

OBD 支持下公交车到达时间的回归预测方法^①

董红召^{②*} 赵龙钢^{*} 赵晨馨^{*} 张亮^{**} 孔娟娟^{**}

(^{*} 浙江工业大学智能交通系统联合研究所 杭州 310014)

(^{**} 浙江嘉科电子有限公司 嘉兴 314000)

摘 要 公交车到达时间的精准预测是共享公交专用道研究的基础,为此基于车辆车载自动诊断系统(OBD)数据提出一种基于证据理论优化高斯过程回归(DS-GPR)的公交车到达时间预测方法。首先分析影响公交车到达时间的影响因素,随后通过历史融合数据训练得到模型的超参数,并且最终通过 D-S 证据理论为高斯过程的回归输出分布分配全局信任度,得到当前时刻信任度最优的输出值。最后通过实例对 DS-GPR 模型的预测性能进行对比分析,证明了 DS-GPR 模型对公交车到达时间的良好预测性能。

关键词 车载自动诊断系统(OBD);数据融合;高斯过程回归(GPR);证据理论;到达时间预测

0 引言

在间歇性共享公交专用道的探索研究中,共享策略是否可以达到预期目标的关键在于对公交车到达时间的掌握程度。因此,对公交车到达时间的预测成为共享公交专用道策略中最关键的环节^[1-5]。为保证共享公交专用道策略的有效性和稳定性,需要一种鲁棒性强、实时性好、预测精度高的到达时间预测方法,而现有的一些时间预测方法研究并不能满足其预测性能要求。

目前在智能交通领域,国内外学者对公交车到达时间预测做了大量的研究,现有时间预测模型主要有基于历史数据预测模型、机器学习预测模型、多源数据预测模型 3 种。其中,赖永炫等人^[6]、Marko 等人^[7]对海量历史数据进行分类加权处理,利用历史数据集提取出的时间特征预测车辆实时的到达时间。此类基于历史数据的预测方法泛化能力较差,无法适应实际交通中随机多变的特点。李华民等人^[8]根据射频识别技术(radio frequency identifica-

tion, RFID)的电子数据对公交车到达时间进行预测。Lin 等人^[9]以时空路径为依据,提出基于长短记忆卷积神经网络模型。Liu 等人^[10]提出基于评价机制的支持向量回归(support vector regression, SVR)预测模型,与其他同类型模型^[11]相比其预测精度大大提高。这类机器学习预测模型可以有效解决高维数据模型搭建的问题并具有较强的泛化能力,但其需要长期的学习训练过程才可以得到符合要求的预测模型,故无法满足预测模型实时性的要求。

随着车辆车载自动诊断系统(on board diagnostics, OBD)的普及,车辆行驶数据可以实时获取,结合视频卡口等传统采集方法,多源数据的引入可以使车辆到达时间的实时预测更加精确,多种检测器得到的全方位数据更可靠和全面。刘迎等人^[12]根据交通数据融合建立自回归平衡模型(autoregressive integrated moving average model, ARIMA)优化改进支持向量机(support vector machines, SVM)行程时间预测模型。张威威等人^[13]把时间过程分为线性和非线性两部分,利用主成分分析支持下回归树

① 国家自然科学基金(61773347)和浙江省公益技术研究(LGF20F030001)资助项目。
② 男,1969 年生,博士,教授;研究方向:智能交通系统,智能机电控制系统;联系人,E-mail: its@zjut.edu.cn
(收稿日期:2020-02-06)

模型 (principal components analysis-gradient boosting decision tree, PCA-GBDT) 方法融合计算得到最终预测结果。范广鹏等人^[14]利用 OBD、路段行程时间等数据,提出基于长短期记忆(long short-term memory, LSTM)神经网络和卡尔曼滤波的数据融合预测模型,大幅提升预测精度。华雪东等人^[15]基于数据融合建立了公交车到达时间的组合预测模型。以上多源数据预测方法模型结构复杂,所需数据集过多,导致计算时间增长,忽略了较难获取的实时流量数据,导致实时性不好,在极端交通环境下,例如恶劣天气、拥堵、突发事件等,因数据质量无法保证,造成模型鲁棒性不佳。

在交通研究中,车辆行驶过程中位置、方向、速度及加速等工况数据体现当前车辆的运动姿态,这些信息直接决定下一刻车辆的位置及状态,相邻时刻数据间相关性极大。而公交车辆到达时间预测的性质为一个时变性极强的非线性的独立预测,现有研究的预测方法通常无法摆脱前后信息相互影响的相关性,造成预测结果无法达到理想效果。高斯过程回归 (Gaussian process regression, GPR) 具有良好的独立性,可以根据动态数据生成多个相互独立的预测分布,与其他模型相比可以克服预测过程中时变性和相关性的影响,但同时这个过程会产生大量的概率随机值,使预测值出现剧烈浮动;为消除这些概率随机值对其预测结果的干扰,研究中引入证据理论 (D-S) 通过当前时刻车辆信息所构成的认知框架 (证据集) 对 GPR 的浮动输出值进行信任度分配,建立证据集优化的高斯过程回归 (DS-GPR) 公交车到达时间预测模型;保证信任度最大值作为模型的输出,实现精确预测车辆到达未来某一点的时间。

1 DS-GPR 公交车到达时间预测模型

1.1 模型建立

高斯过程回归是一种基于高斯随机函数的回归方法,该函数由协方差函数 $k(x, x')$ 和均值函数 $m(x)$ 确定,如式(1)所示。

$$\begin{cases} m(x) = E[f(x)] \\ k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))] \end{cases} \quad (1)$$

