Hindawi
Journal of Sensors
Volume 2017, Article ID 7074143, 15 pages
https://doi.org/10.1155/2017/7074143

Research Article

Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine

Senyan Yang,[1] Jianping Wu,[1] Yiman Du,[1] Yingqi He,[2] and Xu Chen[2]

[1]*Institute of Transportation Engineering, Department of Civil Engineering, Tsinghua University, Beijing 100084, China*
[2]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Senyan Yang; senyanyang@126.com

Received 19 December 2016; Revised 5 March 2017; Accepted 19 March 2017; Published 4 May 2017

Academic Editor: Fanli Meng

Short-term traffic prediction is vital for intelligent traffic systems and influenced by neighboring traffic condition. Gradient boosting decision trees (GBDT), an ensemble learning method, is proposed to make short-term traffic prediction based on the traffic volume data collected by loop detectors on the freeway. Each new simple decision tree is sequentially added and trained with the error of the previous whole ensemble model at each iteration. The relative importance of variables can be quantified in the training process of GBDT, indicating the interaction between input variables and response. The influence of neighboring traffic condition on prediction performance is identified through combining the traffic volume data collected by different upstream and downstream detectors as the input, which can also improve prediction performance. The relative importance of input variables for 15 GBDT models is different, and the impact of upstream traffic condition is not balanced with that of downstream. The prediction accuracy of GBDT is generally higher than SVM and BPNN for different steps ahead, and the accuracy of multi-step-ahead models is lower than 1-step-ahead models. For 1-step-ahead models, the prediction errors of GBDT are smaller than SVM and BPNN for both peak and nonpeak hours.

1. Introduction

Massive traffic data have been constantly collected from a variety of sensors, such as inductive loop detectors, GPS-equipped vehicles, and mobile phones [1], promoting the development of data-driven intelligent transportation systems (ITS) [2]. Short-term traffic prediction is one of the most dynamic and typical researches in ITS, aiming at estimating the traffic state in the near future (within a few minutes) based on the historical traffic data [3, 4]. The prediction traffic information is essentially useful for travelers to make better travel planning in the pretrip stage or reschedule in the en route trip [5]. Accurate short-term traffic prediction is the first important step for real-time route guidance [6] and is quite critical in advanced travelers' information systems (ATIS) and advanced traffic management systems (ATMS) [7].

Traditional statistical approaches for short-term traffic prediction, such as ARIMA [8] and Kalman filtering technique [9], take advantages of the significant temporal dependencies of the historical univariate time series data of traffic variables. These methods usually assume model structures beforehand and estimate model parameters from the historical data later, with enough interpretability. It is easy for the prediction accuracy to be affected by the unstable traffic conditions, such as the traffic condition at peak hours [10].

Nonstationary and nonlinearity are the basic characteristics of traffic variables [11]. A variety of data-driven approaches have been applied for short-term traffic prediction, capturing the nonlinear relationship among the variables. Higher prediction accuracy can be acquired by these nonparametric machine learning (ML) methods, including Back Propagation Neural Network (BPNN) [12, 13], Support Vector Machine (SVM) [14, 15], and k-nearest neighbor algorithm (KNN) [16]. These methods belong to supervised learning method, and the target variables need to be prepared for the dataset beforehand, focusing on learning the relationship between the response and predictors [17]. The underlying information in the massive traffic data can be efficiently captured by these ML methods, achieving good prediction performance, but lacking interpretability [18].

Considering the freeway traffic condition independent of signalization, most short-term traffic prediction algorithms have been conducted and verified based on the freeway traffic data [3]. In the past decades, most researches focus on the prediction of traffic variables at one specific site of interest, solely considering the effect of its own previous traffic information. Actually, the traffic prediction performance for the given site is considerably influenced by the neighboring traffic condition. Spatial and temporal correlations were taken into account when performing short-term traffic prediction [6, 19, 20]. The traffic condition at a specific site is closely related to that of the upstream and downstream traffic condition. Multivariate traffic flow prediction model was constructed, improving the prediction performance by incorporating upstream traffic flow series as the transfer function input of ARIMA [21]. The influence of upstream and downstream traffic on the traffic condition of the given site is not symmetric [22]. The relationship between the current traffic speed at the given location and the past traffic speeds at the upstream and downstream locations was explored through cross correlation analysis [10].

The information provided by the traffic variables of neighboring sites can be used to improve the traffic prediction performance for the given site [10]. In this study, based on the freeway traffic data collected by the detectors, the historical upstream and downstream traffic volume are considered into the variables of prediction models. Actually, the traffic state variation of adjacent detectors is correlative. For many ML models, the effects of the input variables on the model output are difficult to interpret, and when the redundant or irrelevant variables are added, the prediction performance may get worse.

In order to capture the complex nonlinearity of traffic variation and identify the importance of variables, gradient boosting decision trees (GBDT) method, a tree-based ensemble learning method, is proposed to make short-term traffic prediction in this study. GBDT is a relatively new robust and accurate method in the machine learning field, which can cover different types of variables and identify the effects of upstream or downstream traffic on the traffic prediction of the given site, achieving excellent performance over classical methods. The main goal of this study is to identify the relative importance of input variables and enhance the accuracy of short-term traffic prediction.

Ensemble learning is one of the most popular and promising machine learning methods, which can improve the prediction performance by combining large numbers of weak base models [23]. The most commonly used ensemble techniques include boosting, bagging, and stacking. Different with other ML methods, the interaction between the input variables and prediction models can be interpreted, and the relative importance of critical factors can be identified by ensemble learning [24]. Tree-based ensemble methods, combining multiple simple decision trees, have been applied to handle prediction and classification problems in the transportation field, such as random forest, gradient boosting machine, and boosted regression trees. The prediction or classification output of model is the weighted summation or voting of the prediction of base trees. Random forests algorithm into AdaBoost algorithm is applied to estimate and predict traffic flow and congestion [25]. Stochastic gradient boosting is used to identify crashes with a superior classification performance [26]. The nonlinear relationships in the traffic accident data and the main effects of crucial variables are investigated by the boosted regression trees [27].

Additionally, the tree-based models on the basis of the random forest algorithm in the bagging framework are independently trained by uniformly and randomly sampling with replacement from the original dataset, strengthening the robustness, which can be trained by parallel computing. For each splitting node of the based trees, features are randomly selected [28]. Significantly different from the random forest, the tree-based models of GBDT are trained sequentially, and each base model is added to correct the error produced by its previous tree models. For each step, the samples misclassified by previous models are more likely to be selected as the train data, producing more accurate prediction performance. Comparing with the simple single tree model, GBDT is more stable with better prediction performance and interpretability by combining the output results of base trees [24].

The main contribution of this study is that the short-term traffic flow prediction models on the basis of gradient boosting machine are constructed, focusing on the influence of upstream and downstream traffic condition simultaneously and achieving a higher prediction accuracy than conventional machine learning methods. GBDT algorithm provides a flexible framework to adopt different combinations of the upstream and downstream historical traffic volumes as the input variables, which can capture the complex traffic nonlinearity, cover the hidden traffic patterns, and identify the relative importance of variables, and is of good interpretability. In addition, GBDT can resist the outliers of variables and perform well with partly erroneous data without cleaning [26].

2. Methodology

Single decision tree is a fast but instable algorithm, easily affected by the small perturbations in the training data [18], but the performance can be significantly improved by ensemble techniques [26]. Gradient boosting regression trees algorithm (GBDT) is viewed as combining the strengths of boosting algorithms and decision trees. Friedman [29] proposed the gradient boosting machines (GBM), based on a gradient descent formulation of boosting methods, which is suitable for regression and classification problems. Boosting framework is essentially a constructive strategy of ensemble formation, sequentially adding new weak base models which are trained with respect to the error of the former whole ensemble model for each iteration, and these base learners just produce a slightly lower error rate than random guessing [30].

