基于梯度提升回归树的短时交通流预测模型

沈夏炯[1,2] 张俊涛[2] 韩道军[1,2]
(河南大学数据与知识工程研究所 河南 开封 475004)[1]
(河南大学计算机与信息工程学院 河南 开封 475004)[2]

摘 要 短时交通流预测是交通流建模的一个重要组成部分,在城市道路交通的管理和控制中起着重要的作用。然而,常见的时间序列模型(如ARIMA)、随机森林(RF)模型在交通流预测方面由于被构建模型产生的残差和输入变量所影响,其预测精度受到限制。针对该问题,提出了一种基于梯度提升回归树的短时交通预测模型来预测交通速度。首先,模型引入Huber损失函数作为模型残差的处理方法;其次,在输入变量中考虑预测断面受到毗邻空间因素和时间因素相关性的影响。模型在训练过程中通过不断调整弱学习器的权重来纠正模型的残差,从而提高模型预测的精度。利用某城市快速路的交通速度数据进行实验,并使用MSE和MAPE等指标将本文模型与ARIMA模型和随机森林模型进行对比,结果表明,文中所提模型的预测精度最好,从而验证了模型在短时交通流预测方面的有效性。

关键词 短时交通流预测,梯度提升回归树,损失函数,时空相关性
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.040

Short-term Traffic Flow Prediction Model Based on Gradient Boosting Regression Tree

SHEN Xia-jiong[1,2] ZHANG Jun-tao[2] HAN Dao-jun[1,2]
(Institute of Data and Knowledge Engineering, Henan University, Kaifeng, Henan 475004, China)[1]
(School of Computer and Information Engineering, Henan University, Kaifeng, Henan 475004, China)[2]

Abstract Short-term traffic flow prediction is an important part of traffic flow modeling, and it also plays an important role in urban road traffic management and control. However, the common time series model (e.g., ARIMA) and random forest model (RF) are limited in the prediction accuracy due to the residuals generated by the model and the input variables. Aiming at this problem, a short-term traffic flow prediction model based on gradient boosting regression tree (GBRT) was proposed to predict the travel speed. The model (GBRT) first introduces the Huber loss function to deal with residuals. Secondly, the spatial-temporal correlations are also considered in the input variables. The model adjusts the weight of the weak learners in the training process, and corrects the residuals of the model to improve the prediction accuracy. Experiment was conducted by using traffic speed data of a city expressway, and ARIMA model and random forest modle were compared with the proposed model by using MSE, MAPE and other indicators. Results show that the proposed model has the best prediction accuracy, and the validity of the model in short-term traffic flow prediction is verified.

Keywords Short-term traffic flow prediction, Gradient boosting regression tree, Loss function, Spatial-temporal correlations

1 引言

随着社会经济和城市化的不断发展,城市道路交通在给人们日常生活带来便利的同时,也带来了交通拥堵、交通事故等问题;尤其是人口、车辆的日益增加与道路规划不合理、基础设施不到位的矛盾,使得城市道路交通问题日益严峻。值得欣慰的是,随着信息化、计算机、自动控制和人工智能等技术的不断进步,智能交通系统成为了解决城市道路交通问题的有效途径之一。短时交通速度预测是智能交通系统最重要的组成部分[1],同时也是智能交通系统领域的研究热点;其在

到稿日期:2017-04-12 返修日期:2017-07-19 本文受国家自然科学基金资助项目(61272545,61402149),河南省科技攻关计划基金资助项目(142102210390),河南省教育厅科技攻关计划基金资助项目(14A520026),河南省博士后科研项目(2015036)资助。
沈夏炯(1963-),男,博士,教授,CCF会员,主要研究领域为空间数据处理、形式概念分析,E-mail:77230497@qq.com;**张俊涛**(1989-),男,硕士生,主要研究领域为空间数据处理、数据挖掘与分析,E-mail:zhangjuntao1990@126.com(通信作者);**韩道军**(1979-),男,博士,副教授,CCF会员,主要研究领域为空间数据处理、信息安全。