其中 x, x' 属于 R 为任意变量,高斯回归模型可利用核函数将多维数据向量映射到高维空间,核函数为正定协方差矩阵。

高斯过程回归模型输出为概率分布,取到均值概率最大,产生误差概率最小;距均值越远取到的概率越小,产生误差概率越大。现有研究中一般取概率最大点为输出值^[16],但概率最大处的分布值往往存在概率误差,致使最大概率处产生误差浮动,导致预测性能降低。而且,核函数的存在使模型在训练样本时涉及大量矩阵求逆过程,尤其当样本量维度较大时计算量和计算时间会急剧上升。为避免长时间计算及消除误差浮动,利用 D-S 证据理论模型对 GPR 输出分布函数进行优化。对于此类不确定、不精确的交通数据,利用 D-S 证据理论对目标信息构建多角度多维度的认知框架,验证其有效性指标,识别目标参数^[17-19],削减模型计算量。

根据认知框架内的信息证据,寻找全局信任度最高的点作为模型最终输出值,消除以往简单取最大概率点的误差浮动。

利用高斯过程回归预测车辆到达时间,可定义回归过程为

$$y_i = f(x_i) + \varepsilon \sim GP(m(x_i), k(x_i, x'_i)) + \varepsilon \quad (2)$$

式中, $x_i \in R^\lambda$ 为 λ 维的车辆输入,包括车辆历史行驶状态信息,即路径、交叉口位置、公交站位置、车流量、发车时段、发车日期、车辆经纬度、速度、加速度和方向角; y_i 为模型目标输出分布; ε 为均值为 0、方差为 σ_n^2 的高斯噪声,表示罕见事件 (如交通事故、突发事件等) 对车辆到达时间的影响;平方指数协方差函数 (squared exponential covariance function) 核函数具有性能稳定、拟合良好、误差小、形式简单和参数少的特点,故选择平方指数协方差函数作为核函数,如式(3)所示。

$$k(x, x') = \sigma_f^2 \cdot e^{\left[\frac{-(x-x')^T \Lambda^{-1} (x-x')}{2\lambda^2} \right]} \quad (3)$$

其中, Λ 为对角元素为 l^2 的对角阵, l 为尺度参量; σ_f^2 为样本传输方差,表示样本数据的覆盖程度。平方指数协方差函数中输入值的相关性会随距离的变大而不断变小,且无限可微,使模型运行过程更加平稳。

为简化模型,可将高斯噪声在协方差函数内表示,如式(4)所示。

$$k(x, x') = \sigma_f^2 \cdot e^{\left[\frac{-(x-x')^T \Lambda^{-1} (x-x')}{2l^2} \right]} + \sigma_n^2 \cdot \delta(x, x') \quad (4)$$

$\delta(x, x')$ 为克罗内克函数,当输出值相等时,其值为1,否则为0。参数集合 $\xi = [\Lambda, \sigma_f, \sigma_n]$ 为超参数。高斯回归模型的训练目的就在于找到最能拟合样本数据且泛化能力最好的超参数。高斯过程认为样本点 y 与预测值 $\hat{y}$ 可以认为是从高斯联合分布中采样到的点,则样本点 y 的先验分布如式(5)所示。

$$y \sim N(0, \mathbf{K}(X, X) + \sigma_n^2 \cdot \mathbf{I}_n) \quad (5)$$

式中, X 为输入训练集。

据高斯分布关系可得到样本值 y 与预测值 $\hat{y}$ 的分布为

$$\begin{bmatrix} y \\ \hat{y} \end{bmatrix} = N \left(0, \begin{bmatrix} \mathbf{K}(X, X) + \sigma_n^2 \cdot \mathbf{I}_n & \mathbf{K}(X, \hat{x}) \\ \mathbf{K}(\hat{x}, X) & \mathbf{k}(\hat{x}, \hat{x}) \end{bmatrix} \right) \quad (6)$$

其中 $\mathbf{K}(X, X)$ 为 N 阶正定协方差矩阵,如式(7)所示。

$$\mathbf{K}(X, X) = \begin{bmatrix} \mathbf{k}(x_1, x_1) & \mathbf{k}(x_1, x_2) & \cdots & \mathbf{k}(x_1, x_n) \\ \mathbf{k}(x_2, x_1) & \mathbf{k}(x_2, x_2) & \cdots & \mathbf{k}(x_2, x_n) \\ \vdots & \vdots & & \vdots \\ \mathbf{k}(x_n, x_1) & \mathbf{k}(x_n, x_2) & \cdots & \mathbf{k}(x_n, x_n) \end{bmatrix} \quad (7)$$

$\mathbf{K}(X, \hat{x}) = \mathbf{K}(\hat{x}, X)^T$ 为测试集 $\hat{x}$ 与训练集 X 的协方差矩阵,如式(8)所示。

$$\mathbf{K}(\hat{x}, X) = [\mathbf{k}(\hat{x}, x_1) \quad \mathbf{k}(\hat{x}, x_2) \quad \cdots \quad \mathbf{k}(\hat{x}, x_n)] \quad (8)$$

$\mathbf{k}(\hat{x}, \hat{x})$ 为测试集自身协方差矩阵, $\mathbf{I}_n$ 为 N 阶单位矩阵。

由此得到到达时间 $\hat{y}$ 的后验分布为

$$\hat{y} |_{X, y, \hat{x}} \sim N(E(\hat{y}), D(\hat{y})) \quad (9)$$

$E(\hat{y})$ 为测试点对应的到达时间的预测分布均值,如式(10)所示。

$$E(\hat{y}) = \mathbf{K}(\hat{x}, X)^T \cdot [\mathbf{K}(X, X) + \sigma_n^2 \cdot \mathbf{I}_n]^{-1} \cdot y \quad (10)$$