The approximation accuracy and execution speed of gradient boosting can be generally improved by randomly subsampling the training data to fit the base learner at each iteration, also called stochastic gradient boosting [31], which is employed to make the short-term traffic volume prediction in this study, simultaneously considering the influence of the

Initialize $f_0(x)$ with a constant, $f_0(x) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$.
For $m = 1$ to M do:
For $i = 1, 2, \ldots, N$ compute the negative gradient g_{im}
$$g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$
End;
Fit a new regression tree $h_m(x)$
Compute the best gradient descent step-size ρ_m
$$\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, f_{m-1}(x) + \rho_m h_m(x))$$
Update $f_m(x) = f_{m-1}(x) + \rho_m h_m(x)$
End;
Output $f_M(x)$

Algorithm 1: Generic gradient boosting decision trees algorithm.

upstream and downstream traffic. The output of short-term traffic prediction model is the traffic volume of the future time at the given site, and the input is the historical volume at the past 1 or 2 or 3 time steps of the given site and its adjacent sites. Similar to other supervised learning methods, GBDT needs to be trained by the dataset with target labels, denoted as $(\mathbf{x}, \mathbf{y})^N$, and $\mathbf{x} = (x_1, \ldots, x_n)$ are the input variables and $\mathbf{y} = (y_1, \ldots, y_n)$ are the corresponding labels of the response variable. To find out the optimal combination of trees, GBDT algorithm adopts the forward stagewise technique and minimizes the loss function by sequentially adding a new base learner (single tree) to the expansion at each iteration without adjusting the parameters of the existing trees that have already been added [23]. The loss function $L(f)$ in using the estimated function $f(x)$ to predict $\mathbf{y}$ based on the training data is defined as

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i)). \tag{1}$$

With regard to the continuous response variables, the classical squared-error L_2 loss is employed in this prediction model, resulting in consecutive error-fitting in the process:

$$L(f)_{L_2} = \sum_{i=1}^{N} \frac{1}{2}[y_i - f(x_i)]^2. \tag{2}$$

In the boosting framework, when the algorithm is repeated for M iterations, the overall ensemble function estimate $\widehat{f}(x)$ is expressed in the additive functional form:

$$\widehat{f}(x) = \sum_{i=0}^{M} \widehat{f}_i(x), \tag{3}$$

where $\widehat{f}_0(x)$ is the initial guess and $\widehat{f}_i(x)$ $(i = 1, 2 \cdots M)$ are the function increments. The new base learners are constructed to be maximally correlated with the negative gradient of the loss function [30]. For the mth iteration, the negative gradient is defined as

$$g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}. \tag{4}$$

$g_m(x)$ is the local direction where $L(f)$ decreases the most rapidly at $f(x) = f_{m-1}(x)$. $h_m(x)$ denotes the base learner model and the gradient descent step length ρ_m is computed as

$$\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, f_{m-1}(x) + \rho_m h_m(x)). \tag{5}$$

For each step, adding a new base tree is to correct the mistakes made by its previous base learners [18]. Thus, the current model is updated as

$$f_m(x) = f_{m-1}(x) + \rho_m h_m(x). \tag{6}$$

To sum up, the generic gradient boosting decision trees algorithm for regression is shown in Algorithm 1. ($f_0(x)$ is just a single terminal node decision tree.)

In the process of gradient boosting, weighted resampling is carried out to put emphasis on observations which are more difficult to predict accurately. The value of each observation is reestimated once the new regression tree is added. The observations with lower prediction accuracy are assigned with a higher weight. The sampling weight is updated at the end of each iteration, and the observations with lower accuracy would be sampled with higher probability at the next iteration [26].

The input variables are seldom of equal relevance for the prediction performance, and usually only some of them have substantial influence on the model output [32]. Breiman et al. [33] proposed a measure method of relative variable importance for the single decision tree models. The importance of the variable x_i is denoted as $I_j^2(T)$, which is based on the number of times that a variable is selected for splitting in the tree weighted by the squared improvement to the model as a result of each split [32]. As a tree based ensemble method, the importance of the variable x_i for the GBDT model is simply averaged over all trees:

$$I_j^2 = \frac{1}{M} \sum_{m=1}^{M} I_j^2(T_m). \tag{7}$$

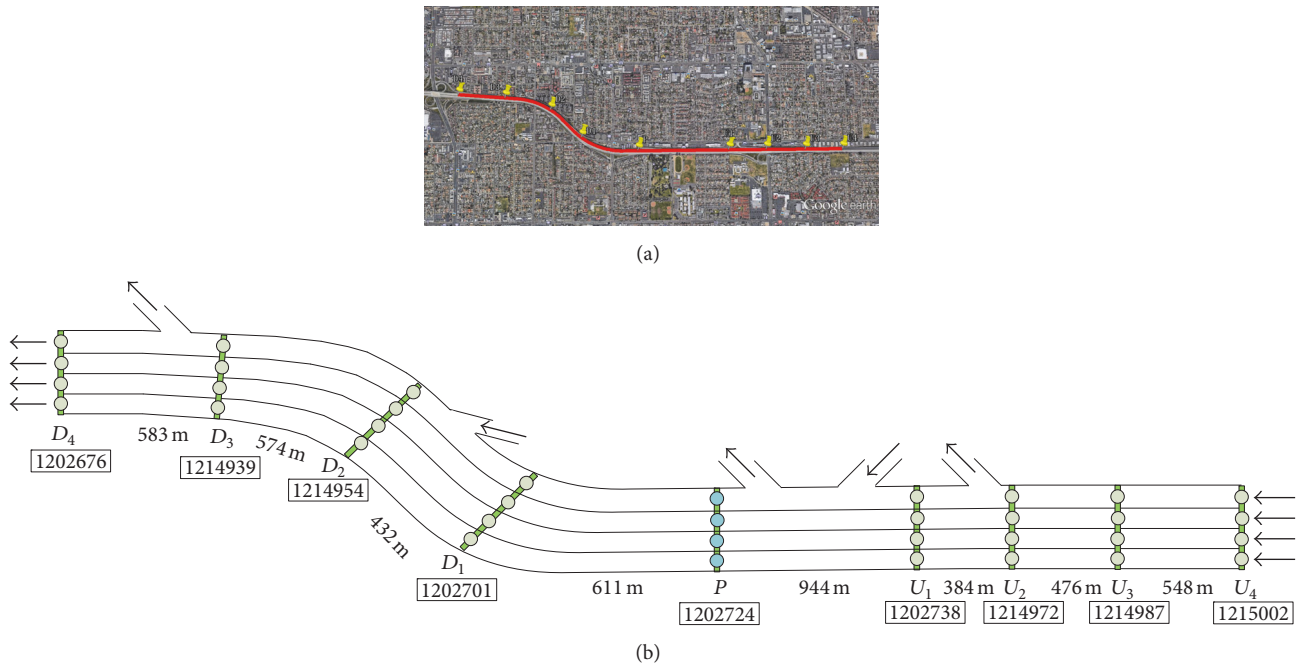

Figure 1: Locations: (a) the selected road segment and (b) detectors.

The importance of all the input variable is further standardized to make sure that they add up to 100%, which can be used for feature selection procedures [30].

3. Data Description

The data used in this study is downloaded from the open-access traffic flow database of Caltrans Performance Measurement System (PeMS) (http://pems.dot.ca.gov/). We collected the traffic volume data of 9 loop detectors located in State Route 22, Garden Grove, USA, from April 4 to June 5, 2016, lasting for 9 weeks. The detailed located information of the selected road segment and 9 detectors is shown in Figure 1. The traffic volume of four lanes is aggregated into one time series, recorded every 5 minutes. The traffic volume data of first eight weeks are used to train the traffic prediction model based on GBDT, while the last week of data serves as the testing set to identify the prediction accuracy of models. Detector P (1202724) is the target detector for traffic prediction, and Detectors U_1 (1202738), U_2 (1214972), U_3 (1214987), and U_4 (1215002) are the upstream detectors of P, while D_1 (1202701), D_2 (1214954), D_3 (1214939), and D_4 (1202676) are the downstream detectors of P. The length of the selected segment is 4552 m, with three exits and two entrances. The distance between two adjacent detectors is shown in Figure 1(b). The traffic volume variation of the given site is closely related to the upstream and downstream traffic condition. The traffic volume profile of 9 detectors on Wednesday, June 1, 2016, is shown in Figure 2. The basic statistics of the collected data for each detector is shown in Table 1, and the traffic volume values of 9 detectors are similar, and tiny differences of the 7 statistical indicators are mainly

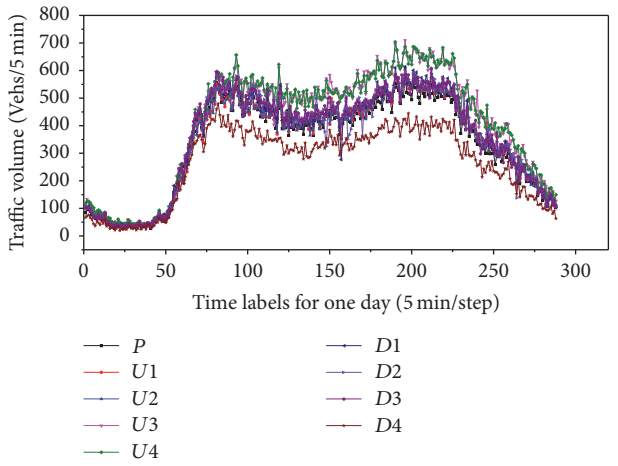

Figure 2: Traffic volume profile of 9 detectors for one day.

generated by the traffic flow at exits and entrances. The "25th," "50th," and "75th" are the 25th, 50th, and 75th percentiles of observations when ranking the traffic volume data in an ascending sort order for each detector.

