文章编号: 1001-0920(2018)11-2080-07 DOI: 10.13195/j.kzyjc.2017.0729

高速公路行程时间Bootstrap-KNN区间预测分析与实证

陈娇娜[1†], 张 翔[2], 张生瑞[3]

(1. 西安石油大学 电子工程学院，西安 710065；2. 中交第一公路勘察设计研究院有限公司，西安 710075；3. 长安大学 公路学院，西安 710064)

摘 要: 针对行程时间点预测不能描述预测结果的可信度问题,以高速公路收费系统作为基础数据源,提出基于Bootstrap的高速公路行程时间区间预测模型,通过范围概率(PICP)、预测区间平均宽度(MPIW)以及综合指标(CWC)反映区间预测性能. 对预测模型建模和Bootstrap置信区间估计方法两个关键步骤进行分析和实证,比较小波神经网络和K最近邻两种常用数据驱动方法的预测误差,并分析4种Bootstrap置信区间估计方法的区间预测性能. 在相同的置信水平下,Percentile Bootstrap-KNN模型的综合指标值CWC最小,说明该模型区间预测性能最佳. 对陕西省高速公路某热点OD进行实例分析,结果表明,采用相同预测算法的区间预测比点预测的误差小,且预测区间宽度可以表征预测结果的可信度和参考价值.

关键词: 交通工程；行程时间；Bootstrap；置信区间；K最近邻；区间预测

中图分类号: TP491 **文献标志码:** A

Analysis and empirical study on highway travel time interval prediction based on Bootstrap-KNN

CHEN Jiao-na[1†], *ZHANG Xiang*[2], *ZHANG Sheng-rui*[3]

(1. School of Electronic Engineering, Xi'an Shiyou University，Xi'an 710065，China；2. CCCC First Highway Consultants Co Ltd，Xi'an 710075，China；3. School of Highway, Chang'an University，Xi'an 710064，China)

Abstract: With the data source from highway charge system, the prediction model is established based on Bootstrap to improve the reliability of point prediction in travel time. Three indexes are used to evaluate the interval prediction performance, including prediction interval coverage probability(PICP), mean prediction interval width(MPIW), and coverage width-based criterion(CWC). Two key steps are analyzed and verified with actual data in modeling. As the methods used frequently in data-driven, the wavelet neural network and K nearest neighbor are compared about prediction error. The confidence interval prediction performance is analyzed among four kinds of Bootstrap methods. Under the same confidence level, the result shows that Percentile Bootstrap-KNN is the best with the minimum CWC. The proposed model is validated by Shanxi expressway in the case study. It is proved that interval prediction is better than point prediction under the same algorithm, as the reliability and value can be reflected by the prediction interval width.

Keywords: traffic engineering；travel time；Bootstrap；confidence interval；K nearest neighbor；interval prediction

0 引 言

行程时间预测是先进的高速公路出行信息系统中不可缺少的部分,行程时间的变化具有非线性和非平稳的特点[1],可靠的预测结果能够帮助出行者决策,而缺乏有效性保障的预测结果将不会被参考或关注. 因此,行程时间预测结果的可靠性量化具有重要的实用价值.

行程时间预测研究的主要方法包括时间序列方法[2]、卡尔曼滤波[3-4]、神经网络[5]、线性回归[6]和支持向量机[7-8]等,车辆检测器[9-10]、收费系统[11-12]、浮动车[13]和蓝牙[14]等多种数据源也在该领域得以应用. Zhang等[15]和毕松等[16]综述了现行的行程时间预测方法,现有的研究主要通过模型组合[17-19]和数据融合[20-22]两个角度进行模型精度的提高. 回顾文献,道路行程时间预测研究大多集中在点预测模型的改进,即只对下一时刻的行程时间进行预测,未提

收稿日期: 2017-06-10；**修回日期:** 2018-02-05.

基金项目: 陕西省交通运输厅科研项目(14-40X).

责任编委: 赵珺.

作者简介: 陈娇娜(1989−), 女, 讲师, 博士, 从事数据挖掘和智能交通的研究；张生瑞(1963−), 男, 教授, 博士生导师,从事综合交通运输等研究.

†通讯作者. E-mail: chenjn@xsyu.edu.cn

供置信水平或可信度等辅助决策信息. 在行程时间预测模型的典型建模过程中,精确掌握所有引起行程时间变化的信息是无法实现的,信息采集的准确性和有限性导致行程时间预测的不确定性. 改进的行程时间点预测模型也无法避免预测结果缺乏可信度保障的问题. 动态交通控制和出行者行为决策都需要进行高速公路行程时间预测,在关注估计值时希望掌握估计的准确程度,量化描述预测值的不确定性具有重要的现实意义.

通过区间预测PI(Interval prediction)来量化描述高速公路行程时间预测值的质量和稳定性,可以使预测结果被出行者有选择性地参考. 文献[23]基于交通流数据采用ARIMA-GARCH模型反映城市主干道行程时间均值的波动性,但是模型适应性有限. 高速公路行程时间样本数据集具有一定的不均衡性和不确定性,Bootstrap不需要对总体分布作任何的假定和限制[24], 文献[25]采用Bootstrap对动态交通网络进行可靠性评估, 文献[26-28]也证实了Bootstrap在其他领域区间预测的应用.