交通管理、控制和指导等方面的应用，能够有效地缓解甚至解决城市交通拥堵、交通事故等问题。

交通系统是由人、车辆和道路组成的相互交织、相互影响的复杂系统，具有高度的不确定性和非线性，尤其是短时交通流预测受随机因素的影响更大，时变性、不确定性更强[2]。交通速度的预测是短时交通流预测的重要组成部分，由于对实时性和准确性要求更高，因此面临巨大困难。随着对问题研究的深入，大量的方法被用于解决交通流预测问题，如统计和回归方法、交通流理论方法、历史平均法、卡尔曼滤波法等。而机器学习技术的发展，使得基于机器学习的有关算法和模型也被广泛应用于交通领域，包括神经网络模型、支持向量机、混合和集成技术等。在众多的模型和方法中，差分自回归移动平均（Autoregressive Integrated Moving Average，ARIMA）模型逐渐成为评价新开发预测模型的基准[3]。当交通流在时间上呈现规律变化时，此模型有很好的预测效果。文献[4]利用 ARIMA 模型对交通事故进行预测，并通过实验证明了其具有很高的预测精度，为政府和交通部门预防交通事故提供了依据。文献[1]采用支持向量机方法来预测短时交通速度，并验证了其有效性。文献[5]采用随机森林模型对短时交通流进行预测，并证明了其有效性和易用性。文献[6]在采用梯度提升回归树训练历史行程时间数据后进行预测，并与其他集成方法进行对比，证明了该方法在高速公路行程时间预测方面的适用性。文献[7]在考虑时空相关的条件下，证明了梯度提升回归树在城市路段行程时间预测方面具有更高的预测精度。文献[8]采用梯度提升回归树对短时地铁的客流量进行预测，并考虑了与地铁相邻的公交车站的客流等一系列因素的影响，以获取短时地铁客流量微妙和隐藏的变化，提高了对短时地铁客流量的预测精度。上述方法尽管在预测精度方面有所提高，但在残差处理方法和输入变量时空相关性分析方面还存在不足。

为此，本文提出一种基于梯度提升回归树的短时交通流预测模型。与文献[7-8]的不同之处在于，本文模型通过引入 Huber 损失函数作为模型残差处理方法来降低样本残差的损失，有效地提高了模型的预测精度。其次，模型考虑了输入变量中预测断面受到毗邻空间因素和时间因素相关性的影响，并就时空相关因素对模型预测性能的影响进行对比和分析。利用某城市快速路上相邻 5 个微波检测器所采集的连续 15 天的速度数据集进行实验，将其中前 14 天的数据作为训练数据集输入模型进行训练，最后一天的数据作为测试集输入模型进行验证。将模型速度的预测值与观测值之差转换成均方差误差（MSE）、对称平均绝对百分比误差（SMAPE）、均方根误差（RMSE）等指标，并将本文模型与时间序列模型（ARIMA）和随机森林（RF）模型的结果进行对比。对比结果表明，本文模型的预测精度误差最小，从而验证了本文模型在短时交通流预测方面的有效性。

2　预测算法

2.1　梯度提升算法

梯度提升的思想源于 Breiman[9]，其被理解为基于误差函数的一个优化算法。梯度提升是解决分类和回归问题的机器学习技术，通过集成弱预测模型的形式生成一个强预测模型，如决策树。梯度提升回归算法随后被 Friedman 开发出来[11-12]，其核心在于每次计算由一个基本模型完成，而下次计算是为了减小上次模型的残差，并在残差减小的梯度方向新建立一个基本模型。因此，通过不断调整和优化弱学习器的权重，使之成为强学习器，可以对损失函数进行极小化优化。

2.2　ARIMA

ARIMA 模型，即差分自回归移动平均模型，是 20 个世纪 70 年代初被提出的著名时间序列预测方法。ARIMA（p，d，q）模型中，AR 是自回归，p 为自回归项；MA 为移动平均，q 为移动平均项数，d 为时间序列平稳时所做的差分次数。所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型。ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，分为移动平均过程（MA）、自回归过程（AR）、自回归移动平均过程（ARMA）以及 ARIMA 过程。ARIMA 模型的基本思想是：将预测对象随时间推移而形成的数据序列视为一个随机序列，并用一定的数学模型来近似描述这个序列。该模型一旦被识别，便可根据时间序列的过去值和现在值预测未来值[15]。

2.3　随机森林

随机森林，是利用多棵树对样本进行训练并预测的一种分类器，于 2001 年由 Breiman 提出并扩展[14]。其结合了两大机器学习技术，即 Breiman 的“Bagging”思想和随机特征选取[6]，并成为了数据挖掘领域的重要成员。随机森林的“Bagging”思想是利用 Bootstrap 重抽样方法从原始样本集中抽取多个样本集，并分别对每个样本集进行决策树建模，预测时对每棵决策树都会给出一个预测结果。随机森林的随机特征选取的过程是指，在每个决策树模型的生成过程中，对于每个节点，通过从随机选取的特征中比较最优分割进行节点分裂。随机森林中两大技术的目的都在于保持所构建决策树样本集的多样性。文献[14]利用数学理论和大量数据测试证明了随机森林不会轻易出现过拟合现象，并且其泛化误差也小于决策树。随机森林所结合的两大技术集合了多棵决策树，具有预测准确率高、对异常值和噪声数据容忍度高等优点。