The predictor response of the short-term traffic volume prediction models is the traffic volume of Detector P at time step t, denoted as V_t, which is related to the previous historical traffic volume of Detectors P, U_1, U_2, U_3, U_4, D_1, D_2, D_3, and D_4. All the possible variables used as the input are as follows: $V_{t\text{-}1}, V_{t\text{-}2}, V_{t\text{-}3}$ are the traffic volume of Detector P at time steps t-1, t-2, and t-3; U_1-$V_{t\text{-}1}$, U_1-$V_{t\text{-}2}$, U_1-$V_{t\text{-}3}$, U_2-$V_{t\text{-}1}$, U_2-$V_{t\text{-}2}$, U_2-$V_{t\text{-}3}$, U_3-$V_{t\text{-}1}$, U_3-$V_{t\text{-}2}$, U_3-$V_{t\text{-}3}$, U_4-$V_{t\text{-}1}$, U_4-$V_{t\text{-}2}$, and U_4-$V_{t\text{-}3}$ are the traffic

Table 1: Statistics of traffic volume data (Vehs/5 min).

Detectors	Mean value	Standard deviation	Min	Max	25th	50th	75th
P	312	162	15	620	161	358	440
$U1$	323	166	14	635	173	372	453
$U2$	324	167	14	619	172	372	453
$U3$	386	198	17	726	204	453	541
$U4$	389	198	18	738	207	458	543
$D1$	335	176	13	669	173	384	473
$D2$	338	176	15	669	173	389	480
$D3$	389	198	18	738	207	458	543
$D4$	248	130	12	535	128	286	351

volume of the 4 upstream detectors at time steps t-1, t-2, and t-3; D_1-V_{t-1}, D_1-V_{t-2}, D_1-V_{t-3}, D_2-V_{t-1}, D_2-V_{t-2}, D_2-V_{t-3}, D_3-V_{t-1}, D_3-V_{t-2}, D_3-V_{t-3}, D_4-V_{t-1}, D_4-V_{t-2}, and D_4-V_{t-3} are the traffic volume of the 4 downstream detectors at time steps t-1, t-2, and t-3; lastly, considering that the traffic volume varies greatly across different time period during one day, the time of day should be considered as an input variable, which is represented by time step *Time*. Each time step is 5 min, and there are 288 time steps for one day.

4. Experiments and Discussion

In this section, the experiment results of the short-term traffic prediction models based on GBDT are discussed in detail. The subsampling fraction is set as 0.5, signifying that 50% of the training data observations are randomly selected to propose the next tree in the expansion at each iteration. On account of randomness, similar but different fits are acquired when running the same model, and thus the prediction accuracy for each model is set as the average of 20 groups of experimental results slightly fluctuating in a small range. The minimum number of observations in the tree terminal nodes is set as 10.

4.1. Parameter Optimization. The performance of GBDT algorithm varies with the different parameter settings, including number of trees M, the maximum depth of variable interactions J, and learning rate R. In order to acquire the optimal prediction model, the effect of different parameter setting on the prediction performance is studied in this section. To uncover the influence of parameters setting on the prediction performance, the input variables and data are set to be the same for the experiments. $V_{t-1}, V_{t-2}, V_{t-3}$ and *Time* are selected as the input variables. Mean absolute percentage error (MAPE) and mean absolute error (MAE) are used to measure the prediction error, which are defined as

$$\begin{aligned} \text{MAPE} &= \left(\frac{1}{N} \sum_{t=1}^{N} \left| \frac{\widehat{V}_t - V_t}{V_t} \right| \right) \times 100\%, \\ \text{MAE} &= \frac{1}{N} \sum_{t=1}^{N} \left| \widehat{V}_t - V_t \right|, \end{aligned} \tag{8}$$

where V_t and $\widehat{V}_t$ are the real and predicted traffic volume at time t of the given site, respectively.

The maximum depth of variable interactions J refers to the number of nodes in a tree, signifying the tree complexity. More complex variable interactions hid in data can be captured by the larger J. Number of trees M is equivalent to the number of iterations and the number of base models in the additive expansion. When the other parameters are fixed, the larger M is, the more complex the model is, and more computational time will be required, which may cause overfitting more easily and produce poor performance on the observations not included in the training dataset [18]. In order to prevent the overfitting, the number of gradient boosting iterations needs to be controlled. In this study, 5-fold cross-validation is applied to check the prediction performance and acquire the optimal iteration number. For example, with the parameter setting of $J = 3$ and $R = 0.05$, MAPE and MAE varying with the increasing of M are shown in Figure 3, and it can be seen that when $M > 100$, the errors fluctuate slightly.

In order to achieve a better prediction performance, the range of J and R is set as $3 \leq J \leq 6$ and $0.001 \leq R \leq 0.5$ through conducting the preliminary experiments. Figure 4 indicates the influence of variable interactions J and learning rate R on the optimal iteration number and prediction errors. The complexity of base trees is represented by the variable interaction J. For a given learning rate R, the higher J is, the more complex the model is, and the fewer trees are needed to be added. Thus, the larger iteration number is preferable when setting a smaller J to produce high prediction accuracy.

The contribution of each base model can be adjusted by learning rate R. When the learning rate R is set to be a higher value, the prediction errors dropped to the lowest with fewer iterations, but the prediction errors are significantly higher than those with a smaller R setting. For example, when $R = 0.5$, the optimal iteration number is less than 200, but MAPE is higher than 0.08, and MAE is higher than 18. More trees need to be added for the smaller R setting, requiring more computational time. Overall, through weighing the computation and accuracy, $R = 0.01$ or 0.05 is more suitable for these traffic prediction models to produce better prediction performance with fewer iterations. Overfitting can be prevented by setting a smaller R to restrict the contribution

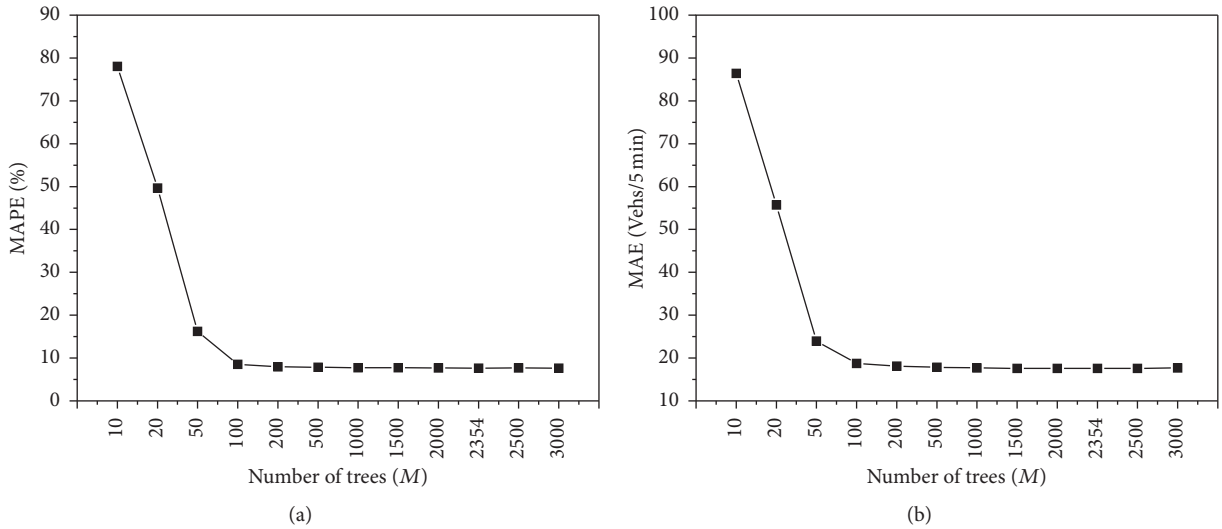

Figure 3: Prediction errors varying with M: (a) MAPE and (b) MAE.

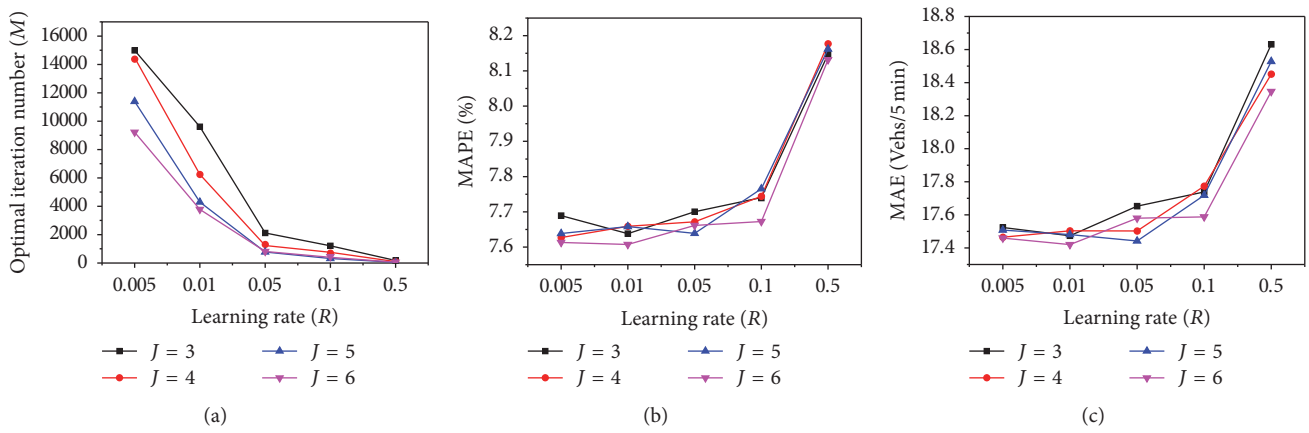

Figure 4: Influence of parameters on (a) optimal iteration numbers, (b) MAPE, and (c) MAE.

of each base tree. In addition, the optimal setting of the parameters M, J, R varies with training datasets, and the prediction models based on GBDT need to be retrained for other road segments.