利用高速公路收费系统记录数据可以较为准确地计算进出站之间的实际行程时间,本文以高速公路收费系统作为数据来源,提出一种基于Bootstrap策略的高速公路行程时间区间预测方法,通过预测区间的宽度反映预测结果的可信度,即预测区间的范围越窄表明预测结果越可靠. 最后,采用陕西省2015年收费系统历史数据进行分析和验证.

1　高速公路行程时间区间预测

高速公路行程时间区间预测利用原始样本X建立Bootstrap样本,采用基于数据驱动的点预测模型计算每个Bootstrap样本的行程时间预测值,通过该预测值序列来构造一个可能包含真实值的估计范围.

1.1　区间预测模型构建

假设高速公路行程时间总体的分布未知,但已知有一个样本容量为n的来自总体的数据样本$X=\{x_1,x_2,\cdots,x_n\},n\in N^+$. 具体建模步骤如下.

Step 1: 构建原始样本数据集X. 以高速公路收费系统为数据来源,计算2015年每条记录的起讫点行程时间,获得原始样本$X=\{x_1,x_2,\cdots,x_n\},n\in N^+$.

Step 2: Bootstrap重采样. 从原始样本X中有放回地抽取数量为m的样本$X^*=\{x_1^*,x_2^*,\cdots,x_m^*\}$, $m\in N^+$,通常取$m=n$,该样本称为Bootstrap样本. 建立B个行程时间Bootstrap样本$X_1^*,X_2^*,\cdots,X_B^*$.

Step 3: 构建点预测模型. 对每个Bootstrap样本X^*进行行程时间预测建模,获得$\hat{t_i}=f_i(x)+\varepsilon,i=1,2,\cdots,B$.

Step 4: 预测值序列计算. 通过上一个步骤中B个点预测模型分别对测试样本进行预测,获得包含B个预测值的序列$\hat{\theta}^*=\{\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_B^*\}$.

Step 5: Bootstrap区间预测. 计算$\{\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_B^*\}$的Bootstrap置信区间估计,即得到行程时间预测区间$[\mathrm{low}(\hat{t}),\mathrm{up}(\hat{t})]$.

由建模步骤可知,Step 3的点预测模型和Step 5的Bootstrap置信区间估计方法将直接影响行程时间区间预测的性能和质量,因此需要对这两个关键步骤进行深入讨论和分析.

1.2　Bootstrap置信区间的常用估计方法

记$\hat{\theta}$为基于原始样本X的预测值,$\hat{\theta}^*$为基于Bootstrap样本X^*的预测值,$\hat{\theta}_i^*$为基于第i个Bootstrap样本X_i^*的预测值. 参考文献[25]给出了Bootstrap置信区间估计的4种常用方法.

1) 标准差的区间估计(Standard error, SE). 计算$\{\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_B^*\}$的均值和方差,即$\bar{\theta}^*=\dfrac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$, $\mathrm{Var}(\theta^*)=\dfrac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^*-\bar{\theta}^*)^2$. 当$\hat{\theta}$服从或近似服从正态分布时,$\mathrm{Var}(\theta^*)$为$\mathrm{Var}(X)$的估计值. 当显著性水平为$\alpha$时,用$u_{1-\alpha/2}$表示标准正态分布的$1-\alpha/2$百分位数,则$\theta$的标准差Bootstrap置信区间为$(\hat{\theta}-u_{1-\alpha/2}\mathrm{Var}(\theta^*),\hat{\theta}+u_{1-\alpha/2}\mathrm{Var}(\theta^*))$.

2) 百分位数区间估计(Percentile bootstrap, PB). 将$\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_B^*$由小到大排序得$\hat{\theta}_{(1)}^*\leqslant\hat{\theta}_{(2)}^*\leqslant\cdots\leqslant\hat{\theta}_{(B)}^*$,将$\hat{\theta}^*$的分布作为$\theta$分布的近似. $\hat{\theta}^*$的近似分位数$\hat{\theta}_{\alpha/2}^*$、$\hat{\theta}_{1-\alpha/2}^*$使得$P(\hat{\theta}_{\alpha/2}^*<\hat{\theta}^*<\hat{\theta}_{1-\alpha/2}^*)=1-\alpha$, 则近似可得$P(\hat{\theta}_{\alpha/2}^*<\theta<\hat{\theta}_{1-\alpha/2}^*)=1-\alpha$. 令$n_1=\left[B\times\dfrac{\alpha}{2}\right],n_2=\left[B\times\left(1-\dfrac{\alpha}{2}\right)\right]$,则$\theta$在$1-\alpha$置信水平下的百分位数Bootstrap置信区间为$(\hat{\theta}_{(n_1)}^*,\hat{\theta}_{(n_2)}^*)$.

