Freeway Short-Term Travel Time Prediction Based on LightGBM Algorithm

Fujian Wang¹*, Huiling Cheng¹, Hongliang Dai² and Haihang Han²

¹ Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, 310058, China
² Zhejiang Scientific Research Institute of Transport, Hangzhou, 310023, China
*Corresponding author’s e-mail: ciewfj@zju.edu.cn

Abstract. In order to realize the prediction of freeway travel time, a short-term travel time prediction model based on LightGBM (Light Gradient Boosting Machine) is proposed under the influence of weather factors, time period factors, and traffic factors. These factors are called as the features used for increase prediction accuracy. The travel time of a single vehicle is determined by license plate recognition data of two adjacent video monitors in Shaoxing section of Shanghai-Hangzhou-Ningbo Freeway, and a better travel time data set is constructed by data preprocessing. The feature data are determined by Pearson correlation. Based on the analysis of main optimization parameters in LightGBM, the short-term average travel time is predicted, the MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) obtained by experiments are satisfying, indicating that LightGBM model has high accuracy and good fit. Finally, through comparison with KNN (K-Nearest Neighbor) model and GBDT (Gradient Boosting Decision Tree) model, the prediction accuracy and training speed both show that LightGBM has good advantages in predicting short-term freeway travel time.

1. Introduction

Travel time is not only a superior evaluator of freeway traffic flow status, but also a direct indicator of traffic efficiency. To effectively manage the huge highway traffic system, the highway video monitoring system tends to be perfect. The prediction of freeway travel time, which can take advantage of these massive traffic monitoring data, provides more reliable route choices for travellers, and improves travel comfort and efficiency.

At present, there are numerous scholars proposing various methods and models for travel time prediction. Zhang et.al [1] utilized a time-varying coefficient linear model to predict trip travel time. Their prediction time step size varies from 0-30min or more. In small data set, prediction accuracy is good. However, in a larger data set, the MAPE ranged from 8% to 13%, which means the accuracy needs to be improved. Fei et.al [2] proposed an adaptive dynamic linear model (ADLM) based on Bayesian theory to predict short-term travel time on freeways. 5 minutes as the prediction step size, the minimum MAPE is 2% and the maximum MAPE is 14% on the 12-day test set, indicating that prediction stability is not high enough. In the intelligent transportation environment, prediction models based on machine learning method have been proposed continuously. Wang [3] utilized the combination model of SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) to carry out prediction. The combined model performs better than a single model, but it also costs more computing power. Among many machine learning algorithms, a class of serialization methods represented by boosting began to appear in travel
time prediction. Zhang et al. [4] utilized GBM (Gradient Boosting Machine) to conduct travel time prediction experiment on Maryland freeway. Using 5min as prediction step, in the non-peak period, the prediction accuracy of the model can be maintained between 2%-4%, in peak periods, the model's prediction accuracy is between 9%-12%, which reflects that the prediction performance of GBM needs to be improved in the case of congestion. Besides, GBR (Gradient Boosting Regression Tree) [5] and GBDT (Gradient Boosting Decision Tree) [6] are gradually introduced into the traffic field and give full play to their advantages.

Among them, the tree-based ensemble learning algorithm can not only produce higher prediction accuracy, but also provide an explanation for the prediction results. LightGBM, which is proposed in recent years, belongs to this kind of model. It makes up for the shortcomings of other boosting algorithms, and its advantages are [7]: (1) Fast training; (2) Less memory consumption; (3) Higher model accuracy; (4) Support parallel learning; (5) Strong ability to process massive data. It has not been used in freeway travel time prediction, but it has been developed in medical and other fields. Wang et al. [8] used LightGBM to classify the characteristics of tumor patients and normal patients, and the recognition accuracy of LightGBM is better than RF (Random Forest) and XGBoost (eXtreme Gradient Boosting). Therefore, LightGBM has superiority in classification algorithm.

In the existing short-term travel time prediction, prediction step size is mostly in the range of 5min to 30min. In this paper, 5 minutes are selected as the short-term prediction step length, and combined with the driving data from Shaoxing section of Shanghai-Hangzhou-Ningbo Freeway, a short-term travel time prediction model based on LightGBM algorithm is proposed. At the same time, this paper shows the applicability and superiority of LightGBM in short-term travel time prediction.