4.2. Prediction Performance. GBDT provides a flexible framework to adopt various combination of different types of attributes as input variables for the prediction models. Firstly, 5 min (1-step) ahead short-term prediction models based on GBDT algorithm are built to uncover the effects of the upstream and downstream traffic condition on the prediction accuracy. The detailed information of 15 models is shown in Table 2. In order to compare the prediction performance of different models, balancing the computation and prediction accuracy, the parameters setting for the 15 models is $J = 5$ and $R = 0.05$. The input variables are the different combinations of historical traffic volume of Detectors P, U_1, U_2, U_3, U_4, D_1, D_2, D_3, and D_4 at time steps t-1, t-2, and t-3, and the response is the traffic volume of Detector P at the next time step t. Through comparing the prediction accuracy of different models, the optimal variable combination can be acquired for the freeway short-term traffic prediction model.

The prediction accuracy of GBDT models is ranked as shown in Table 2. The top three high-accuracy models are Model 10, Model 15, and Model 4, signifying that the upstream traffic condition has more positive impact on the prediction accuracy of GBDT models. In particular, MAPE and MAE reach the minimum at Model 10, just considering the influence of upstream historical traffic volume. Interestingly, the prediction accuracy of Model 11 and 14 is the lowest, just taking the downstream traffic volume as the input of GBDT. Generally, the prediction accuracy of GBDT models is lower when considering more downstream traffic variables as the input variables.

Furthermore, the prediction accuracy of short-term prediction for a given site is influenced by the upstream and downstream traffic condition on the freeway. The GBDT models considering the neighboring traffic condition tend to outperform the traditional simple temporal prediction

Table 2: "5 min ahead" prediction models based on GBDT.

Model	Detectors for prediction	Number of input variables	Spatial factors	MAPE	MAE	Accuracy sorting
Model 1	P	4	No	0.07652	17.53	11
Model 2	P, U_1	7	Upstream	0.07570	17.3	7
Model 3	P, D_1	7	Downstream	0.07648	17.51	10
Model 4	P, U_1, U_2	10	Upstream	0.07480	17.22	3
Model 5	P, D_1, D_2	10	Downstream	0.07690	17.61	12
Model 6	P, U_1, D_1	10	Upstream and downstream	0.07580	17.32	8
Model 7	P, U_1, D_1, U_2	13	Upstream and downstream	0.07510	17.3	4
Model 8	P, U_1, D_1, D_2	13	Upstream and downstream	0.07630	17.48	9
Model 9	P, U_1, U_2, D_1, D_2	16	Upstream and downstream	0.07560	17.4	6
Model 10	$P, U1, U2, U3$	13	Upstream	0.07458	16.99	1
Model 11	$P, D1, D2, D3$	13	Downstream	0.08583	22.14	15
Model 12	$P, U1, U2, U3, D1, D2, D3$	22	Upstream and downstream	0.07738	18.15	13
Model 13	$P, U1, U2, U3, U4$	16	Upstream	0.07550	17.17	5
Model 14	$P, D1, D2, D3, D4$	16	Downstream	0.08504	21.98	14
Model 15	$P, U1, U2, U3, U4, D1, D2, D3, D4$	28	Upstream and downstream	0.07464	17.20	2

models (Model 1), and the prediction performance can be enhanced by adding the neighboring traffic information to the input of models.

4.3. Relative Importance of Variables. In the training process of the GBDT models, the number of times that a variable is selected for splitting in the trees can be described by the relative importance. The relative importance of each variable for Models 1~15 based on GBDT can be conveniently computed, identifying the effects of input variables on the model output and prediction accuracy, as shown in Figure 5. The contribution of the same variables to the performance of different models is diverse. For example, the relative importance of $V_{t\text{-}1}$ in Model 1 is 76.7%, while that in Model 9 is 55.0%. The ranking of the variable importance also varies greatly among different models. For example, the importance of D_2-$V_{t\text{-}1}$ ranks fourth in Model 8 and ranks second in Model 9.

The immediate previous traffic volume $V_{t\text{-}1}$ of Detector P is the most important variable for the 15 GBDT models, and we could consider that $V_{t\text{-}1}$ is the most frequently selected variable to split the terminal nodes in decision trees when training the GBDT models, which is also in accordance with the actual situation that the traffic state in the near future tends to be influenced by the traffic just happening previously [18]. The variable $V_{t\text{-}2}$ of Detector P is the second important input variable for Model 1, Model 2, Model 3, and Model 11, while U_2-$V_{t\text{-}1}$ is for Model 4, Model 7, Model 10, and Model 13, D_1-$V_{t\text{-}1}$ is for Model 6, Model 8, and Model 12, D_2-$V_{t\text{-}1}$ is for Model 5 and Model 9, and D_4-$V_{t\text{-}1}$ is for Model 14 and Model 15. Moreover, when more variables of upstream or downstream detectors are considered for prediction, the models show less reliance on the historical temporal variables of themselves. For example, the importance of $V_{t\text{-}1}$ in Model 15 is about 45%, much lower than that of the other models. With the increasing of the neighboring traffic information, the importance of upstream and downstream traffic variables is improved in the GBDT models, and the prediction performance is enhanced simultaneously. The importance of upstream traffic condition on the traffic prediction accuracy for the given site is not equal to that of downstream traffic. Considering both the prediction accuracy and importance ranking of variables, the historical traffic variables of adjacent detectors should be added to the short-term traffic prediction models.

From the temporal perspective, the importance of traffic volume of the 9 detectors at time steps t-2 and t-3 is lower than that at time step t-1 in GBDT models. The importance of variable *Time* is significant for the 15 models, for the reason that the traffic volume of each detector varies greatly across the different time periods, and the fluctuation in the short term is irregular and complex. Therefore, the prediciton models for peak and nonpeak hours would be discussed in the following.

4.4. Multi-Step-Ahead Traffic Prediction Models. The Support Vector Machine (SVM) and Back Propagation Neural Network (BPNN) have been widely used in short-term traffic prediction on the freeway, which are trained for each combination of input variables in Table 2. The accuracy of 5 min (1-step) ahead GBDT prediction models is compared with that of SVM and BPNN based on 20 groups of repeated experiments for each model, as shown in Figure 6. The prediction errors of GBDT are significantly smaller than those of SVM and BPNN.

To identify the performance of GBDT, SVM, and BPNN approaches with different prediction horizons, the 10 min (2-step) and 15 min (3-step) ahead traffic prediction models are built to compare with the 5 min (1-step) ahead prediction model. The accuracy of 10 min and 15 min ahead