3) t百分位数区间估计(t-Percentile Bootstrap, B-t). 对每个Bootstrap样本X_i^*计算T统计量,即$X_i^*=\dfrac{\hat{\theta}_i^*-\hat{\theta}}{\sqrt{\mathrm{Var}(\theta^*)}},i=1,2,\cdots,B$. 将$T_1^*,T_2^*,\cdots,T_B^*$由小到大排序得$T_{(1)}^*,T_{(2)}^*,\cdots,T_{(B)}^*$. 令$n_1=\left[B\times\dfrac{\alpha}{2}\right]$, $n_2=\left[B\times\left(1-\dfrac{\alpha}{2}\right)\right]$,当显著性水平为$\alpha$时,$\theta$的$t$百分位数Bootstrap置信区间为$(\hat{\theta}-T_{(n_1)}^*\mathrm{Var}(\theta^*),\hat{\theta}+T_{(n_2)}^*\mathrm{Var}(\theta^*))$.

4) 加速偏差修正区间估计(Bias-corrected and accelerated, BCa). 定义$\hat{\varphi}$和$\hat{z_0}$分别表示加速因子和

修正偏差,用来修正可能存在的潜在偏差. 当显著性水平为α时,θ的BCa Bootstrap置信区间为$(\hat{\theta}^{*(\varphi_1)}, \hat{\theta}^{*(\varphi_2)})$. φ_1和φ_2的计算方法为

$$\varphi_1 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{\varphi}(\hat{z}_0 + z^{(\alpha)})}\Big), \tag{1}$$

$$\varphi_2 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\varphi}(\hat{z}_0 + z^{(1-\alpha)})}\Big). \tag{2}$$

其中: z^{α}是标准正态分布的第100α分位数, $\Phi(\cdot)$是标准正态分布的累积函数.

由文献[29-30]可知$\hat{z}_0$和$\hat{\varphi}$的计算方法为

$$\hat{z}_0 = \Phi^{-1}\Big(\frac{\#\{\hat{\theta}_i^* < \hat{\theta}\}}{B}\Big),\ i = 1, 2, \cdots, B; \tag{3}$$

$$\hat{\varphi} = \frac{\sum\limits_{i=1}^{M}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^3}{6\Big(\sum\limits_{i=1}^{M}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^2\Big)^{3/2}},\ \hat{\theta}_{(\cdot)} = \sum_{i=1}^{M}\hat{\theta}_{(-i)}. \tag{4}$$

其中: $\Phi^{-1}(\cdot)$表示$\Phi(\cdot)$的反函数; $\#\{\hat{\theta}_i^* < \hat{\theta}\}$表示满足$\hat{\theta}_i^* < \hat{\theta}$条件的元素个数; $X_{(-i)}$表示原始样本X删除第i个观测值后的样本数据集; $\hat{\theta}_{(-i)}$表示基于样本$X_{(-i)}$的预测值; M表示一定的样本数,可取$M = [10\% \times n]$.

1.3 预测区间质量评估方法

预测区间PI被定义为在置信水平α下预测值范围的估计. 参照文献[31]采用范围概率(Prediction interval coverage probability, PICP)、平均预测区间宽度(Mean prediction interval width, MPIW)以及综合指标(Coverage Width-based Criterion, CWC)衡量预测质量.

PICP是预测准确性的表征,PICP越高,预测区间包含的真实值越多. PICP的计算公式为

$$\mathrm{PICP} = \frac{1}{N}\sum_{j=1}^{N} c_j; \tag{5}$$

$$c_j = \begin{cases} 1,\ \mathrm{low}(\hat{t}_j) \leqslant t_\alpha(j) \leqslant \mathrm{up}(\hat{t}_j); \\ 0,\ \mathrm{else}. \end{cases} \tag{6}$$

其中: N表示测试样本的数量, $t_\alpha(j)$表示第j个测试样本的真实值, $\mathrm{low}(\hat{t}_j)$和$\mathrm{up}(\hat{t}_j)$分别表示第j个预测区间的下限和上限.

MPIW的计算公式为

$$\mathrm{MPIW} = \frac{1}{N}\sum_{j=1}^{N}|\mathrm{low}(\hat{t}_j) - \mathrm{up}(\hat{t}_j)|. \tag{7}$$

区间预测的目标是尽可能高的PICP和尽可能小的MPIW. CWC是一个评价预测区间质量的综合指标,CWC越小越好,计算公式为

$$\mathrm{CWC} = \mathrm{MPIW}(1 + \gamma\cdot\mathrm{PICP}\cdot \mathrm{e}^{-\eta(\mathrm{PICP}-\mu)}); \tag{8}$$

$$\gamma = \begin{cases} 0,\ \mathrm{PICP} \geqslant \mu; \\ 1,\ \mathrm{PICP} < \mu. \end{cases} \tag{9}$$

η和μ为常数,一般μ与置信水平一致,取$\mu = 1-\alpha$; η称为惩罚参数,通常取一个较大的值($\eta = 50$),以便于放大PICP与μ的区别. 指数项$\mathrm{e}^{-\eta(\mathrm{PICP}-\mu)}$的作用如下: 当$\mathrm{PICP} \leqslant \mu$时,CWC快速增大,即获得的PI质量较差; 当$\mathrm{PICP} \geqslant \mu$时,指数项影响被消除,CWC等于MPIW.