2. LightGBM model

LightGBM is a fast and high-performance distributed gradient boosting framework based on decision tree algorithm. LightGBM is a variant of GBDT model, and the basic principle is still boosting tree algorithm. The boosting tree algorithm is shown as follows:

Initialize the boosting tree as shown in equation (1):

\[
f_0(x) = 0
\]

Iterate \( m \) times to get the boosting tree that contains \( M \) decision tree:

\[
f_m(x) = f_{m-1}(x) + T(x, \theta_m)
\]

\[
f_m(x) = \sum_{l=1}^{M} T(x, \theta_m)
\]

where \( f_{m-1}(x) \) is the existing model when the \( m \)th iteration, \( f_m(x) \) is the model obtained after the \( m \)th iteration, \( f_m(x) \) is the boosting tree including \( M \) decision trees, \( T(x, \theta_m) \) is the \( m \)th decision tree, and \( \theta \) is the parameter of the decision tree.

When the forward distribution algorithm proceeds to the \( m \)th step, given the current model \( f_{m-1}(x) \), solve the minimum loss function to obtain the \( m \)th decision tree \( T(x, \theta_m) \):

\[
\min \ L(f_m(x), y) = \sum_{i=1}^{N} L(f_{m-1}(x) + T(x, \theta_m), y_i)
\]

where \( y_i \) is the true value of the sample data, \( L(f_m(x), y) \) is the loss function.

GBDT employs the steepest descent approximation method and applies the negative gradient of loss function to fit the base learner. LightGBM adds Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to the GBDT model to reduce the time complexity of processing samples.

(1) GOSS is a sampling algorithm, which can achieve a good balance with the accuracy of decision tree while reducing the number of data. From the point of view of reducing samples, considering that the samples with large gradient have greater influence on information gain, the data with large information gain contribution are retained, and the data with small gradient are selected randomly. Random selection of data with smaller gradient may change the distribution of the original data set, LightGBM solves this by multiplying the small gradient samples with coefficients and increasing their weight, so as to pay more attention to the samples with insufficient training.

(2) EFB algorithm reduces the dimension of feature by binding features, and uses greedy strategy to establish weighted undirected graph by using the relationship between features. According to the degree
order of feature nodes, traverses the features, then traverses all feature clusters for each feature, selects some features to bind, and then merges mutually exclusive features through histogram algorithm. Through EFB algorithm, the training speed is improved without losing its accuracy.

In addition, the LightGBM algorithm selects "Leaf-wise" leaf growth strategy with depth restriction. This leaf node expansion method can reduce the training error and get better accuracy. Simultaneously, LightGBM performance has greatly improved by technology optimizations such as support for parallel learning, support of class features, and so on.

3. Data preparation

3.1. Data description
In this paper, the experiment utilized travel time data of two video monitors (No. 20301 and No. 20302) in the direction from Shangyu to Keqiao on the Shaoxing section of the Shanghai-Hangzhou-Ningbo Freeway. The collection time is from September 2nd to September 29th, 2019, which includes information such as license plate number, arrival time, and video monitor name. Calculate the travel time of each vehicle by matching the upstream and downstream license plates of the two video monitors, that is, the time difference between the two video monitors along the freeway.

3.2. Data preprocessing
There are abundant abnormal data in the actual data generated by the video recognition. The data should be filtered preliminarily: (1) Eliminate license plate recognition errors, abnormal characters or recognition failure data; (2) Eliminate the data whose travel time is obviously beyond the acceptable range, such as more than one day or negative value. The daily driving time of freeways generally follows a certain stable state. Avoiding the interference of long-time parking and serious speeding, the quartile method is used to further process the data. The quartile method arranges all the data from small to large and divides them into four equal parts. The first quartile \( Q_1 \) (at the position of 25% of the data), the median and the third quartile \( Q_3 \) (at the position of 75% of the data) are taken out. Let \( IQR = Q_3 - Q_1 \), the vehicle travel data within the reasonable range \([(Q_1 - 1.5IQR, Q_3 + 1.5IQR)]\) is reserved for prediction.

Through data preprocessing, the abnormal samples were removed, leaving 562,085 single vehicle travel data. The travel time is basically near 150-350s, and the data quality is improved and conforms to the real driving conditions.