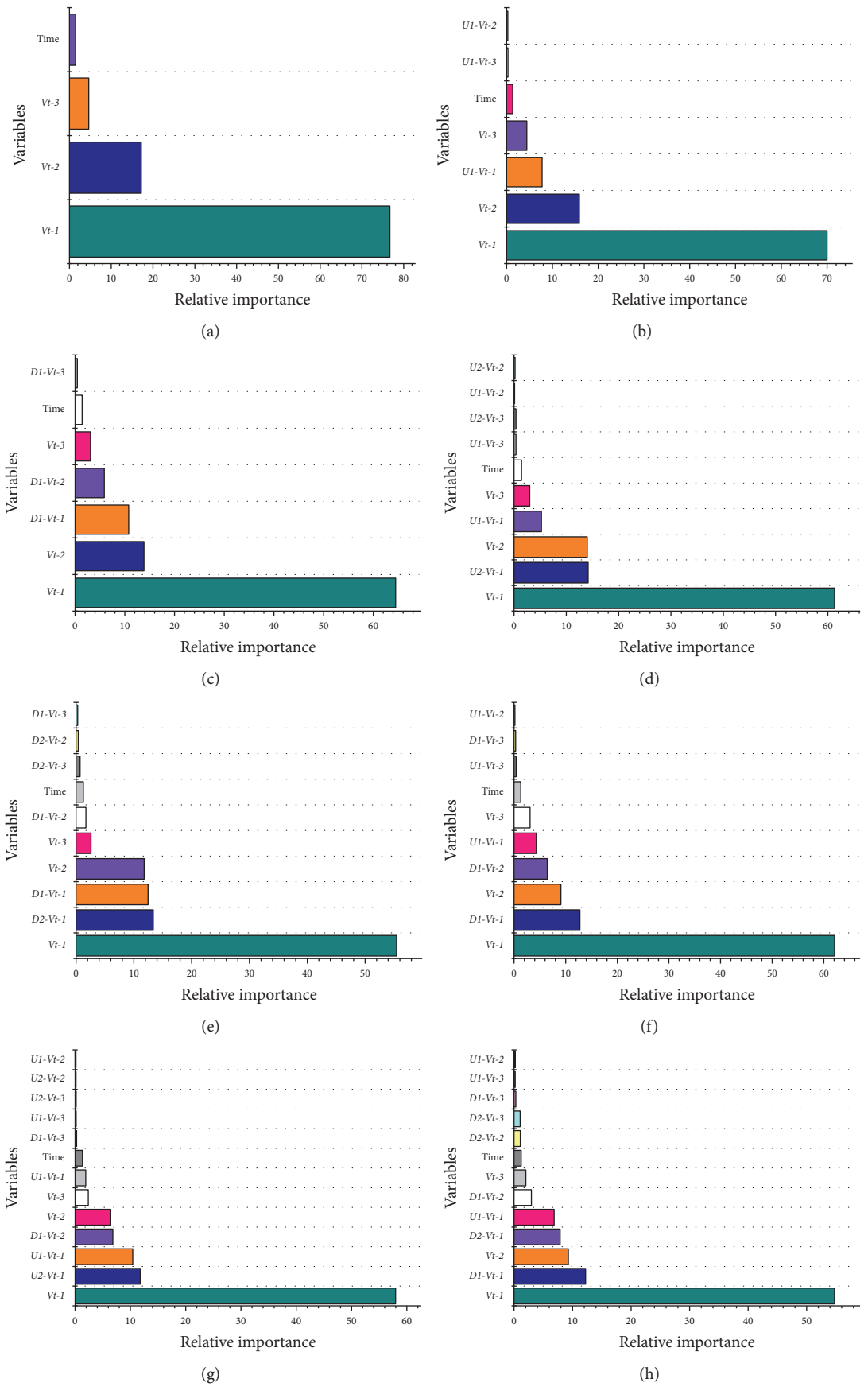

Figure 5: Continued.

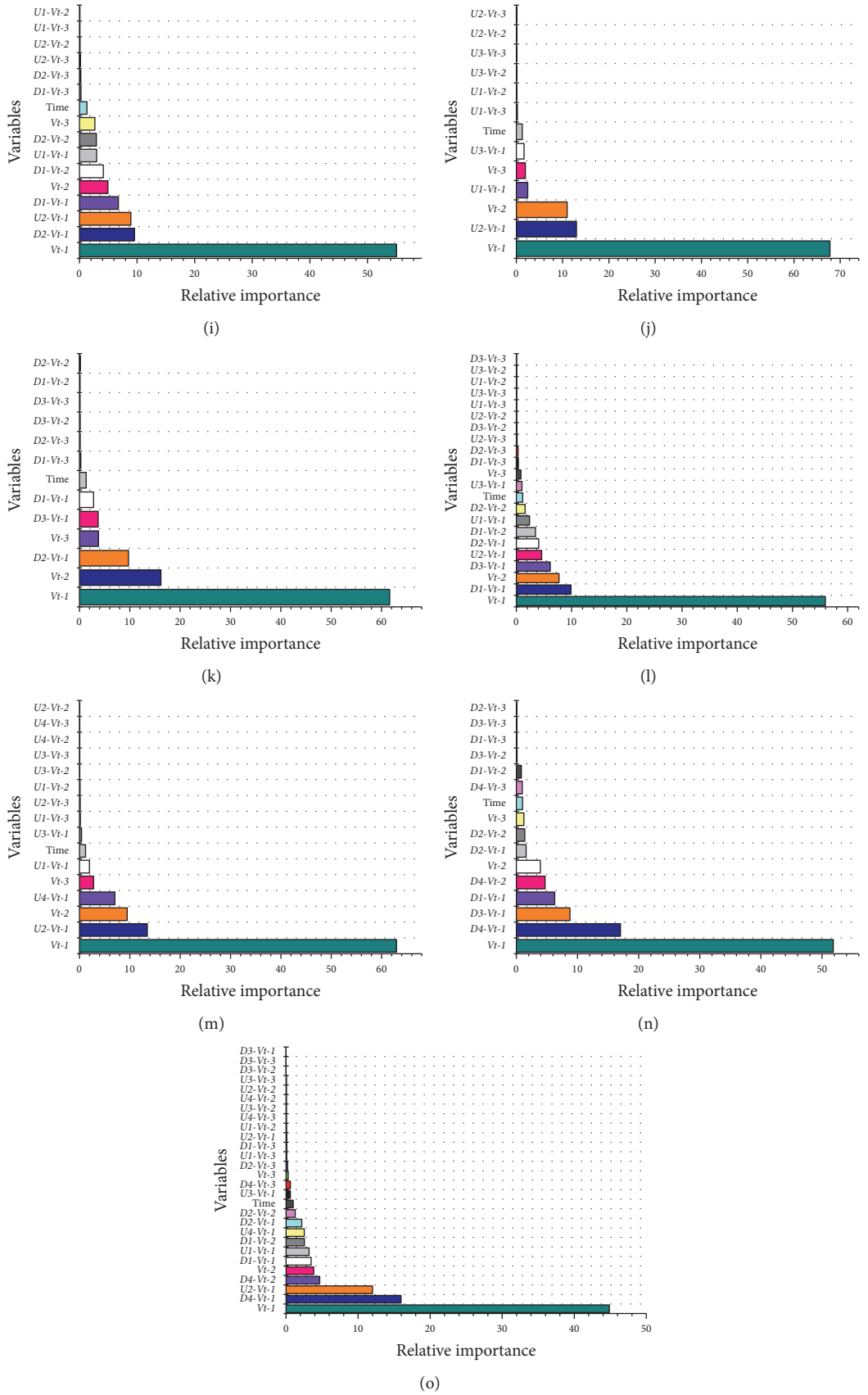

Figure 5: Relative importance of variables in GBDT models (5 min ahead) (%): (a) Model 1, (b) Model 2, (c) Model 3, (d) Model 4, (e) Model 5, (f) Model 6, (g) Model 7, (h) Model 8, (i) Model 9 (j) Model 10, (k) Model 11, (l) Model 12, (m) Model 13, (n) Model 14, and (o) Model 15.

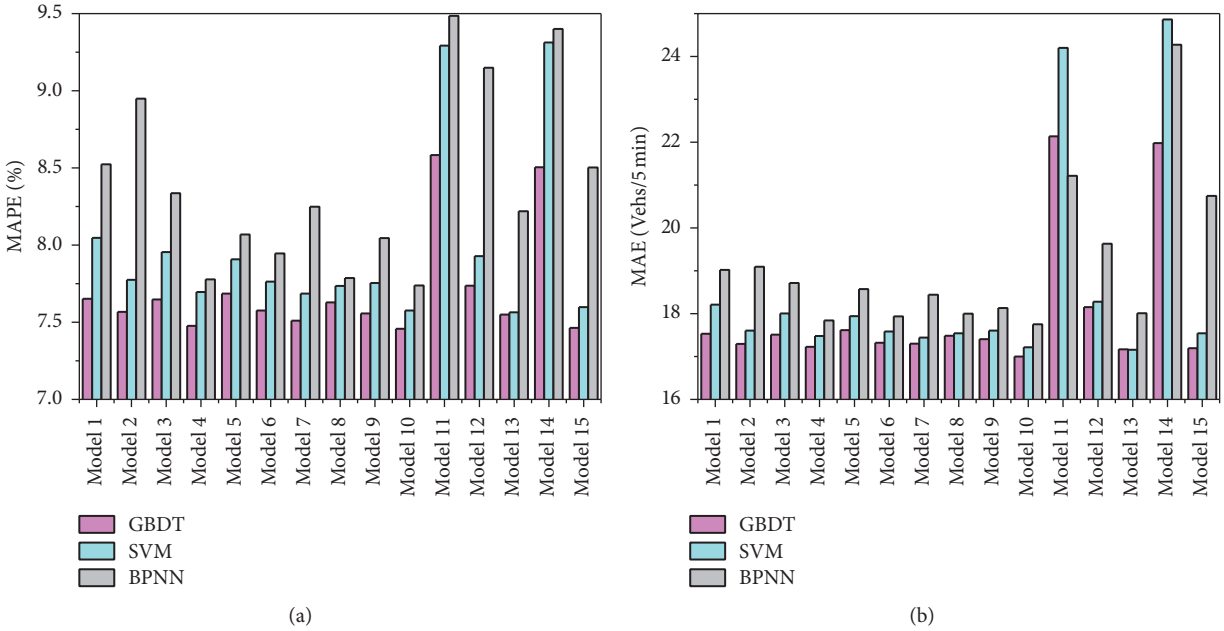

Figure 6: Prediction accuracy comparison of GBDT, SVM, and BPNN models: (a) MAPE and (b) MAE (5 min ahead).