$D(\hat{y})$ 为对应的到达时间预测分布方差,如式(11)所示。

$$D(\hat{y}) = \mathbf{K}(\hat{x}, \hat{x}) - \mathbf{K}(\hat{x}, X)^T \cdot [\mathbf{K}(X, X) + \sigma_n^2 \cdot \mathbf{I}_n]^{-1} \cdot \mathbf{K}(\hat{x}, X) \quad (11)$$

将样本 $\Gamma = \{\tau_1, \tau_2, \cdots, \tau_i\}$ 作为认知框架,其每一个样本数据均视为认知框架上的一个证据,根据预测分布取得预测值集合为 $\{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_i\}$, 令数据源可信度为1,预测值信任度向量为 $\mathbf{M} = \{m_1, m_2, \cdots, m_i\}$ 。

假设各样本值提供错误预测的概率符合均匀分布,如式(12)所示。

$$p(\hat{y}_i | \tau_i) = \begin{cases} \frac{1}{i} & \hat{y}_i = y_i \\ 1 - \frac{1}{i} & \hat{y}_i \neq y_i \end{cases} \quad (12)$$

并且认为各样本对预测值的信任度相互独立,以后验概率最大处的值作为标准值,利用贝叶斯展开得到预测值成立的后验概率为

$$p(\Gamma | \hat{y}_i) = \frac{p(\Gamma) \prod_1^i p(\hat{y}_i | \tau_i)}{\sum_1^i [p(\Gamma) p(\hat{y}_i | \Gamma)]} \quad (13)$$

由于各样本点独立提供预测成立后验概率,即:

$$p(\hat{y}_i | \Gamma) = \prod_1^i \frac{1}{i} \cdot \prod_1^i (1 - \frac{1}{i}) \quad (14)$$

则基于式(14)可以建立该预测值在认知框架 Γ 中的基本信任分配函数:

$$m_i(\Gamma) = \int_{\Gamma} p(\Gamma | \hat{y}_i) d\tau_i \quad (15)$$

由式可知 $m_i(\Gamma) \in [0, 1]$, $\sum_1^i m_i(\Gamma) = 1$ 且 $m_i(\emptyset) = 0$, 满足 D-S 信任分配函数的特征。

预测分布中全局信任分配函数最优值点表示认知框架中所有数据对其值支持度最大,即认为该点预测值最接近真实值。

1.2 模型流程

DS-GPR 模型包含训练、更新、预测 3 个基本模块。

利用历史数据集训练出模型的超参数,训练次数由每次训练的学习率决定,最终得到使学习率收敛的超参数;取最新的 20 个 GPS、OBD 多源数据作为预测窗口数据集,每次多源数据更新后,模型输出

一个预测值,随后窗口数据集抛弃一个“最靠后”的数据,增加一个“最靠前”的数据,再基于最新的 20

个数据进行预测。模型流程如图 1 所示。
整个过程中DS-GPR模型实现了对车辆数据的

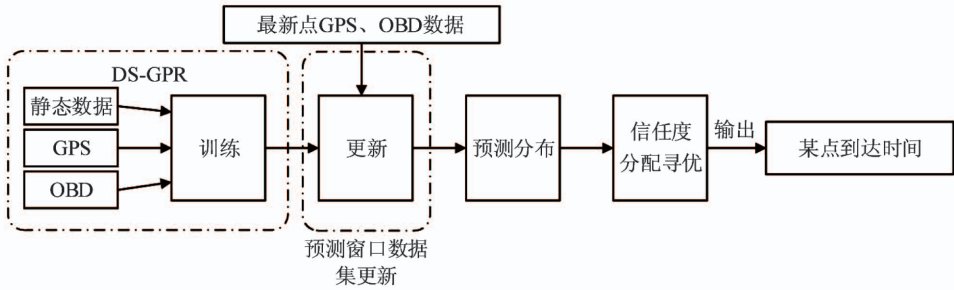


图 1 DS-GPR 模型执行流程

实时更新,结合车辆当前状态给出到达时间实时预测,经多次测试,该模型每次计算时间均小于1 s,大幅提高了输出结果的实时性和准确性。

2 DS-GPR 应用

2.1 实例数据

车辆到达时间预测真实时序数据为杭州市某线路全部公交车 2019 年 7 月 1 日至 15 日的某线路公

交车辆运营时间内的 OBD 数据和途径道路卡口数据,OBD 数据每 10 s 更新一次,每一条数据包括位置、时间、速度、加速度、方向角信息;由于公交车线路确定,所以通过路径、路况、交叉口信息已知,途径卡口数据经处理可以得到实时道路车流量数据。该公交车线路途径火车站、城市主城区、主要景区、大型商业区等,交通情况复杂,路况多变且高峰低谷明显,如图 2 所示,适宜作为实例验证分析对象。



图 2 公交车线路示意图

在其线路上共划分 18 个节点,包括交叉口或路段中心节点,平均每一节点相距 400 ~ 800 m,如表 1

所示。实例通过实时预测公交车到下一节点的时间来对比模型预测性能。

表 1 各节点划分结果

1. 建国南路-西湖大道	2. 中河高架路-西湖大道	3. 延安路-西湖大道
4. 解放路-延安路	5. 平海路-延安路	6. 庆春路-延安路
7. 庆春路-环城西路	8. 北山街-环城西路	9. 北山街-石函路
10. 北山街-北偃塔前山路	11. 北山街-共产党历史馆	12. 北山街-孤山路
13. 北山街-栖霞岭路	14. 灵隐路-植物园	15. 灵隐路-玉古路
16. 灵隐路-灵松山庄	17. 灵隐路-灵溪南路	18. 灵隐路-灵竺路