3　基于梯度提升回归树的短时交通流预测模型

为提升短时交通流预测的精度，文中提出一种基于梯度提升回归树的短时交通流预测模型。模型的每个基本模型是一棵回归树，用以纠正上一次迭代计算过程产生的残差。残差是预测值与观测值之间的差，其值越小代表模型的预测性能越好。为提高模型的性能并降低残差值，本文从损失函数和输入变量两个方面进行改进。一方面，通过引入 Huber 损失函数来降低样本的残差损失。Huber 损失函数是一种使用鲁棒性回归的损失函数[16]，其中变量 a 被表述为残差：$a=y-f(x)$，则其表达式为：

$$L_\delta(y,f(x))=\begin{cases}\frac{1}{2}(y-f(x))^2, & |y-f(x)|\leqslant\delta\\ \delta|y-f(x)|-\frac{1}{2}\delta^2, & \text{otherwise}\end{cases} \tag{1}$$

另一方面，当对交通速度进行预测时，模型的输入变量分别以不考虑空间因素和时间因素（Case1）、仅考虑时间因素（Case2）、仅考虑空间因素（Case3）、考虑空间和时间因素（Case4）4 种情况作为输入。研究了空间因素和时间因素对模型预测精度的影响，并分析了空间因素和时间因素在模型中的相关重要性。

基于梯度提升回归树的短时交通流建模流程如图 1 所示，利用训练数据建立稳定模型，并利用测试数据对模型的性能进行评价。

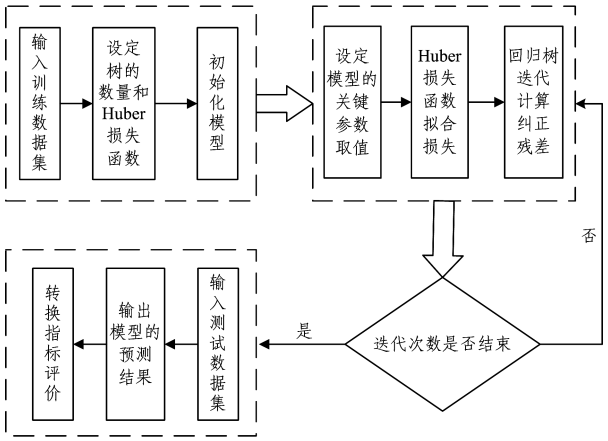


图 1 基于梯度提升回归树的短时交通流建模流程图

Fig. 1 Flowchart of short-term traffic flow modeling based on gradient boosting regression tree

已知训练数据集 $(x,y)=\{(x_i,y_i)\}_{i=1}^n$，其中 x 是输入变量，y 是相对应的输出变量。假设每棵回归树的叶子是 J_m，将输入空间划分为 J_m 个不相交区域：$R_{1m},R_{2m},\cdots,R_{jm}$，并在每个区域上确定输出的常量值[6,10]，如 b_{jm} 为区域 R_{jm} 的常量值。其回归树表达式为：

$$h_m(x)=\sum_{j=1}^{J_m}b_{jm}I(x\in R_{jm}) \tag{2}$$

$$\text{Where } I(x\in R_{jm})=\begin{cases}1, & \text{if } x\in R_{jm}\\ 0, & \text{otherwise}\end{cases}$$

模型的目标是向着函数 $f(x)$ 寻找一个近似 $\hat{F}(x)$，并使给定的损失函数 $L(y,f(x))$ 在数据集上的期望值最小化。

$$\hat{F}(x)=\arg\min_{f(x)}\sum_{i=1}^{N}L(y,f(x)) \tag{3}$$

因此，对模型初始化一个常量函数 f_0，引入 Huber 损失函数进行拟合，且模型树的数量为 M。初始化模型如下：

$$f_0=\arg\min_{f(x)}\sum_{i=1}^{M}L_\delta(y,f(x)) \tag{4}$$

模型沿梯度下降的方向进行迭代，目的是降低残差值。对于模型的第 m 次迭代，在模型中计算损失函数负梯度的当前值，并将它作为残差的估计值，则负梯度被定义为：

$$\gamma_{mi}=-\left[\frac{\partial L_\delta(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} \tag{5}$$