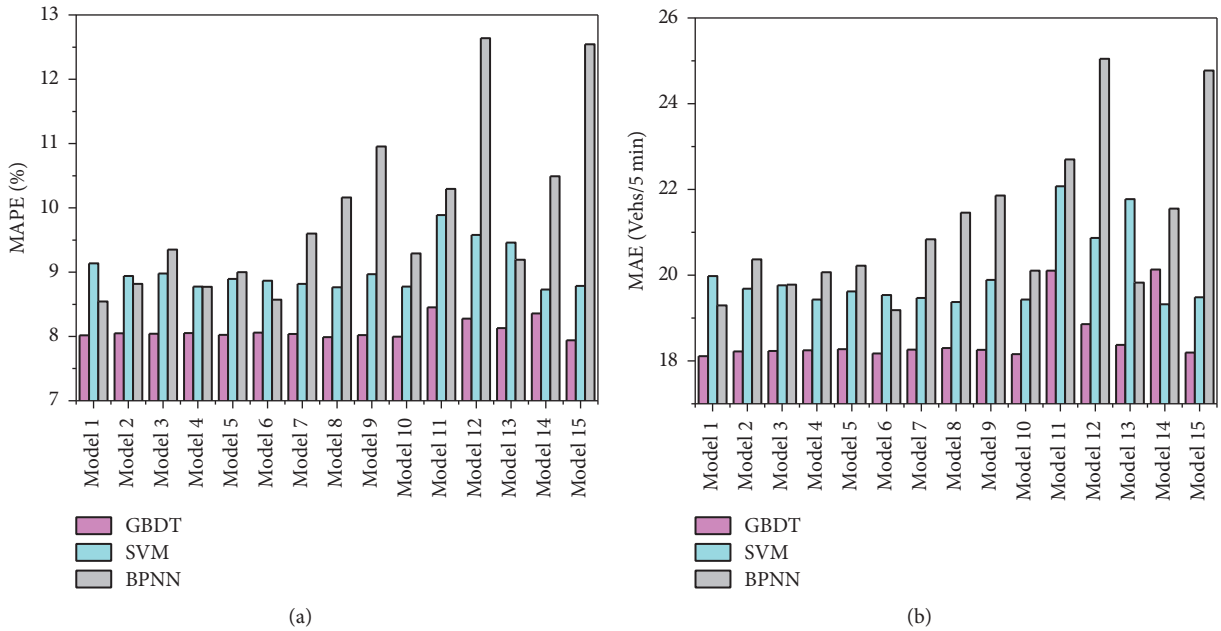

Figure 7: Prediction accuracy comparison of GBDT, SVM, and BPNN models: (a) MAPE and (b) MAE (10 min ahead).

prediction models is shown in Figures 7 and 8, respectively. It is obvious that the prediction accuracy tends to be reduced for the multi-step-ahead models in comparison with 1-step-ahead models. As a whole, the prediction errors of 5 min ahead prediction models are smaller than those of 10 min and 15 min ahead prediction models. Generally, from the perspective of prediction accuracy, GBDT models perform relatively better than SVM and BPNN models in the short-term traffic prediction for the three horizons.

The computational time for 5 min, 10 min, and 15 min ahead traffic prediction models based on GBDT, SVM, and BPNN is shown in Figure 9. GBDT algorithm costs more time than SVM for the reason that it needs to train large numbers of decision trees. As for BPNN models, the computational time varies greatly for different input variables.

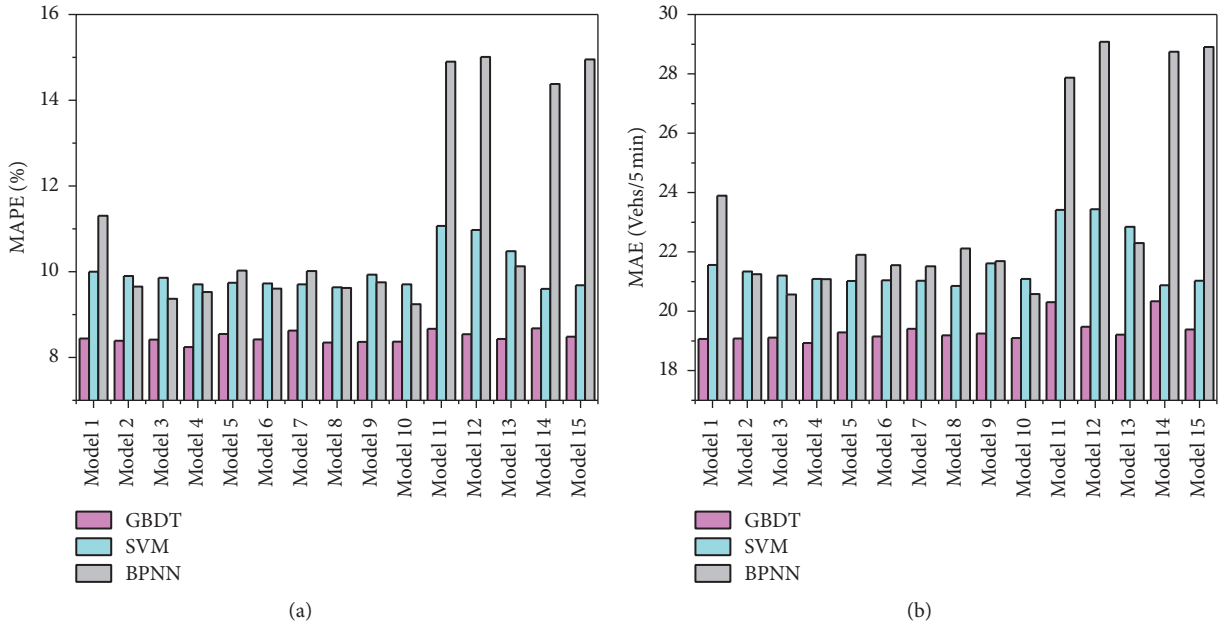

Figure 8: Prediction accuracy comparison of GBDT, SVM, and BPNN models: (a) MAPE and (b) MAE (15 min ahead).

Among all the models, the prediction errors reach the minimum at Model 10 (5 min ahead) based on GBDT, considering the historical traffic volume of P, U_1, U_2, and U_3 as the input. The traffic volume of the 9th week at Detector P is estimated based on Model 10 (5 min ahead) for GBDT, SVM, and BPNN methods respectively. The predicted traffic volume is compared with the real observations, with the total number of time steps $288 \times 7 = 2016$, as shown in Figure 10.

4.5. Traffic Prediction Models for Peak and Nonpeak Hours. To identify how the GBDT, SVM, and BPNN approaches perform under different conditions, we build the short-term traffic prediction models under the congested and smooth traffic condition by selecting the traffic volume data at peak hours (7:00–9:00, 17:00–19:00) and nonpeak hours (4:00–6:00, 21:00–23:00) as the dataset.

The prediction accuracy comparison of GBDT, SVM, and BPNN (5 min ahead) models for peak hours (7:00–9:00, 17:00–19:00) and nonpeak hours (4:00–6:00, 21:00–23:00) is shown in Figures 11 and 12. Generally, the prediction errors of GBDT models are lower than SVM and BPNN for the traffic condition at both peak hours and nonpeak hours. Moreover, MAPE of the prediction models at peak hours is lower than that of nonpeak hours for GBDT, SVM, and BPNN models, while MAE is the opposite.

Computational time comparison of GBDT, SVM, and BPNN models (5 min ahead) for peak and nonpeak hours is shown in Figure 13. As a whole, GBDT algorithm costs more time than that of SVM and less time than BPNN for peak and nonpeak hours. Generally, three prediction models for nonpeak hours cost less computational time than those for peak hours. In addition, the prediction performance is significantly improved by training the prediction models separately for different time periods of one day, such as peak or nonpeak hours, comparing with the traffic prediction models for the whole day.

5. Conclusions

This study indicates that gradient boosting machine is suitable for the short-term traffic prediction of freeway, providing a flexible framework to adopt different combinations of variables referring to the neighboring traffic information for the prediction models. The performance of GBDT is influenced by the parameter settings. Considering the computation and accuracy, the three main parameters M, J, R are optimized to produce better prediction performance with fewer iterations and avoid overfitting.

GBDT models perform better than the classical SVM and BPNN models in the short-term traffic prediction. The prediction accuracy is affected by adding the upstream or downstream traffic information to the prediction models, and the highest accuracy is produced by Model 10 for GBDT algorithm, just considering the influence of upstream traffic condition. The relative importance of variables varies considerably in the GBDT models with different variable combination. The previous traffic volume of the same site $V_{t\text{-}1}$ is the most important variable for the GBDT models, and the importance of upstream traffic condition on the traffic prediction of the current site is not equal to that of downstream traffic condition. From the temporal perspective, the importance of traffic condition at time steps t-2 and t-3 is lower than that at time step t-1.