4. Construction of prediction model

4.1. Input variable selection
The travel time is often influenced by weather factors, time period factors, and traffic factors, these factors are called as features in LightGBM. To construct a complete LightGBM model, the input data includes travel time data and feature data, the feature data can increase the prediction accuracy. In this paper, the travel time between the two video monitors is mainly concentrated in 200-300s. So, the prediction step is set as 5 minutes, and a day is divided into 288 time steps, \( t=1-288 \), \( t \) is the time step index. On the one hand, this step can reflect the change of travel time between video monitors, on the other hand, it can meet the actual forecast demand. So, the travel time is the average of all vehicle travel times in each interval of 5min. To determine the feature data, Pearson correlation analysis is performed on multiple features. The correlation is shown in the following table 1:

| Feature  | hour  | week | weather | \( T_{t-1} \) | \( T_{t-2} \) | \( T_{t-3} \) |
|----------|-------|------|---------|-------------|-------------|-------------|
| Correlation coefficient | 0.317 | 0.022 | -0.201  | 0.767       | 0.779       | 0.775       |

Through the feature correlation analysis, this paper finally selects the hour, week, weather and traffic status of the previous several steps as the feature inputs. Specifically the hour period of the day is indexed...
from 0-23; the week period of the day is indexed from 1-7; the weather of the day is sunny or rainy, 0: sunny, 1: rain; the output of the LightGBM model is travel time at time step \( t \) denoted as \( T_t \), the three traffic status variables that are used as input to predict travel time at time step \( t \) are as follows: \( T_{t-1} \), \( T_{t-2} \), and \( T_{t-3} \) are three most recent travel time observations at time steps \( t-1 \), \( t-2 \), and \( t-3 \).

4.2. Analysis of model parameters
The main optimization parameters in LightGBM are as follows:

(1) num_leaves: the number of leaves per tree, the larger num_leaves is, the greater the complexity of the tree model, which is the main parameter to improve the accuracy;

(2) min_data_in_leaf: the minimum number of data in a leaf, its appropriate value is related to the number of training data and num_leaves;

(3) max_bin: the maximum number of features subpackage. The larger max_bin is, the higher the accuracy will be, but the training speed will be slowed down;

(4) max_depth: the maximum depth of each tree. The deeper the tree is, the higher the accuracy is. However, if max_depth is set too large, it is easy to overfit. In LightGBM, leaf-wise tree is very easy to be overfitted. In order to get the best leaf-wise tree, num_leaves and max_depth are usually controlled. Considering num_leaves and max_depth together, num_leaves < \( 2^{\text{max depth}} \) is generally followed.

The parameters mentioned above have different adjustment methods for different optimization purposes. After understanding their influence on the performance of the model, the best parameters can be found by grid search method and other parameter optimization methods.

5. Model analysis

5.1. Model result analysis
In this paper, the data of the first three weeks from September 2nd to September 22nd, 2019 are used as the training set, and the last week as the test set. Grid search method is used to find the appropriate parameters. The final parameters are as follows: learning_rate=0.1, num_leaves=31, min_data_in_leaf=35, max_bin=350, max_depth=-1. The average travel time obtained from the test is used as the predicted value. MAE and MAPE are common indicators to measure the accuracy of prediction. This paper chooses to use them to measure the prediction accuracy of LightGBM. Their calculation formula are as follows:

\[
\text{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |Y_i - y_i| \\
\text{MAPE} = \frac{100}{n} \cdot \sum_{i=1}^{n} \left| \frac{Y_i - y_i}{y_i} \right|
\]

where \( Y_i \) is the predicted value, \( y_i \) is the actual value, and \( n \) is the number of samples.

The smaller the MAE and MAPE are, the higher the prediction accuracy will be. Table 2 shows: MAE and MAPE are not only within the acceptable range, but also maintained at a relatively good level, which indicates that the LightGBM model performs well in the short-term travel time prediction of the freeway.

| Table 2. Prediction error |
|--------------------------|
| Error index      | MAE (s) | MAPE (%) |
| value            | 5.46    | 2.55     |

The travel time examples are shown in figure 1. By comparing the predicted and actual values, especially when the travel time fluctuates greatly, the prediction model can fit such changes well. It can be seen from the figure that the model is very effective in predicting the travel time of this road segment.
5.2. Model comparison analysis

By comparing the prediction errors of several prediction models, the applicability and superiority of the LightGBM method for travel time prediction are verified. The other two types of models used in this section are as follows:

1) KNN model, a commonly used supervised learning algorithm, has been widely used in the field of traffic prediction [9]. The prediction is based on the information of K training samples nearest to the measured samples. Because it only saves the samples in the training stage, the KNN model has the characteristics of simple and fast training speed.