对于残差，梯度提升回归树模型将为其拟合一个回归树 $h_m(x)$，则模型梯度下降法的步长计算如下：

$$\rho_m=\arg\min_{\rho}\sum_{i=1}^{n}L_\delta(y_i,f_{m-1}(x_i)+\rho_m h_m(x_i)) \tag{6}$$

为了纠正上一个回归树的残差，模型在每一步增加一棵新的回归树，因此模型最终的更新如下：

$$f_m=f_{m-1}(x)+\rho_m h_m(x) \tag{7}$$

梯度提升回归树沿梯度下降的方向构建一个新的模型，并通过最小化损失函数的期望值来不断更新模型[7]，从而使模型最终趋于稳定，最后用测试数据预测未来值来进行验证。本文模型在训练过程中不仅引入 Huber 损失函数作为拟合残差的方法，还考虑模型的预测断面受空间因素和时间因素相关性的影响，同时通过不断调整基本模型的权重来提高预测精度。模型有 3 个关键的参数：树的数量（M）、学习效率（L）、树的深度（D）。一般来说，随着树的增加，模型能够减小训练误差，然而如果树的数量太多，将会导致其泛化能力大大减弱[6,10]，模型也会因为出现过拟合现象而预测性能降低。因此，通过优化模型树的数量将预测误差降到最低很有必要，并且过拟合现象也可以通过控制树的数量来避免。另外，梯度提升算法引入了收缩系数（Shrinkage），即每次以一小步长逐渐逼近最佳结果，以避免过快逼近结果而造成过拟合。因此，模型方程式被修改为：

$$f_m=f_{m-1}(x)+L*\rho_m h_m(x),0<L\leqslant 1 \tag{8}$$

参数 L 被称作“学习效率”，即每个弱分类器的权重缩减系数。采用较小的学习效率（$L<0.1$）的梯度提升算法能更有效地提高模型的泛化能力[10,13]，但是很小的学习效率（$L<0.001$）需要较多的树才能使模型收敛，而且会降低预测的精度。而树的深度，即基本模型树的复杂度，也是影响预测精度的一个因素。本文模型通过不断优化树的数量（M）、学习效率（L）和树深度（D）的组合来提高模型的预测精度，这 3 个关键参数的选优过程将在后文中详细描述。

4 实验分析与结果

本实验采用微波检测器采集得到的数据集对模型进行训练，不断地调整模型的参数组合以寻找出最优组合，并使用测试数据集对模型进行验证。将预测值与观测值之间的差转换成 MAPE 等指标，对模型以及模型属性进行分析，并将本文所提模型与 ARIMA 模型、随机森林模型的预测精度进行对比，以验证其在短时交通流预测方面的有效性。

4.1 实验数据集

本文利用安装在固定位置的微波检测器来采集城市道路上的车辆行驶速度、车辆流量、车辆占有率等信息作为实验数据集。在某城市一条长度约为 10 kM 的道路上，由南向北的 5 个固定位置分别安装远程微波检测器，将这 5 个固定位置分别标记为 h,j,f,w,x。其中，h 和 j 之间的距离为 0.9 kM，j 和 f 之间的距离为 1.5 kM，f 和 w 之间的距离为 1.9 kM，w 和 x 之间的距离为 2.3 kM。微波检测器每间隔 5 min 采集一次数据，连续获取 2016 年 6 月 21 日到 2016 年 7 月 5 日共 15 天的交通数据信息。本文考虑预测断面受到上下游断面

空间因素以及历史时间因素的影响，共选取 20 个数据属性，其中前 19 个数据属性作为模型的输入变量，最后 1 个数据属性作为模型的期望输出。其数据集的格式如表 1 所列。

表 1　数据集格式

Table 1　Format of data sets

h_{t-1}	h_{t-2}	h_{t-3}	j_{t-1}	j_{t-2}	j_{t-3}	f_{t-1}	f_{t-2}	f_{t-3}	w_{t-1}
56.65	53.67	55.74	67.96	64.14	64.54	10.62	10.67	12.81	70.66
56.63	56.65	53.67	62.32	67.96	64.14	13.05	10.62	10.67	69.55
54.61	56.63	56.65	64.42	62.32	67.96	13.03	13.05	10.62	67.01
55.46	54.61	56.63	67.73	64.42	62.32	12.13	13.03	13.05	66.29
…	…	…	…	…	…	…	…	…	…

w_{t-2}	w_{t-3}	x_{t-1}	x_{t-2}	x_{t-3}	$\triangle f_{t-1}$	$\triangle f_{t-2}$	day of week	time of day	f_t
69.92	69.89	60.55	58.93	67.29	−0.04	−2.15	5	1	13.05
70.66	69.92	72.51	60.55	58.93	2.43	−0.04	5	2	13.03
69.55	70.66	64.00	72.51	60.55	−0.02	2.43	5	3	12.13
67.01	69.55	72.91	64.00	72.51	−0.91	−0.02	5	4	25.42
…	….	…	…	…	…	…	…	…	…