2 点预测模型的选取

行程时间具有历史回归性,基于数据挖掘的预测方法可以通过历史数据来推理未来的交通状态,不要求复杂训练和精确建模. 针对小波神经网络(Wavelet neural network, WNN)和K最近邻算法(K nearest neighbour, KNN)两种常用的预测方法,本文对其进行优化和比较分析. 为达到较优的预测区间性能,选择误差较小的方法作为建模Step 3中的点预测模型.

2.1 小波神经网络模型

采用紧致型小波神经网络,用小波基函数代替BP神经网络中隐含层神经元的传统传递函数,网络结构见图1.

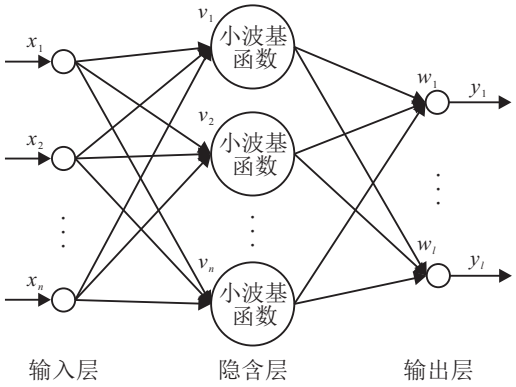

图1 紧致型小波神经网络结构

$x_1, x_2, \cdots, x_p$为输入参数; $y_1, y_2, \cdots, y_l$为小波神经网络的预测输出; v_j为输入层和隐藏层的连接权值, $j = 1, 2, \cdots, q$, q为隐含层节点数. 令$h(j)$为隐含层第j个节点输出,得

$$h(j) = h_j\Big[\frac{\sum\limits_{i=1}^{p} v_j x_i - \chi_j^b}{\chi_j^a}\Big]. \tag{10}$$

其中: h_j为小波基函数, χ_j^a和χ_j^b分别为伸缩因子及平移因子. 母小波基函数选用Morlet小波函数$y = \cos(1.75x)\mathrm{e}^{-\frac{x^2}{2}}$.

假设$O(k)$为输出层第k个节点, w_k为隐含层到输出层的网络权值, $k = 1, 2, \cdots, l$, l为输出层节点数,则输出层的计算公式表示为

$$O(k)=\sum_{j=1}^{l} w_k h(j). \tag{11}$$

为了改进小波神经网络易陷入振荡效应和局部极小的问题,利用遗传算法进行优化,以降低预测效果的随机波动,改进后的GAWNN流程见图2.

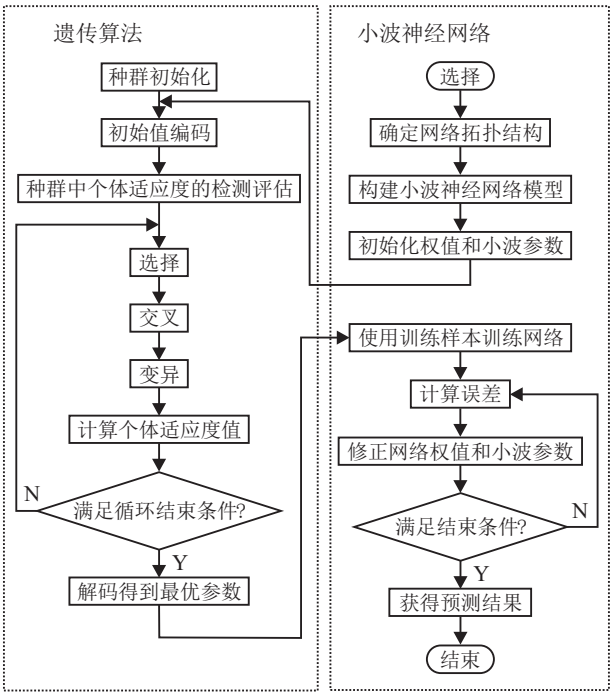

图 2　GAWNN预测模型基本流程

2.2　K 最近邻模型

K 最近邻算法利用历史数据中相似记录对预测值的贡献,采用特征向量最相似的 K 个近邻进行预测. 一般分为历史数据集构建、特征向量选择、K 值标定、相似程度距离计算和贡献度加权估计5个步骤[12].

1) 历史数据集构建. 根据交调标准将车型分为6类,选择分析时段为7:00～00:00(次日),将历史数据集以15 min为间隔,区分工作日与节假日,获得 $2\times 18\times 4\times 6=816$ 组样本.

