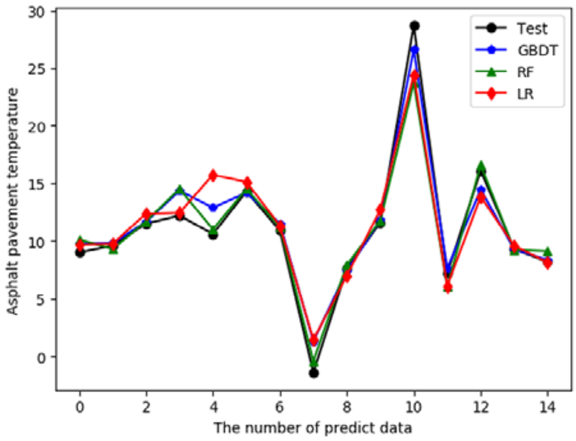

Figure 4. Asphalt pavement temperature prediction model test results.

In order to verify the accuracy of each asphalt pavement temperature prediction model, the Mean Square Error (MSE) of the prediction results of GBDT, random forest and linear regression models was calculated. The calculation formula is as shown in Equation (3), and the MSE variation trend of each temperature prediction model is shown in Figure 5.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{3}$$

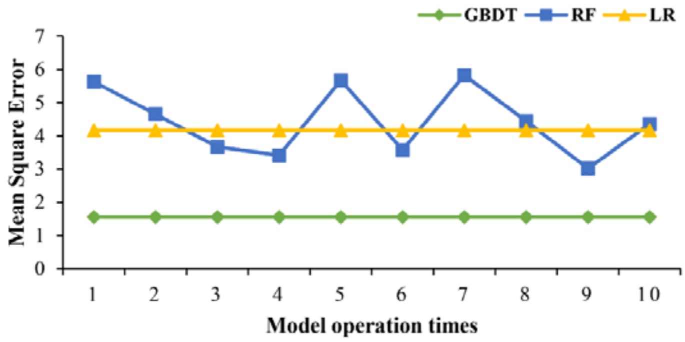

Figure 5. MSE variation trend for each temperature prediction model.

The analysis results show that the mean square error of GBDT prediction results is the smallest and the robustness is high, which has great advantages in asphalt pavement temperature prediction; the mean square error of random forest prediction result is always in the state of fluctuation; the prediction model built by linear regression is as stable as the GBDT model, but has little advantage in prediction accuracy. In conclusion, the model built by GBDT has good prediction effect, and has good generalization ability and robustness. It has good applicability in the prediction of asphalt pavement temperature.

4. Conclusion

The following major findings were achieved from this study:

(1) Both GBDT and Random Forest belong to the decision tree integration algorithm. In the case that the training samples have constraints, the GBDT has higher prediction accuracy than the random forest, so GBDT is more suitable for establishing the prediction model.

(2) The asphalt pavement temperature prediction model established by GBDT has good generalization ability and robustness, can predict the temperature change of asphalt pavement more accurately, and has certain feasibility for the prediction of pavement temperature. The study would serve as a technical support for the application of machine learning algorithms in prediction.

Acknowledgements

The authors are grateful to the financial supports from the Natural Science Foundation of Zhejiang Province (LY18E080020).

References

[1] Barber E S 1957 Calculation of Maximum Pavement Temperatures from Weather Reports Highway Research Board Bulletin
[2] Huber G A 1993 Weather Database for the Superpave™ Mix Design System Binders
[3] Krsmanc R, Slak A S and Demsar J 2013 Statistical Approach for Forecasting Road Surface Temperature Meteorological Applications **20** pp 439-446
[4] Friedman J H 2001 Greedy Function Approximation: A Gradient Boosting Machine Annals of Statistics **29** pp 1189-1232
[5] Friedman J H 2002 Stochastic Gradient Boosting Elsevier Science Publishers B. V. pp 367-378
[6] Liu L, Ji M and Buchroithner M 2017 Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra. Remote Sensing **9** pp 1299-1317
[7] Hong T, Pinson P and Fan S 2016 Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and Beyond International Journal of Forecasting **32** pp 896-913
[8] Ding C, Wang D and Ma X 2016 Predicting Short-Term Subway Ridership and Prioritizing its Influential Factors Using Gradient Boosting Decision Trees Sustainability **8** pp 1100-1116
[9] Yang S, Wu J and Du Y 2017 Ensemble Learning for Short-term Traffic Prediction Based on Gradient Boosting Decision Trees Journal of Sensor **2017** pp 1-15