表 1 共获取了 4320 个时段的速度数据，且每个时段的数据有 20 个属性值。5 个断面在 2016 年 6 月 21 日的速度波动情况如图 2 所示。数据集中的 h_{t-1}，h_{t-2}，h_{t-3} 分别是当前时刻 t 的前 5 min，10 min，15 min 在断面 h 处所采集到的车辆速度；类似可知在断面 j，f，w，x 处所采集的前 5 min，10 min 和 15 min 的车辆速度。f_t 是当前时刻 t 在断面 f 处的车辆速度，$\Delta f_{t-1}=f_{t-1}-f_{t-2}$ 和 $\Delta f_{t-2}=f_{t-2}-f_{t-3}$ 是断面 f 在当前时刻 t 之前的速度变化。day of week 为每周周一到周日的索引；而 time of day 为 1 到 288 的索引，表示每天以 5 min 的频率获取一个时段的速度数据。实验以数据集中前 14 天的数据作为训练数据，而将最后 1 天的数据作为测试数据用于验证模型的预测精度。模型在当前时刻 t 对断面 f 处的车辆速度进行预测并将其作为模型的输出变量 y，而将采集的历史速度数据作为模型的输入变量 x，输入变量考虑预测断面 f 受到上游断面 h 和 j 处以及下游断面 w 和 x 处的空间因素以及时间因素的影响。对模型的 3 个关键参数即树的数量(M)、学习效率(L)和树的深度(D)进行调优组合，以寻求模型的最优预测精度。

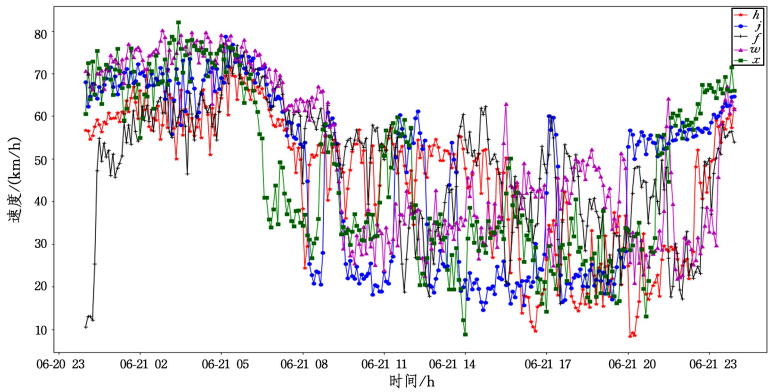


图 2　2016 年 6 月 21 日的速度变化分布

Fig. 2　Distribution of velocity on July 21st, 2016

4.2　参数调优

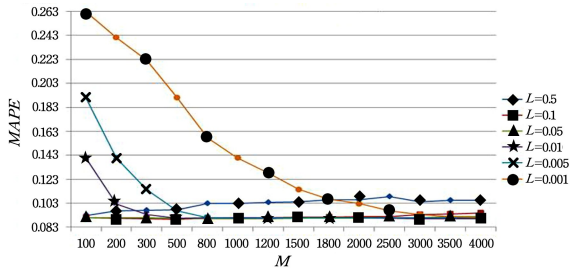
短时交通流预测是一个未来事件，对智能交通系统的交通控制和交通诱导具有非常重要的意义。优化梯度提升回归树参数，找到不同参数组合对模型性能的影响，至关重要[6-7]。本节将 14 天的训练数据集信息输入到模型中，通过不断调整参数组合，用 MAPE 指标表示测试数据集中观测值与预测值的差，当寻找到的 $MAPE$ 值最小时，即找到模型的最优参数组合。基于梯度提升回归树的短时交通流预测模型采用 Huber 损失函数，对模型的 3 个关键参数给出多个适当的取值：树的数量 M 在 100～4000 之间取值；学习效率 L 取值为 0.5，0.1，0.05，0.01，0.005，0.001；树的深度 D 取值为 2，3，4，5。为了研究参数对模型预测误差的影响，引入平均绝对百分比误差(MAPE)作为指标，其表达式如式(9)所示：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{r_i-p_i}{r_i}\right| \tag{9}$$