图 3 为处理后的原始行驶数据,表示公交车在每次信息更新时到达相邻下一节点的时间。可以看出到达时间波动较大,但大部分数据集中于某一定

值附近,具备一定的相似性。为保证模型的泛化能力,需将所有数据按影响因素进行聚类分析,划分原始数据集,提高预测精度。

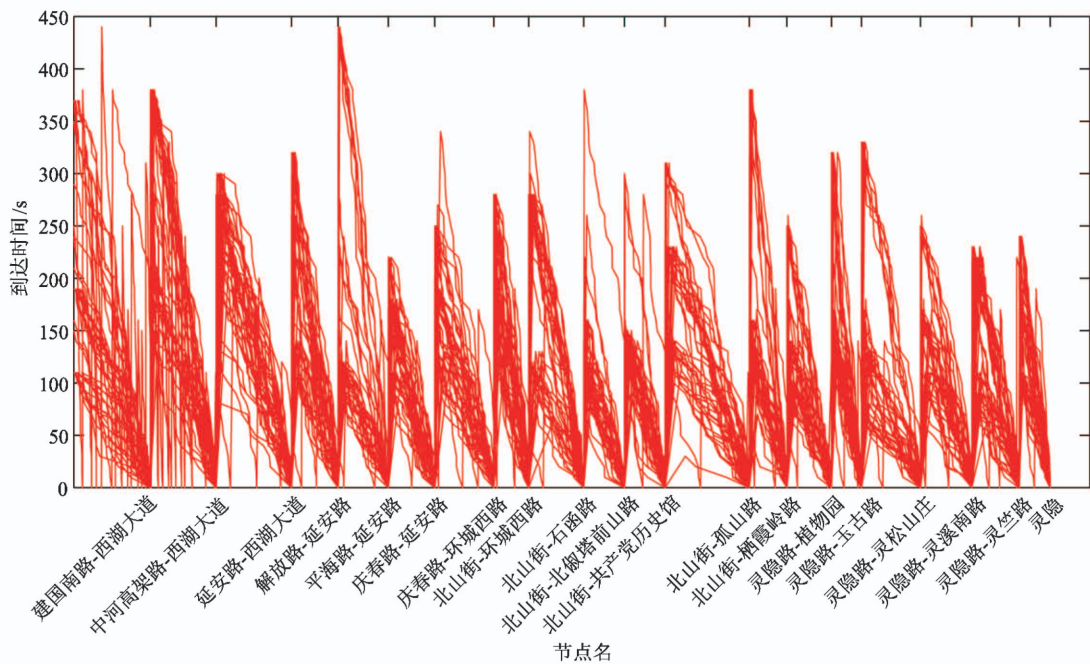


图 3 公交车行驶数据示意图

不同交通检测器数据采集频率不一致,使交通数据在时间序列和空间位置上不能同步。按照时间、空间特性匹配的原则收集城市道路网络中的各交通检测器数据,设置最小时间间隔阈值 t_{max} ,以同路段为约束将其转换为同时段、同间隔的关联数据;运用关联数据拟合预测模型,根据上一时段的融合数据实时预测下一时刻目标值,得到融合后的最优预测序列。

2.2 影响因素分析

2.2.1 节假日因素

由于节假日出行人数激增,同时路段相较于工作日交通更加拥堵,交通情况更加复杂,交通峰值持续时间更长,使得节假日对公交车到达时间的影响十分明显。取得部分节假日与工作日同时段的到达时间数据,如图 4 所示,可以看出节假日的到达时间明显大于工作日,尤其进入景区道路(灵隐路)之后,节假日到达时间更是显著高于工作日。可见节假日对交通特性影响较大,故将节假日因素作为预测到达时间的影响因素之一;输入参数当节假日时为 0,工作日为 1。

2.2.2 发车时间因素

在同一天中,各个时段道路交通情况也有很大差异,高峰时道路拥堵对到达时间也会产生一定影响。如图 5 所示,分别取早高峰(9:00)、中午小高峰(13:00)、平峰(15:00)、晚高峰(17:00)4 个时段的车辆行驶数据进行对比分析,可看出高峰时段内车辆到达时间与平峰时段差异较大;并且早晚高峰间也有一定差异,其原因在于早晚高峰的主要交通方向不同,具有潮汐性。可见发车时间与公交车的到达时间有一定的相关性,故将发车时间作为预测到达时间的影响因素之一。

2.3 模型验证

将现有 1 日至 14 日的数据作为训练集,15 日的数据作为验证集合;输入变量集合 $x_i \in R^{\lambda}$ 为分别包括位置、时间、速度、加速度、方向角、发车时间等特征值的 λ 维向量;超参数初始值设为 $\xi = [100, 1, -1]$,模型学习速率随训练次数更新如式(16)所示。

$$\rho(m) = \rho_{max} - \frac{\eta(\rho_{max} - \rho_{min})}{\eta_{max}} \tag{16}$$