2) 特征向量选择. 采用主成分分析法来确定特征向量,既可以避免主观性选取,又能降低算法的时间复杂度. 备选主成分包括三均值、均值、标准差、方差、分位数(10%、25%、75%和90%)、中位数、众数、变异系数、偏度系数和峰度系数. 定义三均值为25%、50%和75%三个分位数的加权求和,对应权值为0.25、0.5和0.25. 选取2015年某热点OD工作日小型客车样本数据较多的某一时段作为分析样本,计算252组分析样本的备选主成分,构建 252×13 的历史数据标准化矩阵以及 13×13 的相关系数矩阵,采用雅可比法求出相关系数矩阵对应的13个特征值 $\lambda_i(i=1\sim 13)$,则 $\lambda_g\Big/\sum_{i=1}^{13}$ 即为第 g 个主成分的贡献率. 各主成分贡献率与累计贡献率计算结果如表1所示.

表 1　备选主成分贡献率与累计贡献率

序号	备选主成分	贡献率	累计贡献率
1	三均值	0.6233	0.623 3
2	均值	0.209 7	0.832 9
3	标准差	0.129 6	0.962 6
4	方差	0.026 5	0.989 1
5	中位数	0.005 6	0.994 7
6	众数	0.001 7	0.996 4
7	变异系数	0.001 4	0.997 9
8	偏度系数	0.001 0	0.998 9
9	峰度系数	0.000 5	0.999 4
10	10%分位数	0.000 4	0.999 8
11	25%分位数	0.000 1	1.000 0
12	75%分位数	0.000 0	1.000 0
13	90%分位数	0.000 0	1.000 0

由表1可知,三均值、均值和标准差的累计贡献率已经达到96%,选择这3个样本特征作为特征向量的最终主成分.

由于行程时间的序列自相关性,选择具有较高相关系数的前几期特征参数作为当前时段的特征向量. 选取2015年某热点OD工作日小型客车 $T-9$ 至 T 时段的 252×10 组数据进行相关性分析,以三均值、均值、标准差作为每组样本的特征参数,分别计算 T 时段与前几期特征参数的相关系数. 由图3可知,当期时段特征参数与前3期的相关系数均大于0.9,考虑到计算效率,故选择当前时期的前3期行程时间特征参数作为特征向量.

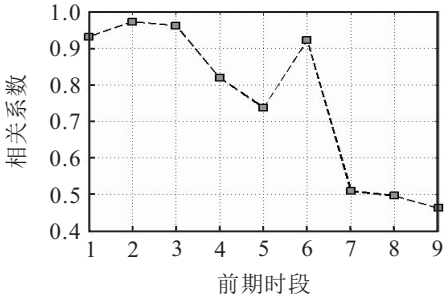

图 3　当前时期与前期时段的相关性分析

3) K 值标定. K 值是直接影响模型预测结果的惟一参数. 本文采用交叉验证法确定各历史数据集中 K 值的最好预测效果取值,具体过程略.

4) 相似程度距离计算. 选择欧几里德距离来表示两个特征向量之间的相似程度,相似程度距离的计算公式为

$$d=\sqrt{\sum_{f=1}^{F}\lambda_f\cdot(F_{P,f}-F_{A,f})^2}. \tag{12}$$

其中: F 表示特征向量个数, λ_f 表示第 f 个属性的主成分贡献率, $F_{P,f}$ 和 $F_{A,f}$ 分别表示历史记录特征向量和预测时刻特征向量的第 f 个属性.

5) 贡献度加权估计. 不同相似程度的近邻具有不同的预测贡献, 越相似的记录对预测值的影响越大. 采用 K 个历史值加权获得行程时间预测值 t_p, 计算公式为

$$t_p = \sum_{k=1}^{K_0} \tau_k t_a(k),\ k = 1, 2, \cdots, K; \tag{13}$$

$$\tau_k = \frac{\exp(-d_k)}{\sum_{j=1}^{K_0} \exp(-d_j)},\ k = 1, 2, \cdots, K. \tag{14}$$

其中: τ_k 表示第 k 个近邻的权重; d_k 表示第 k 个近邻与预测值之间特征向量的距离; $t_a(k)$ 表示第 k 个近邻历史记录的行程时间; K_0 表示最优 K 值, 采用交叉验证法确定.

2.3 点预测模型的误差比较

以陕西省2015年收费数据为原始数据, 以某热点OD的小型客车工作日行程时间(单位: s)为分析对象. 选取7:45 ~ 14:00为研究时段, 15 min为间隔, 共25个时段. 每个时段样本进行预处理后随机抽取50个工作日样本作为测试样本集 D_2, 其余工作日的样本为训练样本集 D_1. 采用相同的训练样本、测试样本和预测输入, 分别利用WNN、GAWNN和KNN进行预测, 通过平均绝对误差(E_{MAE})和平均相对误差(E_{MAPE})比较3种模型的预测精度. 其中: 输入为预测时段前3期的三均值、均值和标准差, 预测时段当期三均值的取值定义为实际值. EMAE和EMAPE的计算公式为

$$E_{\mathrm{MAE}} = \frac{1}{N}\sum_{i=1}^{N} |t_p(i) - t_a(i)|, \tag{15}$$

$$E_{\mathrm{MAPE}} = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{t_p(i) - t_a(i)}{t_a(i)} \times 100\,\% \right|. \tag{16}$$

其中: $t_p(i)$ 为第 i 个样本的预测值, $t_a(i)$ 为第 i 个样本的实际值, N 为测试样本个数.