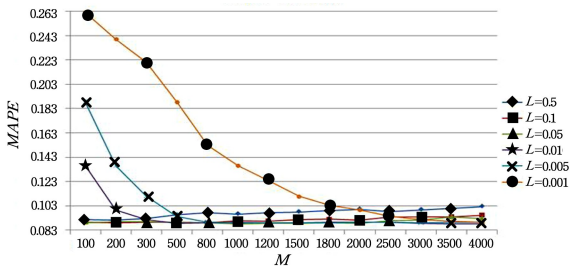
其中，n 表示测试数据的样本量，r_i 表示测试数据的观测值，p_i 表示测试数据的预测值。

为了研究树的数量对模型预测精度的影响，首先确定学习效率值和树的深度值，通过不断改变树的数量进行实验研究。树的数量 M 和 $MAPE$ 值的关系如图 3 所示。树的数量表示模型中基学习器的数量，随着树数量的变化可以获取任意 $MAPE$ 值的输出。当树的数量增加时，$MAPE$ 值会减小。然而，当树的数量继续增加(如 M 在 1500 以后增加)甚至逼近于数据集时，$MAPE$ 值变化微弱甚至会增大。导致这种现象的原因是模型出现了过拟合，从而造成预测精度下降，而基本模型的增加也造成了计算时间的浪费；并且学习效率和树

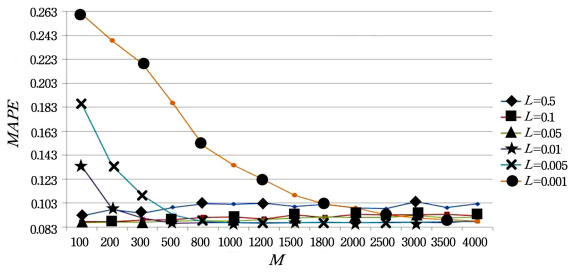
深度的变化，也使树的数量对 *MAPE* 值的影响不同。



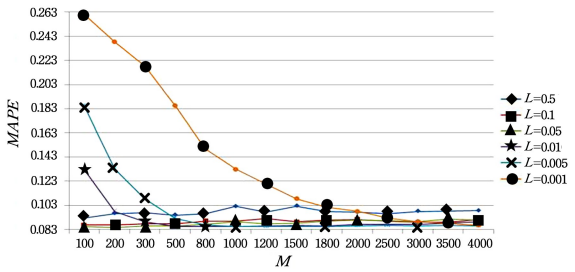
(a)Depth of tree(D=2)



(b)Depth of tree(D=3)



(c)Depth of tree(D=4)

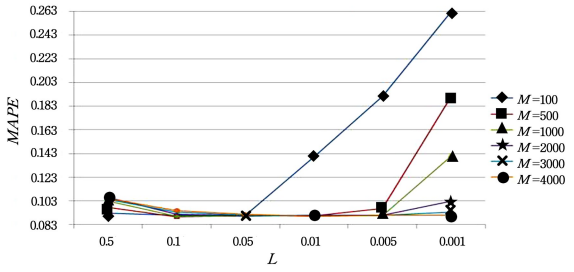


(d)Depth of tree(D=5)

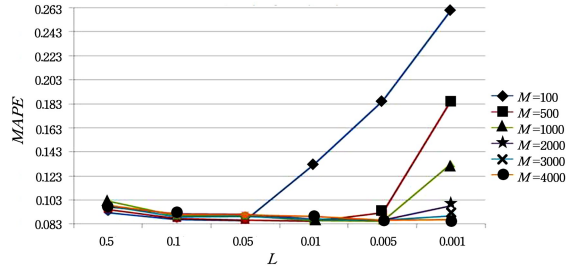
图 3 树的数量 M 和 *MAPE* 值的关系

Fig. 3 Relationship between number of trees and *MAPE*

为研究学习效率对模型预测精度的影响，首先确定模型树的数量和树的深度，得到学习效率 L 与 *MAPE* 值的关系，如图 4 所示。学习效率是模型中回归树的权重，控制着模型逼近最优结果的步长，同时也抑制着树的过度增加，以避免出现过拟合现象，从而提高模型的泛化能力。由图 4 可知，当学习效率值较小(如 $0.005<L<0.1$)时，模型的学习相对比较充分，*MAPE* 值会随着学习效率的降低而变小，因此模型有很好的预测精度。然而，当学习效率继续变小(如 $L\leqslant 0.001$)时，*MAPE* 值减小微弱甚至变大，在模型的预测精度下降的同时也要求更多的基本模型来保证模型收敛，并导致了计算时间的增加。因此，学习效率较小(如 $L\leqslant 0.001$)时，会造成过拟合，不仅会降低模型的预测性能，还将浪费计算时间。因此，选取学习效率值时，需要平衡预测精度与计算时间代价。