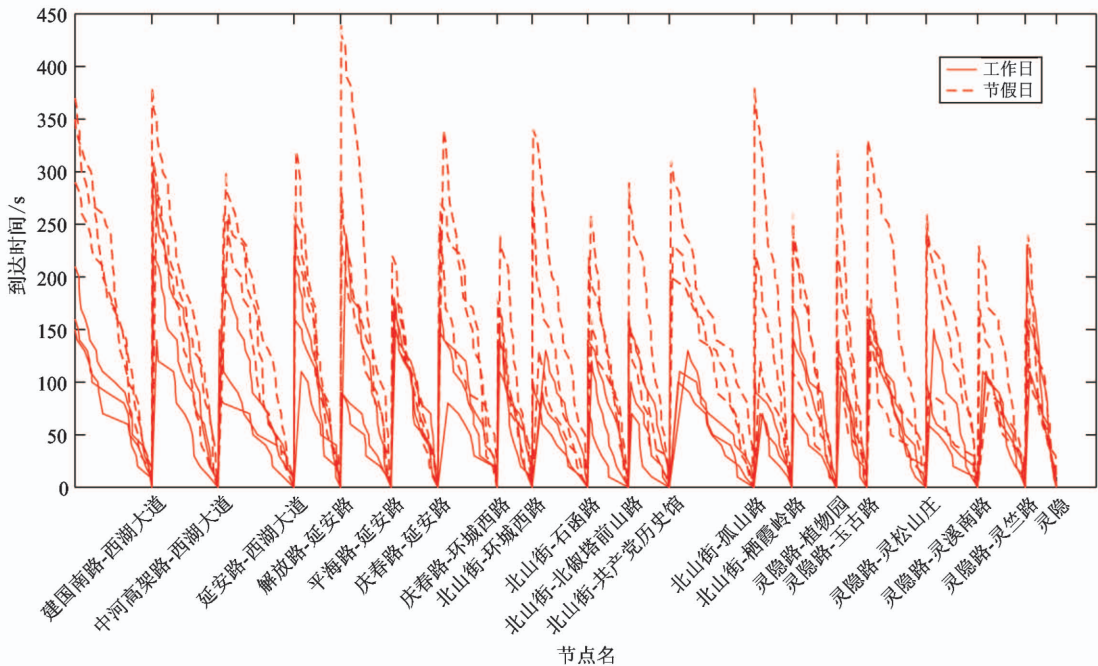


图 4 节假日、工作日公交车到达时间对比图

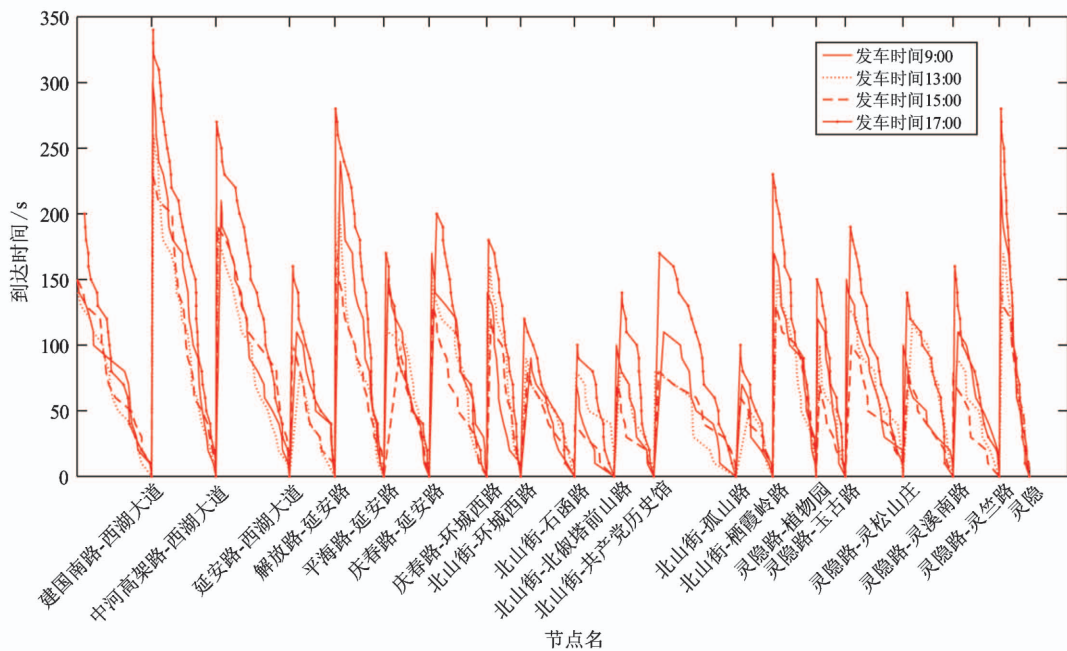


图 5 高峰、平峰公交车到达时间对比图

其中, $\rho_{\max}$ 、 $\rho_{\min}$ 分别为最大、最小学习率, η 为当前训练次数, $\eta_{\max}$ 为最大训练次数。

利用现实数据训练至其学习率收敛, 此时超参数取值为 $\xi = [62.34, 0.045, -0.0019]$ 。其他各个对比模型的参数均取利用本数据集训练过程中的最优取值。使用 7 种预测方法得到测试集到达时间的预测数据与实际数据进行对比, 部分测试集到达

时间预测对比图如图 6 所示, 图中横坐标为各个节点名, 各节点间距离按比例展示在横坐标中; 纵坐标表示对应的到达时间; 同时为了解 7 种模型的预测值相对于实际值的偏离程度, 取各个模型的残差值 (预测值与实际值的差) 进行对比, 如图 7 所示。