As a whole, GBDT performs relatively better than SVM and BPNN algorithms for the 5 min, 10 min, and 15 min ahead prediction models, and the prediction errors of 5 min ahead prediction models are smaller than that of 10 min and

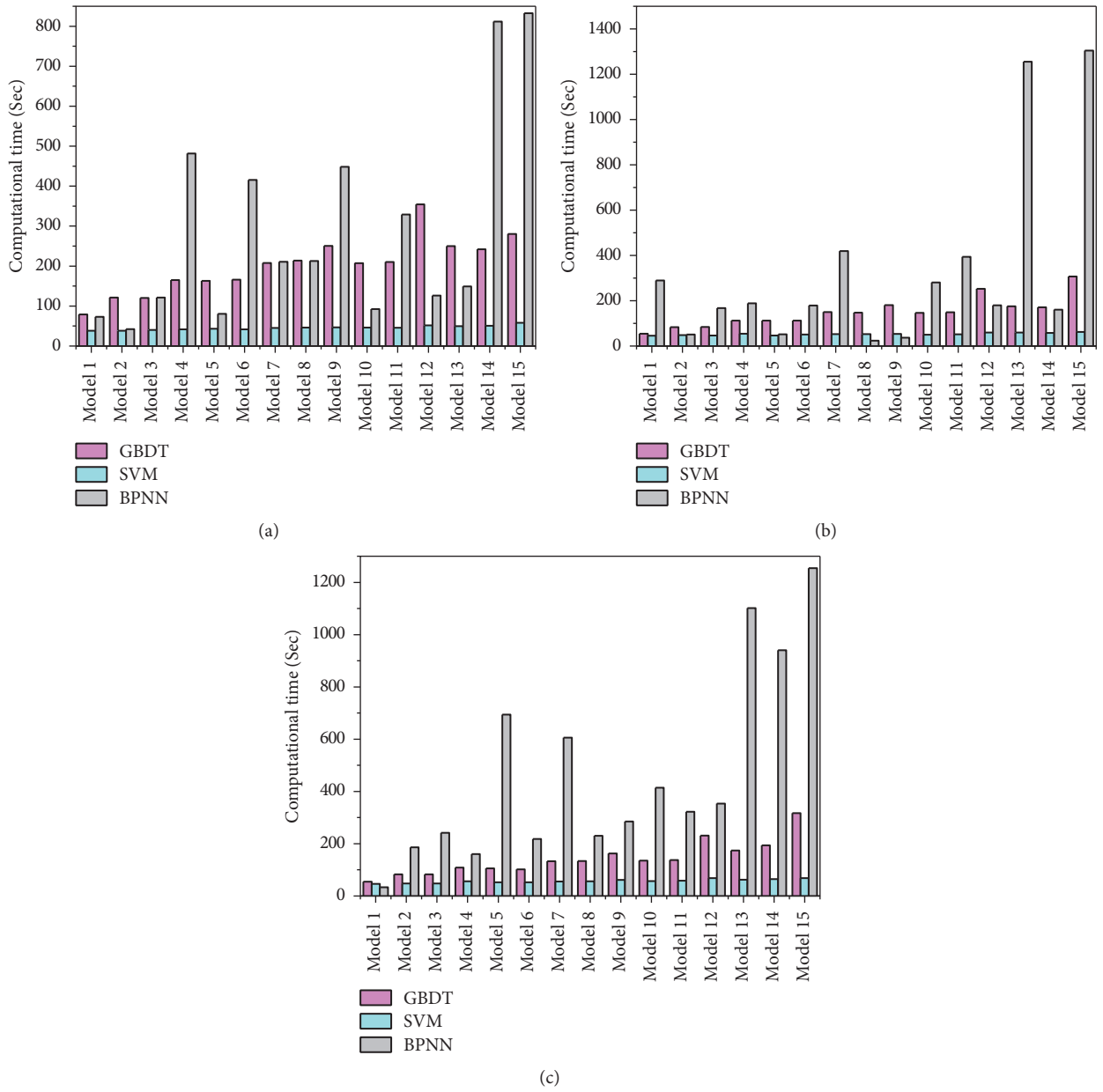

Figure 9: Computational time comparison of GBDT, SVM, and BPNN: (a) 5 min ahead; (b) 10 min ahead; (c) 15 min ahead.

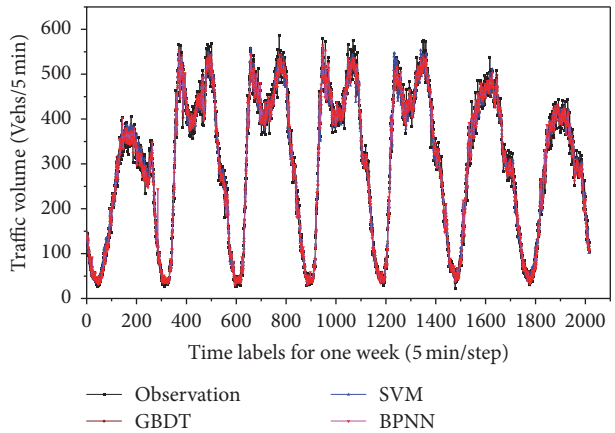

Figure 10: Comparison of the real observations and predicted values (5 min ahead).

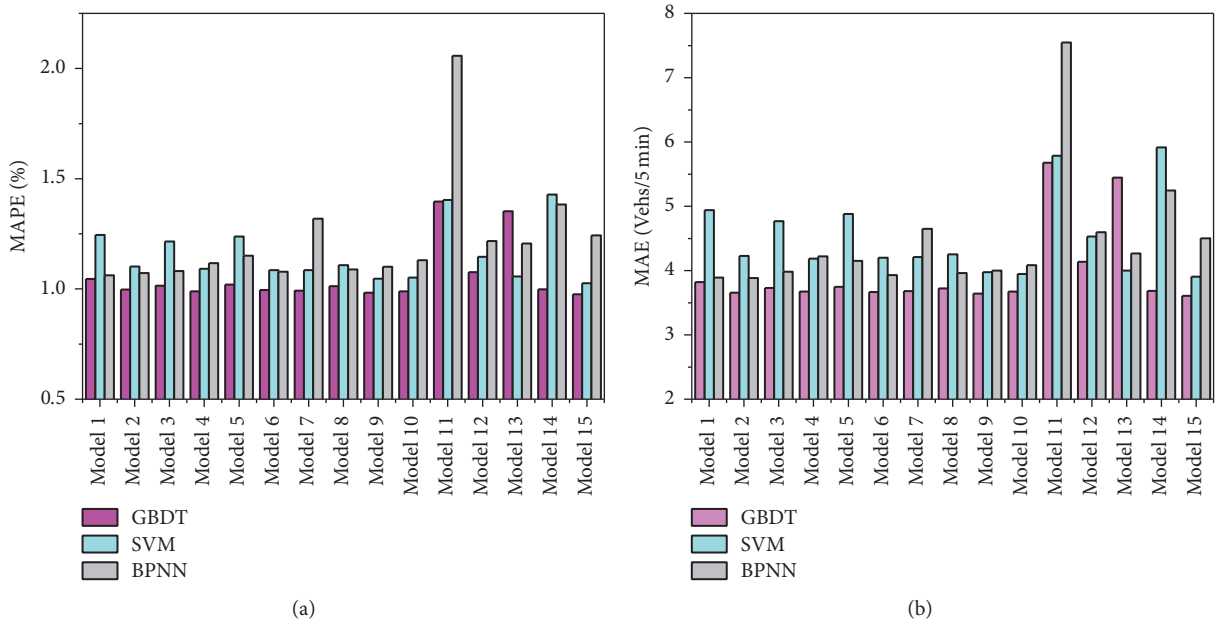

Figure 11: Prediction accuracy comparison of GBDT, SVM, and BPNN models for peak hours: (a) MAPE and (b) MAE (5 min ahead).