计算各时段50个测试样本的预测误差, 得到每个时段的预测误差指标 E_{MAE} 和 E_{MAPE}. 计算25个时段 E_{MAE} 与 E_{MAPE} 的均值和标准差, 如表2所示.

表 2 3种预测模型的误差比较

指标		KNN	WNN	GAWNN
E_{MAE}	均值	2.778 0	8.483 1	4.260 3
	标准差	1.750 2	12.026 7	2.500 0
E_{MAPE}	均值	5.964 0	19.308 6	9.324 7
	标准差	3.304 6	27.739 0	4.760 5

由表2可知: GAWNN能够改进WNN的预测误差及其误差波动性; KNN的预测精度在3种模型中最佳, E_{MAE} 和 E_{MAPE} 的均值都为最小, 且不同时段的预测误差波动性最小. 可见, KNN是一种用于高速公路短时行程时间点预测的有效方法, 选择KNN模型作为点预测模型可以获得较小的预测误差.

3 实例分析

采用2.3节中的训练样本集 D_1 和测试样本集 D_2 为分析数据集, 构建基于Bootstrap-KNN的高速公路行程时间区间预测模型, 通过预测区间宽度表征预测结果的不确定性. 采用样本容量 $m = n$, 抽样次数 $B = 1\,000$.

3.1 不同Bootstrap策略区间预测性能分析

通过实证分析来选择最佳Bootstrap置信区间估计方法. 在相同的置信水平95 %下, 采用4种常用的Bootstrap置信区间估计方法, 计算25个时段测试样本的区间预测性能PICP、MPIW和CWC, 指标均值如表3所示. 结果表明, 百分位数区间估计PB的PICP最高, 且CWC最小. 说明PB估计方法在4种方法中最佳, 估计结果更为可靠. 因此, 在基于Bootstrap-KNN的建模过程中, 选择PB作为Bootstrap置信区间估计方法可以获得较优的区间预测性能.

表 3 4种Bootstrap方法的测试样本集性能指标均值

性能指标	SE	PB	B-t	BCa
PICP	0.434 4	0.529 6	0.416 0	0.283 3
MPIW	62.884 0	4.738 7	48.506 7	2.869 9
CWC	7.95e+14	3.9e+12	1.82e+15	1.00e+17

3.2 KNN与Bootstrap-KNN的预测误差比较分析

分别采用KNN和Percentile Bootstrap-KNN对25个时段的测试样本集进行预测, 比较和验证区间预测相较于点预测在预测误差方面的改进.

定义MSE/n 表示实际值与预测值的误差平方和均值, 计算结果见表4. 其中: 实际值为预测时段当期的三均值, Percentile Bootstrap-KNN的预测值取区间预测上、 下限的均值. 由表4可知, Percentile Bootstrap-KNN的MSE/n 值比KNN小. 结果表明, 采用相同的点预测模型, 基于Bootstrap的区间预测能够有效减小预测误差.

表 4 2种预测模型的MSE/n 值

指标	KNN	PB-KNN
MSE/n	95.018 1	85.971 9

3.3 Bootstrap-KNN区间预测误差分析

对测试样本集 D_2 中某一时段采用Percentile Bootstrap-KNN模型进行区间预测, 并与预测时段当

期样本数据集的三均值进行比较,比较结果如图4所示. 结果表明,PB-KNN模型能够有效跟踪实际行程时间的变化. 当行程时间剧烈波动变化时,预测区间相较于平稳时期较宽,即该预测值的可信度较低. 可见,PB-KNN模型既可以提供预测值的估计范围,又包含了预测结果的可信度. 预测区间宽度越小,则预测结果的可靠性越高,结果越可信.

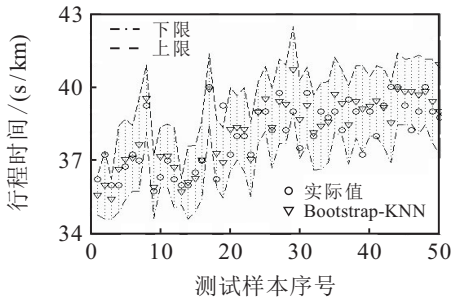

图 4　基于Percentile Bootstrap-KNN的行程时间区间预测

4　结　论

目前行程时间短时预测多为点预测,不能描述预测结果的可信度. 为避免点预测的缺陷,以高速公路收费系统为数据来源,提出了基于Bootstrap的高速公路行程时间区间预测模型,为衡量预测结果的可靠性提供了一种有效的评估手段;深入讨论了预测模型和Bootstrap置信区间估计方法这两个关键步骤,以寻求较优的区间预测性能;通过比较预测误差和预测区间质量指标表明,PB-KNN区间预测性能模型在降低点预测误差和量化预测值可信度方面均有较优表现. 行程时间的预测模型数量众多, Bootstrap策略为衡量各预测模型的可靠性提供了一种手段,因此需要进一步考虑.