从两图中可以直观地看出, 相同数据样本下, 相比其他 3 种预测方法, DS-GPR 预测最贴近实际值,

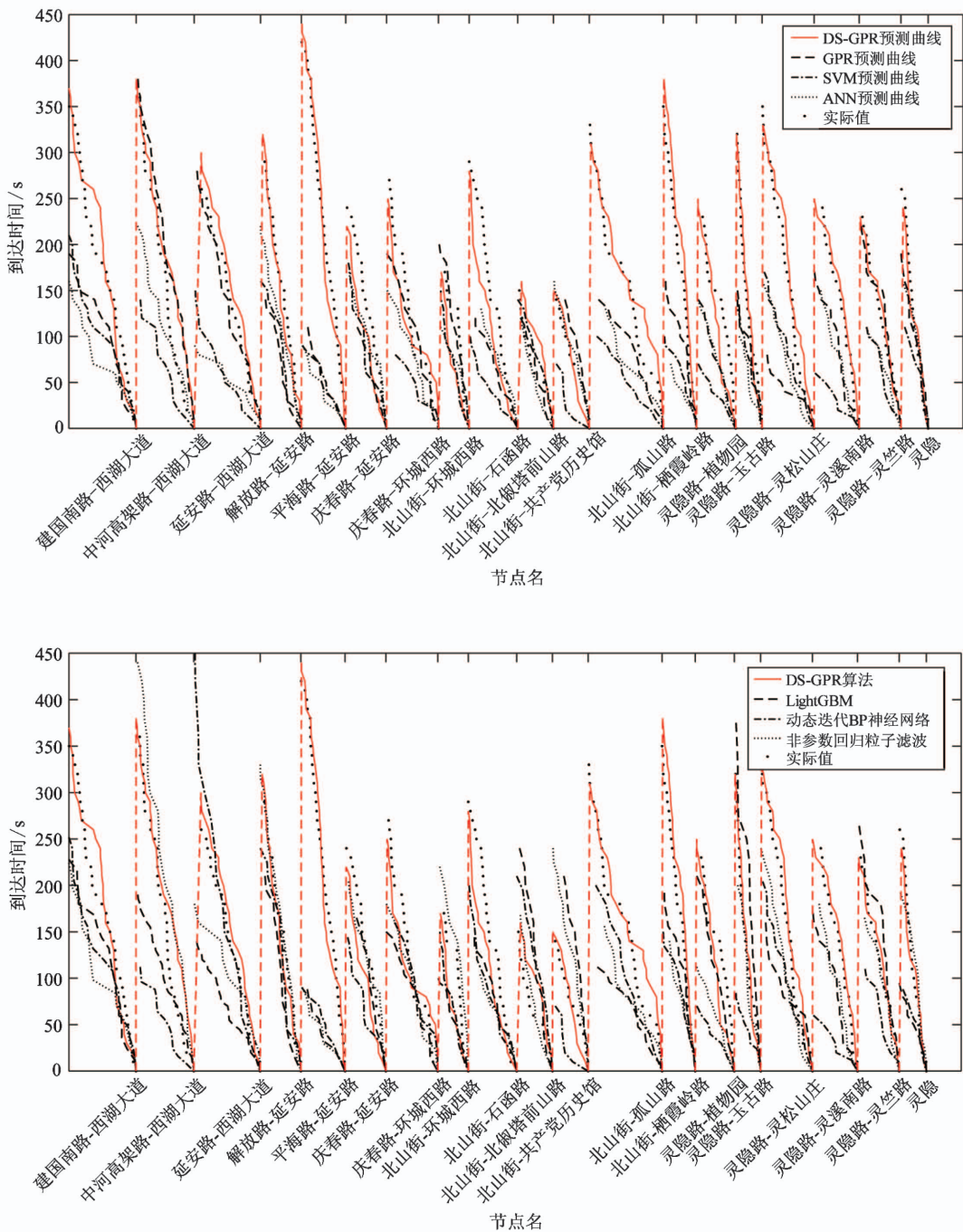


图 6 模型预测结果对比图

动态迭代反向传播(back propagation, BP)神经网络预测效果最不理想;与 DS-GPR 模型相比,若要相同预测效果,其他模型所需的训练时间更长,所需要样本量、样本数据种类更多。

为了比较 DS-GPR 与 GPR、SVM、人工神经网络(artificial neural network, ANN)、LightGBM、动态迭代 BP 神经网络以及非参数回归粒子滤波的预测性能^[20-22],选用评价指标分别对各模型预测结果进行

分析,利用以下 3 个指标评价预测结果。

(1)均方根误差(root mean square error, RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N [\hat{y}_i - y_i]^2}{N}} \tag{17}$$

在比较时,取精确度提高百分比 ρ 来计算:

$$\rho = \frac{(RMSE_{DS-GPR} - RMSE_{MODE})}{RMSE_{MODE}} \times 100\% \tag{18}$$