(a)Depth of tree(D=2)

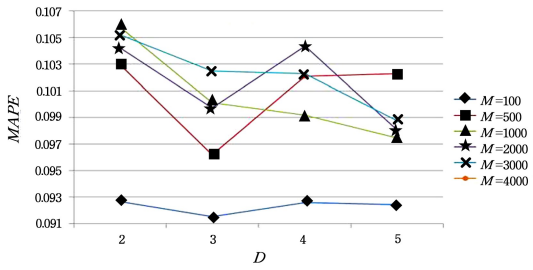


(b)Depth of tree(D=5)

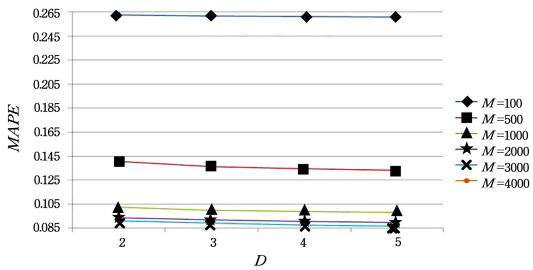
图 4 学习效率 L 和 *MAPE* 值的关系

Fig. 4 Relationship between learning rate L and *MAPE*

树的深度，即模型中回归树的复杂度，也是影响预测性能的一个因素。树的深度 D 与 *MAPE* 值的关系如图 5 所示。当学习效率取值为 0.5 且模型树的数量确定时，随着树深度的增加，*MAPE* 值呈现先减小后增大的趋势，同时 *MAPE* 值也随着树的数量的增加而上升。当学习效率值取为0.005时，随着深度的增加，*MAPE* 值减小得很缓慢；但当树的数量增加时，*MAPE* 值减小得非常明显。因此，在较小的学习效率下，树的深度不仅可以提高模型的预测精度，而且能够减少由于树数量的增加而造成的过拟合问题。



(a)Learning rate(L=0.5)



(b)Learning rate(L=0.005)

图 5 树的深度 D 与 *MAPE* 值的关系

Fig. 5 Relationship between depth of trees and *MAPE*

由实验结果可知，模型的这 3 个关键参数控制着模型预测精度的变化。选取适当小的学习效率和相应较大的树数量是很有必要的，较小的学习效率虽然能够每次收缩一个小步长，不断地逼近最优结果，但也要权衡迭代计算的时间代价。

树的深度是每棵基本模型树的复杂度，它影响着对学习效率和树的数量的选择。对于给定的数据集，树的深度越大，模型越复杂；使模型学习效率变小，会产生更好的预测精度；改变树的数量的选取方式，会增加模型计算的时间。因此，为提高模型的预测精度，不仅要进行调优选择，还需要考虑模型由于计算而付出的时间代价。

4.3　模型分析与模型对比

使用本文模型对速度进行预测，对于模型的输入变量，分别以不考虑空间因素和时间因素（Case1）、仅考虑时间因素（Case2）、仅考虑空间因素（Case3）、考虑空间因素和时间因素（Case4）4 种情况作为输入。模型对 4 种变量的输入情况进行参数调优以寻找预测精度最好的参数组合，并比较 *MAPE* 值。由图 6 可知，在模型的输入变量中，同时考虑预测断面受到毗邻空间因素和时间因素的影响比不考虑空间因素和时间因素、仅考虑时间因素和仅考虑空间因素的预测精度都好，其 *MAPE* 分别降低了 6.4%，2.2%和 3.6%。

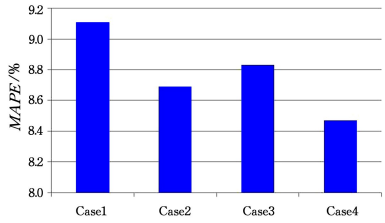


图 6　模型输入变量对预测精度的影响

Fig. 6　Impact of input variables on prediction accuracy

图 7 显示了考虑空间因素和时间因素的情况下，模型中每个输入变量的相对重要性，输入变量对预测值的影响是多样性的。正如我们所预期的，f_{t-1} 作为与当前时刻最接近的速度值，其所占的比重也最大。Δf_{t-1} 和 Δf_{t-2} 是断面 f 前 5 min 和 10 min 的速度变化，它们对断面 f 当前时刻的速度预测也有很大的影响。time of day 和 day of week 所占的比重反映了速度的周期性和规律性。图 6 中的 *MAPE* 值也表明了时间因素对预测值的影响。观察上游断面 h 和 j 以及下游断面 w 和 x 的相对重要性可知，断面 f 受到与之紧紧相邻的断面 j 和断面 w 的影响比其他两个断面的大，这反映了预测断面受到的空间因素影响与断面之间所处的不同位置有关。