参考文献(References)

[1] Kasai M, Warita H. Refinement of pattern-matching method for travel time prediction[J]. Int J of Intelligent Transportation Systems Research, 2014, 13(2): 84-94.

[2] Billings D, Yangt J. Application of the ARIMA models to urban roadway travel time prediction — A case study[C]. Int Conf on Systems, Man, and Cybernetics. Taipei: IEEE, 2006: 2529-2534.

[3] 王宝杰, 王炜, 杨敏, 等. 基于Kalman滤波行程时间预测的BRT车速诱导[J]. 吉林大学学报: 工学版, 2014, 44(1): 41-46.
(Wang B J, Wang W, Yang M, et al. BRT speed induction based on Kalman travel time prediction[J]. J of Jilin University: Engineering and Technology Edition, 2014, 44(1): 41-46.)

[4] 赵建东, 王浩, 刘文辉. 高速公路旅行时间的自适应插值卡尔曼滤波预测[J]. 华南理工大学学报: 自然科学版, 2014, 42(2): 109-115.
(Zhao J D, Wang H, Liu W H. Prediction of expressway travel time based on adaptive interpolation Kalman filter[J]. J of South China University of Technology: Natural Science Edition, 2014, 42(2): 109-115.)

[5] 刘江用, 云美萍, 闫亚文, 等. 基于径向基函数神经网络的城市道路路段行程时间实时预测模型[J]. 交通信息与安全, 2011, 29(5): 31-35.
(Liu J Y, Yun M P, Yan Y W, et al. A real-time travel time prediction model based on RBF neural network[J]. J of Transport Information and Safety, 2011, 29(5): 31-35.)

[6] 杨兆升, 保丽霞, 朱国华. 基于Fuzzy回归的快速路行程时间预测模型研究[J]. 公路交通科技, 2004, 21(3): 78-81.
(Yang Z S, Bao L X, Zhu G H. An Urban express travel time prediction model based on Fuzzy regression[J]. J of Highway and Transportation Researchand Development, 2004, 21(3): 78-81.)

[7] 张娟, 孙剑. 基于SVM的城市快速路行程时间预测研究[J]. 交通运输系统工程与信息, 2011, 11(2): 174-179.
(Zhang J, Sun J. Prediction of urban expressway travel time based on SVM[J]. J of Transportation Systems Engineering and Information Technology, 2011, 11(2): 174-179.)

[8] 邱淳风, 王珊, 王超群. 基于支持向量回归的行程时间预测算法[J]. 计算机时代, 2014, 32(4): 40-42.
(Qiu C F, Wang S, Wang C Q. Travel-time prediction algorithm based on support vector regression[J]. Computer Era, 2014, 36(4): 40-42.)

[9] 刘克. 高速公路的路段行程时间估计与预测方法研究[D]. 北京: 北京交通大学交通运输学院, 2013: 17-34.
(Liu K. Estimation Prediction of link travel time for freeways[D]. Beijing: School of Transportation, Beijing Jiaotong University, 2013: 17-34.)

[10] Willem W E, Chris M J, Bieke M. A parsimonious method for offline freeway travel time estimation from sectional speed detectors[J]. J of Intelligent Transportation Systems, 2014, 18(1): 67-80.

[11] 王浩. 基于收费数据的高速公路旅行时间自适应插值卡尔曼滤波预测研究[D]. 北京: 北京交通大学机械与电子控制工程学院, 2014: 23-35.
(Wang H. Highway travel time prediction research based on adaptive interpolation Kalman filter and toll collection data[D]. Beijing: School of Mechanical and Electronic Control Engineering, Beijing Jiaotong University, 2014: 23-35.)

[12] 王翔, 陈小鸿, 杨祥妹. 基于K最近邻算法的高速公路短时行程时间预测[J]. 中国公路学报, 2015, 28(1): 102-111.
(Wang X, Chen X H, Yang X M. Short term prediction of highway travel time based on K nearest neighbor algorithm[J]. China J of Highway and Transport, 2015, 28(1): 102-111.)

[13] 唐俊. 基于浮动车数据的高速公路路段行程时间预测方法研究及系统实现[D]. 广州: 中山大学软件学院, 2011: 39-68.

(Tang J. Freeway link travel time prediction research and system implementation based on float car data[D]. Guangzhou: School of Software, Zhongshan University, 2011: 39-68.)

[14] Evangelos M, Josep M S, Evangelia C, et al. A robust method for real time estimation of travel times for dense urban road networks using point-to-point detectors[J]. Transport, 2015, 30(3): 264-272.

[15] Zhang Y, Shi W H, Liu Y C. Comparison of several trac forecasting methods based on travel time index data on weekends[J]. J of Shanghai Jiaotong University(Science), 2010, 15(2): 188-193.

[16] 毕松, 车磊, 赵忠诚, 等. 城市路网路段行程时间预测研究综述[J]. 计算机仿真, 2014, 31(7): 157-160.
(Bi S, Che L, Zhao Z C, et al. A survey on the link travel time prediction for urban road net[J]. Computer Simulation, 2014, 31(7): 157-160.)