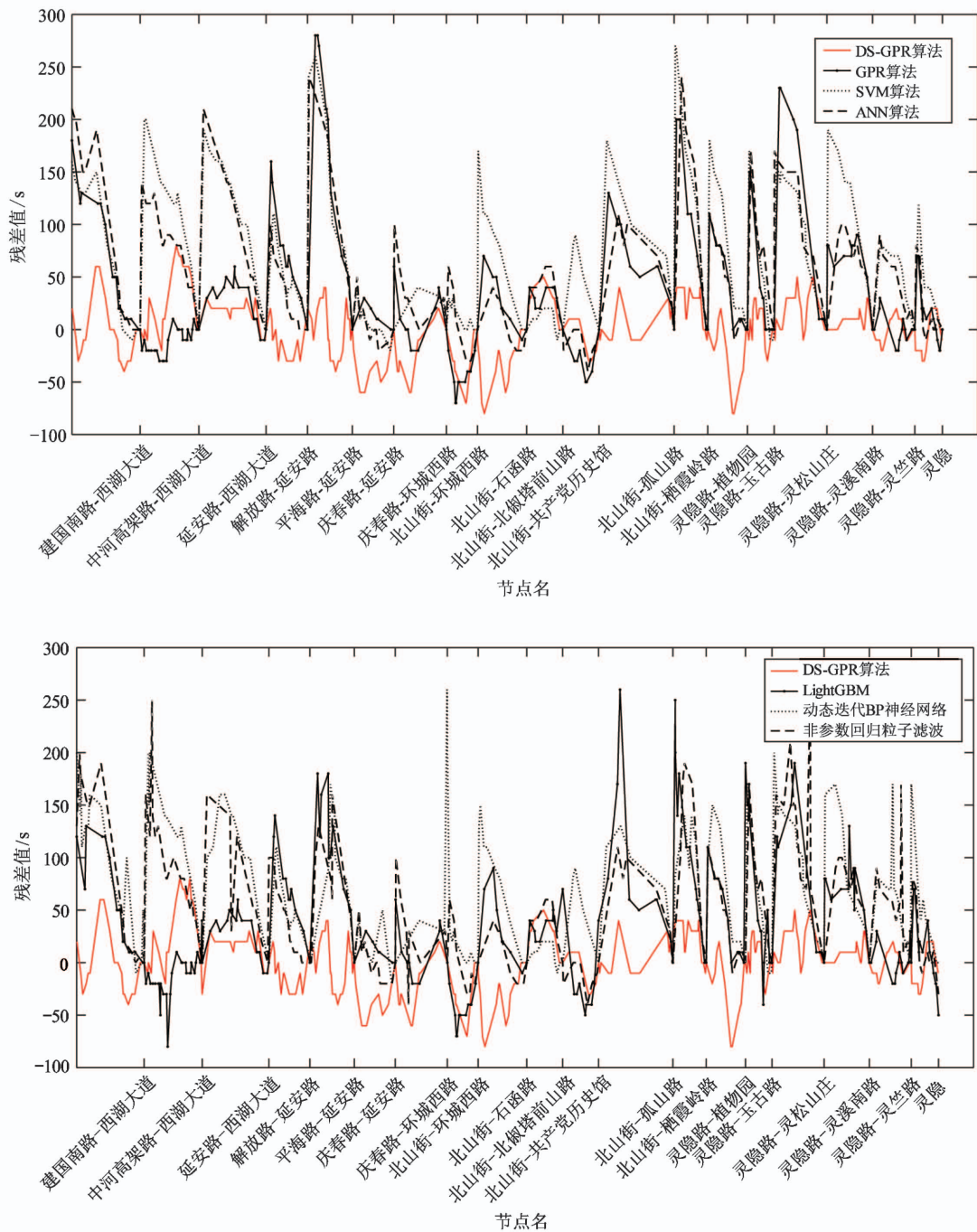


图 7 模型残差对比图

(2) 均方百分比误差 (mean square error, MSP)

$$MSP = \sqrt{\frac{\sum_{i=1}^N \left(\frac{\hat{y}_i - y_i}{y_i} \right)^2}{N}} \tag{19}$$

(3) 可决系数 R^2

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \tag{20}$$

式中, N 为预测数据量, y_i 为实际值, $\hat{y}_i$ 为预测值。 $RMSE$ 表示了预测总体的可靠性, 其值越大, 说明预测误差越大; 其值越小, 表明预测误差越小, 说明预测越可靠。 MSP 则表示了相对的平均偏离程度和可靠性程度。 R^2 表示预测值与实际值之间的相关性, 取值在 0 到 1 之间, 越接近 1 表示两者相关性越好, 即预测效果越好。

根据上述 3 个评价指标的公式计算出的评价结

果如表 2 所示,从表中可以看出本文方法普遍比经典方法预测准确性更好,DS-GPR 模型预测结果的各项评价指标均优于其他算法模型;*RMSE* 值最小说明其预测平均误差可靠程度最高;*MSP* 最小说明其预测结果平均偏离程度最低,即预测可靠性最高;*R²* 最接近 1,说明其预测值与实际值的拟合程度最好,预测结果最准确。综上所述,DS-GPR 模型的预测性能比其他对照模型的预测性能好,更适合共享公交专用道策略的使用。

表 2 不同预测方法预测性能评价对比

预测方法	<i>RMSE</i>	<i>MSP</i>	<i>R²</i>
DS-GPR	28.747436	0.238063	0.92301
GPR	65.20436	0.529945	0.591605
SVM	102.1524	0.560223	0.436933
ANN	99.11509	0.556606	0.582399
LightGBM	74.57218	0.326435	0.726705
动态迭代 BP 神经网络	101.7844	0.562903	0.635783
非参数回归粒子滤波	89.11509	0.353861	0.694569

3 结 论

基于车辆 OBD 的多源数据对公交车的到达时间预测进行研究,分析了时段因素和节假日因素对公交车到达时间的影响。利用证据理论优化了高斯过程回归的算法流程,建立了 DS-GPR 公交车到达时间预测模型。经过实例验证,与 3 种经典方法以及 3 种新方法进行对比分析,DS-GPR 模型的各项预测性能明显优于另外 6 种模型,证明了 DS-GPR 模型预测车辆到达时间的良好预测性能,很好地解决了以往时间预测模型中存在的泛化能力差、对突发情况预测不力的情况,并且该模型能满足预测实时性的要求,可基于此模型预测得到的公交车到达时间对共享公交专用道进行更深入的研究。