2) GBDT model, also known as gradient descent tree model, is a learning model based on regression tree and boosting. GBDT model, which is one of the most commonly used models in supervised learning model, has been already involved in the field of travel time prediction.

The comparison results are shown in table 3. It can be seen from the two indexes that the error values of the three models are all small, which can reasonably predict the travel time of the freeway. However, the two indexes of LightGBM are the smallest, which indicates that the prediction accuracy is better than the other two models. Because the data set selected in this paper is not huge, from the perspective of training speed, KNN as a representative model of lazy learning, its training time complexity is low. But this also confirms its shortcomings: it relies on training data and has less explanatory than GBDT and LightGBM. The training time of GBDT is 0.42s, and the training time of LightGBM is 0.18s, which shows that the training speed of LightGBM is greatly improved compared with the similar boosting models. The comparison with other two models shows that LightGBM has achieved better performance in model accuracy and training speed.

| Model  | KNN (s) | LightGBM (s) | GBDT (s) |
|--------|---------|--------------|----------|
| MAE    | 5.78    | 5.46         | 6.08     |
| MAPE   | 2.69    | 2.55         | 2.83     |

6. Conclusion

This paper uses the data of two video monitors to get the travel time of some sections of freeway. By analysing the influencing factors of travel time, feature data are extracted as model input. Through model parameter analysis, LightGBM model is established to predict the short-term average travel time. Finally comparing with the KNN model and GBDT model, through the display of evaluation indicators such as MAE and MAPE, it shows that the LightGBM proposed in this paper has good advantages in travel time prediction.

In the future research, further comparisons will be made with more prediction models to verify the advantages of LightGBM in travel time prediction. At the same time, LightGBM will be used in a larger
data set to improve its stability and accuracy in travel time prediction, making the model more widely used in traffic forecasting.

Acknowledgments
This work was supported by Science and technology plan project of Zhejiang Provincial Transportation Department [grant number 2019047].

References
[1] Zhang, X.Y., Rice, J.A. (2003) Short-term travel time prediction. Transportation Research Part C: Emerging Technologies, 11: 187–210.
[2] Fei, X., Lu, C.C., Liu, K. (2011) A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transportation Research Part C: Emerging Technologies, 19: 1306–1318.
[3] Wang, X., An, K., Tang, L., Chen, X.H. (2015) Short term prediction of freeway exiting volume based on SVM and KNN. International Journal of Transportation Science and Technology, 4: 337–352.
[4] Zhang, Y.R, Haghani, A. (2015) A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58: 308–324.
[5] Zhan, X.B., Zhang, S.C., Szeto, W.Y., Chen, X.M. (2020) Multi-step-ahead traffic speed forecasting using multi-output gradient boosting regression tree. Journal of Intelligent Transportation Systems, 24: 125–141.
[6] Cheng, J., Li, G., Chen, X.H. (2019) Research on travel time prediction model of freeway based on gradient boosting decision tree. IEEE Access, 7: 7466–7480.
[7] Ke, G.L., Meng, Q., Finley, T., Wang, T.F., Chen, W., Ma, W.D., Ye, Q.W., Liu, T.Y. (2017) LightGBM: A highly efficient gradient boosting decision tree. In: the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA. pp. 3149-3157.
[8] Wang, D.H., Zhang, Y., Zhao, Y. (2017) LightGBM: An effective miRNA classification method in breast cancer patients. In: the 2017 International Conference on Computational Biology and Bioinformatics (ICCBB 2017). Newark, NJ. pp. 7–11.
[9] Robinson, S., Polak, J.W. (2005) Modeling urban link travel time with inductive loop detector data by using the K-NN method. Transportation Research Record Journal of the Transportation Research Board, 1935: 47–56.