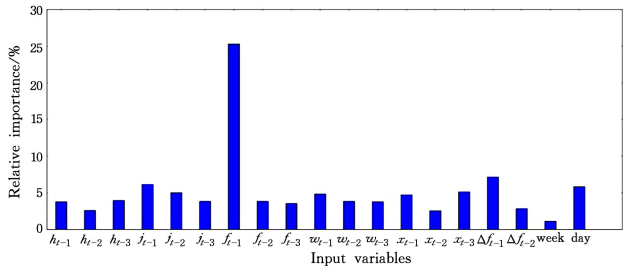


图 7　在 Case4 的情况下输入变量在模型中的相对重要性

Fig. 7　Relative importance of input variables in Case 4

为验证本文模型在预测方面的有效性以及较其他模型更具优势，将其与 ARIMA 模型和随机森林模型进行对比。采用均方差误差（MSE）、对称平均绝对百分比误差（SMAPE）、均方根误差（RMSE）作为对比 3 个模型预测精度误差的指标。

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(p_i-r_i)^2 \tag{10}$$

$$SMAPE=\frac{100\%}{n}\sum_{i=1}^{n}\frac{|p_i-r_i|}{(|r_i|+|p_i|)/2} \tag{11}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(p_i-r_i)^2}{n}} \tag{12}$$

对比结果如表 2 所列。可以看出，本文模型在预测误差方面优于 ARIMA 模型和随机森林模型。

表 2　3 个模型的预测误差的对比

Table 2　Comparison of prediction error among three models

模型 \ 指标	*MSE*	*MAPE*	*SMAPE*/%	*RMSE*
ARIMA	29.1507	0.0922	8.99	5.3991
RF	26.0140	0.0892	8.75	5.1004
GBRT	22.7414	0.0847	8.21	4.7688

结束语　本文利用微波检测器所采集的某城市道路的实时交通速度数据，提出一种基于梯度提升回归树的短时交通流预测模型。所提模型引入了 Huber 损失函数，并考虑了输入变量预测断面受到毗邻空间因素和时间因素相关性的影响，通过在训练过程中不断对弱学习器的权重进行优化，使其成为强学习器，从而提高模型预测的精度。通过对模型的 3 个主要参数（即树的数量、学习效率和树的深度）的不断优化组合，寻找最佳参数组合来提高模型的预测性能。本文将前 14 天的数据作为训练集，最后 1 天的数据作为测试集，对模型输入变量的空间因素和时间因素的相关性进行分析，并将所提模型与 ARIMA 模型和随机森林模型的预测性能进行对比。本文模型较其他模型果有明显的效果提升，从而验证了本文模型在短时交通流预测方面的有效性。今后将对异常事件下交通流的预测进行进一研究。

参 考 文 献

[1] YAO B, CHEN C, CAO Q, et al. Short-Term Traffic Speed Prediction for an Urban Corridor[J]. Computer-Aided Civil and Infrastructure Engineering, 2017, 32(2): 154-169.

[2] WANG J, SHI Q X. Summary of short-term traffic flow prediction model[J]. Its Communication, 2005, 1(1): 10-13. (in Chinese)
王进，史其信. 短时交通流预测模型综述[J]. Its 通讯，2005，1(1)：10-13.

[3] ZHANG Y, ZHANG Y, HAGHANI A. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatilitymodel[J]. Transportation Research Part C: Emerging Technologies, 2014, 43(1): 65-78.

[4] ZHANG J, LIU X M, HE Y L, et al. Application of ARIMA Model in Forecasting Traffic Accidents[J]. Journal of Beijing University of Technology, 2007, 33(12): 1295-1299. (in Chinese)
张杰，刘小明，贺玉龙，等. ARIMA 模型在交通事故预测中的应用[J]. 北京工业大学学报，2007，33(12)：1295-1299.

[5] CHENG Z, CHEN X F. The model of short term traffic flow prediction based on the random forest[J]. Microcomputer & Its Applications, 2016, 35(10): 46-49. (in Chinese)
程政，陈贤富. 基于随机森林模型的短时交通流预测方法[J]. 微型机与应用，2016，35(10)：46-49.

（下转第 264 页）