同时,DS-GPR 模型的训练速度和预测精度还可以通过调整核函数进一步提高,接下来可以从优化核函数类型、精简数据结构和完善参数权重等方面对模型进行进一步改进研究,使其预测性能更加优越。

参考文献

[1] Eichler M, Daganzo C F. Bus lanes with intermittent priority: strategy formulae and an evaluation[J]. *Transportation Research, Part B (Methodological)*, 2006, 40 (9):731-744

[2] Li C P, Han Y, Wang X Y. Traffic state estimation method of bus route in connected vehicle environment[J]. *Transport Research*, 2018, 4(5):29-34

[3] 董红召, 赵玉婷. 基于 PARAMICS 的公交专用道时分复用方法研究及优化[J]. *浙江工业大学学报*, 2012, 40(1):65-69

[4] 董红召, 凌越, 刘冬旭. 车路集成环境下 BRT 专用道时分复用的车路协作方法研究[J]. *中国工程科学*, 2012, 14(7):68-73

[5] Dong H Z, Zhao C X, Fu F J. Sharing bus lanes: a new lanes multiplexing-based method using a dynamic time slice policy[J]. *Proceedings of the ICE-Transport*, 2018: 1-38

[6] 赖永炫, 杨旭, 曹琦, 等. 一种基于 Gradient Boosting 的公交车运行时长预测方法[J]. *大数据*, 2019, 5 (5):58-78

[7] Marko Č, Lep M. Bus-arrival time prediction using bus network data model and time periods[J]. *Future Generation Computer Systems*, 2018, 110:164-171

[8] 李华民, 吴俊美, 孙棣华, 等. 基于 RFID 电子车牌数据的公交行程时间预测方法[J]. *中国公路学报*, 2019, 32(8):165-174

[9] Lin Y, Yang X, Zou N, et al. Real-time bus arrival time prediction: case study for Jinan, China[J]. *Journal of Transportation Engineering*, 2013, 139(11):1133-1140

[10] Liu Y, Ji Y, Chen K. Support vector regression for bus travel time prediction using wavelet transform[J]. *Journal of Harbin Institute of Technology (New Series)*, 2019 (3):26-34

[11] Maiti S, Pal A, Pal A, et al. Historical data based real time prediction of vehicle arrival time[C] // *IEEE International Conference on Intelligent Transportation Systems*, New York, USA, 2014:1837-1842

[12] 刘迎, 过秀成, 周润瑄. 基于多源数据融合的干线公交车辆行程时间预测[J]. *交通运输系统工程与信息*, 2019, 19(4):124-129, 148

[13] 张威威, 李瑞敏, 谢中教. 基于 PCA-GBDT 的城市道路旅行时间预测方法[J]. *公路工程*, 2017, 42(6):6-

11

[14] 范光鹏, 孙仁诚, 邵峰晶. 基于 LSTM 和 Kalman 滤波的公交车到站时间预测[J]. 计算机应用与软件, 2018, 35(4):91-96

[15] 华雪东, 阳建强, 王炜, 等. 基于数据融合的公交到站时间组合预测模型[J]. 公路交通科技, 2019, 36(2):112-120

[16] 詹随. 基于高斯过程回归的 GPS 失效车辆位置预测[D]. 长沙:湖南大学信息科学与工程学院, 2017:13-30

[17] 陆百川, 舒芹, 马广露. 基于多源交通数据融合的短时交通流预测[J]. 重庆交通大学学报(自然科学版), 2019, 38(5):17-23, 60

[18] 王力. 基于 DS 证据理论的多传感器数据融合算法研究与应用[D]. 太原:太原理工大学信息工程学院, 2015:45-51

[19] 安宝坤, 周欢欢. 基于混沌理论的短时公交到站时间预测[J]. 厦门大学学报(自然科学版), 2018, 57(4):565-571

[20] 王芳杰, 王福建, 王雨晨, 等. 基于 LightGBM 算法的公交行程时间预测[J]. 交通运输系统工程与信息, 2019, 19(2):116-122

[21] 刘靖, 肖冠烽. 基于 Spark 与粒子滤波算法的公交到站时间预测系统[J]. 计算机应用, 2019, 39(2):429-435

[22] 韩勇, 周林, 高鹏, 等. 基于 BP 神经网络的公交动态行程时间预测方法研究[J]. 中国海洋大学学报(自然科学版), 2020, 50(2):142-154

Regression prediction method of bus arrival time supported by OBD

Dong Hongzhao*, Zhao Longgang*, Zhao Chenxin*, Zhang Liang**, Kong Juanjuan**

(* Joint Institute of Intelligent Transportation System, Zhejiang University of Technology, Hangzhou 310014)

(** Zhejiang Jiake Electronics Co., Ltd., Jiaxing 314000)

Abstract

The accurate prediction of bus arrival time is the basis of the study of shared bus lanes. Based on the vehicle on board diagnostics (OBD) data, Gaussian process regression (DS-GPR) for bus arrival time prediction with evidence theory optimization is proposed. Firstly, the influencing factors of bus arrival time are analyzed, and then the hyper parameters of the model are obtained by historical data training. Then, the optimal output values are selected from the regression output distribution of Gaussian process with the global trust degree allocated according to D-S theory. Finally, the prediction performance of DS-GPR model is analyzed by comparing with other models. It is proved that DS-GPR model has good prediction performance for bus arrival time.

Key words: on board diagnostics (OBD), data fusion, Gaussian process regression (GPR), evidence theory, arrival time prediction