[17] 陈旭梅, 龚辉波, 王景楠. 基于SVM和Kalman滤波的BRT行程时间预测模型研究[J]. 交通运输系统工程与信息, 2012, 12(4): 29-34.
(Chen X M, Gong H B, Wang J N. BRT vehicle travel time prediction based on SVM and Kalman filter[J]. J of Transportation Systems Engineering and Information Technology, 2012, 12(4): 29-34.)

[18] 丁宏飞, 李演洪, 刘博, 等. 基于BP神经网络与SVM的快速路行程时间组合预测研究[J]. 计算机应用研究, 2016, 33(10): 1-6.
(Ding H F, Li Y H, Liu B, et al. Expressway's travel time prediction based on combined BP neural network and support vector machine approach[J]. Application Research of Computers, 2016, 33(10): 1-6.)

[19] 田甜, 王秀玲, 吕芳. 基于Kalman和ARIMA组合模型的路段行程时间预测[J]. 信息技术, 2016, 40(3): 148-150.
(Tian T, Wang X L, Lv F. Urban travel time prediction based on the Kalman and ARIMA model[J]. Information Technology, 2016, 40(3): 148-150.)

[20] 李嘉, 刘春华, 胡赛阳, 等. 基于交通数据融合技术的行程时间预测模型[J]. 湖南大学学报: 自然科学版, 2014, 41(1): 33-38.
(Li J, Liu C H, Hu S Y, et al. A travel time prediction model based on traffic data fusion technology[J]. J of Hunan University: Natural Sciences, 2014, 41(1): 33-38.)

[21] 江周, 张存保, 许志达, 等. 基于多源数据的城市道路网络行程时间预测模型[J]. 交通信息与安全, 2014, 32(3): 27-31.
(Jiang Z, Zhang C B, Xu Z D, et al. Development of a travel time prediction model for urban road using multi-source data[J]. J of Transport Information and Safety, 2014, 32(3): 27-31.)

[22] 赵建东, 徐菲菲, 张琨, 等. 融合多源数据预测高速公路站间旅行时间[J]. 交通运输系统工程与信息, 2016, 16(1): 52-57.
(Zhao J D, Xu F F, Zhang K, et al. Highway travel time prediction based on multi-source data fusion[J]. J of Transportation Systems Engineering and Information Technology, 2016, 16(1): 52-57.)

[23] Cui Q H, Xia J X. Time-varying confidence interval prediction of travel time for urban arterials using ARIMA-GARCH model[J]. J of Southeast University: English Edition, 2014, 30(3): 358-362.

[24] Efron B, Tibshrani R. An introduction to the Bootstrap[M]. New York: Chapman and Hall, 1993: 1-28.

[25] Wu Z F. Measuring reliability in dynamic and stochastic transportation networks[D]. Lincoln: School of Civil Engineering, University of Nebraska Graduate Studies, 2015: 81-201.

[26] 蒋朝辉, 董梦林, 桂卫华, 等. 基于Bootstrap的高炉铁水硅含量二维预报[J]. 自动化学报, 2016, 42(5): 715-723.
(Jiang Z H, Dong M L, Gui W H, et al. Two-dimensional prediction for silicon content of hot metal of blast furnace based on Bootstrap[J]. Acta Automatica Sinica, 2016, 42(5): 715-723.)

[27] 赵玉, 祁春节. 大宗农产品价格风险评估——基于小波神经网络-Bootstrap方法的实证研究[J]. 技术经济, 2014, 33(3): 75-79.
(Zhao Y, Qi C J. Evaluation on prices risk of bulk agricultural products: Empirical study based on wavelet neural network — Bootstrap method[J]. Technology Economics, 2014, 33(3): 75-79.)

[28] 沈盟. 基于Bootstrap方法的金属期货市场风险测度VaR和ES的区间预测[D]. 成都: 西南交通大学理学院, 2015: 21-36.
(Shen M. Bootstrap prediction intervals or value at risk and expected shortfall or metal futures market[D]. Chengdu: School of Science, Southwest Jiaotong University, 2015: 21-36.)

[29] 黎光明, 张敏强. 概化理论方差分量置信区间估计方法的比较[J]. 统计与决策, 2013, 29(9): 14-17.
(Li G M, Zhang M Q. Comparison on confidence interval estimation methods of variance component base on generalizability theory[J]. Statistics & Decision, 2013, 29(9): 14-17.)

[30] 魏艳华, 王丙参, 邢永忠. 基于Bootstrap方法的回归分析的比较[J]. 统计与决策, 2016, 32(3): 77-79.
(Wei Y H, Wang B C, Xing Y Z. The comparison of regression analysis based on the Bootstrap method[J]. Statistics & Decision, 2016, 32(3): 77-79.)

[31] 韩帅, 李树刚. 基于区间预测模型的流感趋势预测[J]. 计算机仿真, 2014, 31(9): 237-242.
(Han S, Li S G. Influenza trends forecast based on interval prediction model[J]. Computer Simulation, 2014, 31(9): 237-242.)

(责任编辑: 齐 霁)