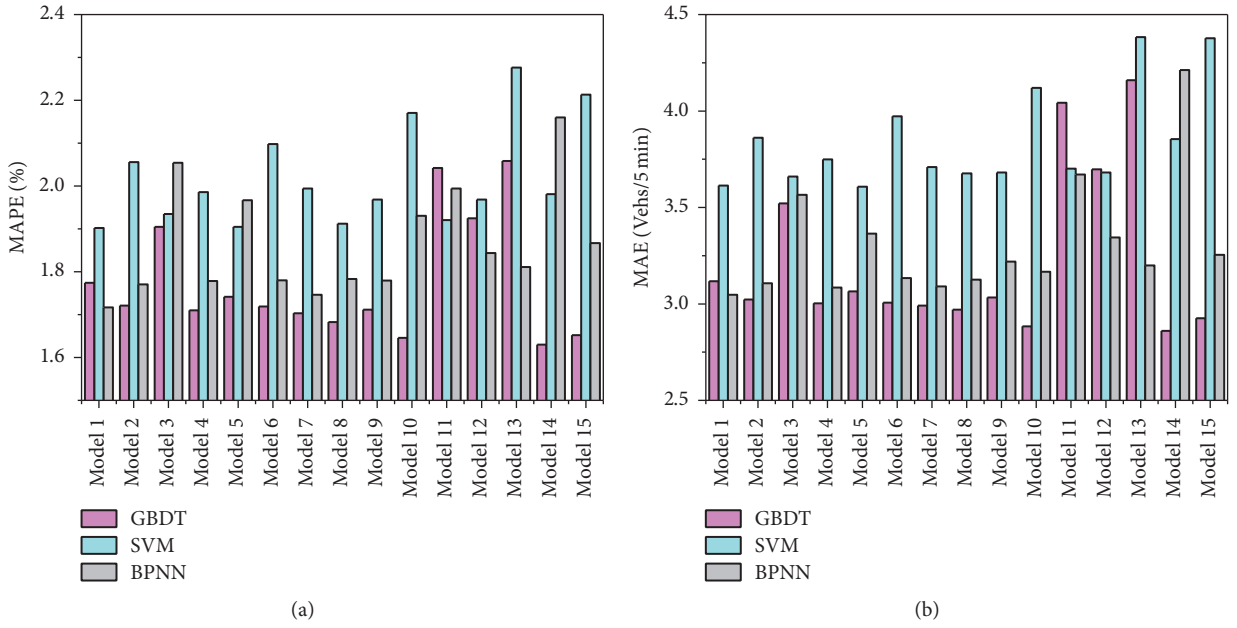

Figure 12: Prediction accuracy comparison of GBDT, SVM, and BPNN models for nonpeak hours: (a) MAPE and (b) MAE (5 min ahead).

15 min ahead prediction models. The prediction errors of GBDT models are lower than SVM and BPNN for the traffic condition at peak and nonpeak hours.

Overall, the superior prediction performance and model interpretability can be achieved by GBDT for the short-term traffic prediction, simultaneously considering the neighboring traffic condition. Short-term traffic prediction is of crucial importance for the traffic management and route guidance at the road network level. Considering the high efficiency and robustness of GBDT algorithm, more spatial and temporal traffic information could be taken into account for the accurate traffic prediction in a larger scale road network in the future work.

Conflicts of Interest

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

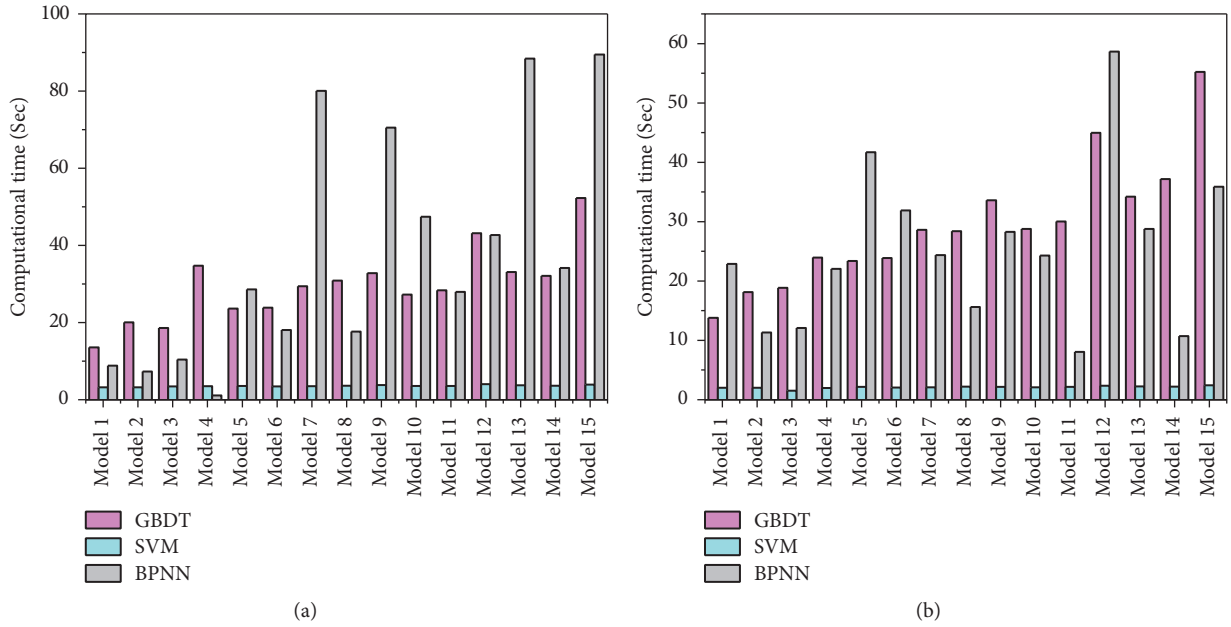

Figure 13: Computational time comparison of GBDT, SVM, and BPNN (5 min ahead): (a) for peak hours; (b) for nonpeak hours.

Acknowledgments

The research reported in this paper was supported by the project "Research on the Traffic Environment Carrying Capacity and Feedback Gating based Dynamic Traffic Control in Urban Network," which is funded by the China Postdoctoral Science Foundation (Project no. 2013M540102).

References

[1] G. Leduc, "Road traffic data: collection methods and applications," Working Papers on Energy, Transport and Climate Change 55, 2008.

[2] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[4] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: overview of objectives and methods," *Transport Reviews*, vol. 24, no. 5, pp. 533–557, 2004.

[5] H. Al-Deek, M. Wang, M. Abdel-Aty, P. Kerr, and S. Ishak, "The impact of real-time and predictive traffic information on travelers' behavior in the I-4 corridor," Tech. Rep., BC355 RPWO #3, 2003.

[6] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[7] C. De Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC '08)*, pp. 197–203, Beijing, China, October 2008.

[8] B. L. Smith, B. M. Williams, and R. Keith Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[9] Y. Xie, Y. Zhang, and Z. Ye, "Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 326–334, 2007.

[10] S. R. Chandra and H. Al-Deek, "Predictions of freeway traffic speeds and volumes using vector autoregressive models," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 13, no. 2, pp. 53–72, 2009.

[11] Y. Kamarianakis, W. Shen, and L. Wynter, "Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO," *Applied Stochastic Models in Business and Industry*, vol. 28, no. 4, pp. 297–315, 2012.

[12] J. W. C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5-6, pp. 347–369, 2005.

[13] W. Zheng, D.-H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: bayesian combined neural network approach," *Journal of Transportation Engineering*, vol. 132, no. 2, pp. 114–121, 2006.

[14] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on Chaos-Wavelet Analysis-Support Vector Machine theory," *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 219–232, 2013.

[15] Y. Zhang and Y. Xie, "Forecasting of short-term freeway volume with v-support vector machines," *Transportation Research Record*, no. 2024, pp. 92–99, 2008.

[16] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An improved K-nearest neighbor model for short-term traffic flow prediction," *Procedia—Social and Behavioral Sciences*, vol. 96, pp. 653–662, 2013.

[17] L. Breiman, "Statistical modeling: the two cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[18] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[19] Y. Kamarianakis and P. Prastacos, "Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches," *Transportation Research Record*, no. 1857, pp. 74–84, 2003.

[20] T. L. Pan, A. Sumalee, R.-X. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal-spatial correlation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1242–1254, 2013.

[21] B. M. Williams, "Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling," *Transportation Research Record*, no. 1776, pp. 194–200, 2001.

[22] J. Yang, L. D. Han, P. B. Freeze, S.-M. Chin, and H.-L. Hwang, "Short-term freeway speed profiling based on longitudinal spatiotemporal dynamics," *Transportation Research Record*, vol. 2467, pp. 62–72, 2014.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2nd edition, 2009.

[24] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[25] G. Leshem and Y. Ritov, "Traffic flow prediction using adaboost algorithm with random forests as a weak learner," *Proceedings of World Academy of Science, Engineering and Technology*, no. 19, pp. 193–198, 2007.

[26] M. Ahmed and M. Abdel-Aty, "Application of stochastic gradient boosting technique to enhance reliability of real-time risk assessment: use of automatic vehicle identification and remote traffic microwave sensor data," *Transportation Research Record*, no. 2386, pp. 26–34, 2013.

[27] Y.-S. Chung, "Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees," *Accident Analysis and Prevention*, vol. 61, pp. 107–118, 2013.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[30] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, article 21, 2013.

[31] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[32] J. H. Friedman and J. J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statistics in Medicine*, vol. 22, no. 9, pp. 1365–1381, 2003.

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, CRC Press, Boca Raton, Fla, USA, 1984.