

基于高斯过程回归的公交到站预测方法

李香云,任 帅,张卫钢,吴娟娟,伍 菁

(长安大学 信息工程学院,陕西 西安 710064)

摘 要:在城市交通管理中,向市民提供更多的公共交通服务,是实现节能减排的有效途径。而研究公交车到站预测问题对提高公共交通服务水平有着重要的现实意义。目前,在公交车到站时间预测的研究中,大都是围绕到站时间精准性问题,而没有对预测结果的不确定性进行定量分析。因此,文中提出了一种基于高斯过程回归的公交到站预测方法,不仅可以对公交车到站时间进行精准预测,还可以根据预测值的方差来确定预测值95%的置信区间,即对不确定性进行研究。实验结果表明,该预测方法不仅与基于支持向量机的预测方法具有相近的预测精度,其中标准误差为13.39,平均绝对误差为12.91,平均绝对百分比误差为0.14,而且能够有效实现公交车到站时间概率意义上的预测。

关键词:城市交通;高斯过程回归;精准预测;置信区间;概率性预测

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2019)10-0021-05

doi:10.3969/j.issn.1673-629X.2019.10.005

A Bus-to-station Prediction Method Based on Gaussian Process Regression

LI Xiang-yun, REN Shuai, ZHANG Wei-gang, WU Juan-juan, WU Jing

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: In urban traffic management, providing more public transportation services to the public is an effective way to achieve energy conservation and emission reduction. The research on bus arrival prediction is of important practical significance for improving the public transportation service level. At present, most of the research on the arrival time prediction of buses is based on the accuracy of the arrival time, but there is no quantitative analysis of the uncertainty of the prediction results. Therefore, we propose a bus-to-station prediction method based on Gaussian process regression, which can not only accurately predict the arrival time of the bus, but also determine the 95% confidence interval of the predicted value based on the variance of the predicted value, that is, the uncertainty is studied. The experiment shows that the prediction method proposed not only has similar prediction accuracy with the prediction method based on support vector machine, in which the root mean square error is 13.39, the mean absolute error is 12.91, and the mean absolute percentage error is 0.14, but also can effectively realize the prediction of the probability of bus arrival time.

Key words: urban traffic; Gaussian process regression; precise prediction; confidence interval; probabilistic prediction

0 引言

目前,在智能公共交通系统领域,国内外学者对公交车到站时间的精准预测做了大量研究,常用且高效的模型主要有以下几种:

第一种是基于历史数据的预测模型。文献[1-2]将要预测的路段分割成若干段,利用历史行驶数据分别计算各子路段的平均行驶时间,求得最终的预测时间;文献[3]在此基础上引入有限状态机完善模型;文献[4]创新地利用GPS数据和路段在空间和时空分布

上的特点,根据历史路段平均到站总时间与当前车辆位置与速度,预测到达下站所需要的时间;文献[5]采用逆向查找法,将瞬时速度与历史平均速度进行融合,利用粒子滤波算法预测公交车到达下站的时间;文献[6]利用交通流的时间变化规律获取交通数据周期性和局部变化的特征,建立了时间序列模型,预测出公交车到站时间。这种模型的预测精度很大程度上取决于采集的历史数据的准确性。

第二种是基于人工神经网络的预测模型。文献

收稿日期:2018-11-18

修回日期:2019-02-21

网络出版时间:2019-04-24

基金项目:国家自然科学基金(61702050,61402052)

作者简介:李香云(1995-),女,硕士研究生,研究方向为数据挖掘、机器学习;任 帅,博士,副教授,研究方向为信息隐藏理论与模型;张卫钢,博士,教授,研究方向为计算机应用技术、智能测控技术、汽车电子技术。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1100.082.html>

[6-7]将行驶距离、时段等作为影响因素,使用神经网络算法进行预测;文献[8]结合调度经验,对神经网络算法进行优化,最终建立了智能调度模型;文献[9]建立了基于小波神经网络的预测模型,改进了神经网络易引起振荡效应的缺点,并采用粒子群算法进行优化,有效避免了其陷入局部最优。

第三种是基于动态的预测模型。文献[10]利用 SVM 从历史数据中预测的时间作为矩阵输入 Kalman 滤波器,利用“更新方程”将最新的观测值加入到预测向量中,有效地提高了预测精度。但由于公交车到站时间是具有长期和短期特性的时间序列数据,因此文献[11]利用具有长短记忆递归神经网络 LSTM 作为静态预测模型,然后利用 Kalman 滤波作为动态模型对预测结果进行调整。

上述研究都是对公交车到站时间的精准预测,但由于客流量、车本身性能、交通拥堵等因素的影响,增加了公交车到站时间的不确定性。面对各种突发情况,市民选择公交出行时,对公交到站时间的区间估计尤为关注,因此有必要对预测值的置信区间进行研究。高斯过程回归(Gaussian processes regression, GPR)是一种基于统计学习理论和贝叶斯理论的非参数^[12]机器学习方法,文中探索性地建立了基于 GPR 的公交车到站时间预测模型(GPR bus arrival time prediction, GPR-BATP),在对公交车到站时间进行精准预测的同时,得到预测值的方差估计值,然后由预测值减去标准差作为下限,预测值加上标准差作为上限,来确定预测值 95% 置信区间,实现对到站时间概率意义上的预测。

1 高斯过程回归模型建立

1.1 高斯过程

高斯过程(Gaussian processes, GP)是一个正态的随机过程,其任意维有限变量的联合分布服从高斯分布。

对于任意有限个 x , 即 $x_1, x_2, \dots, x_n \in N$, 其相对应的有限个随机变量 $(f(x_1), f(x_2), \dots, f(x_n))^T$ 均服从式 1 所示的概率分布, 那么 $F = (f(x_1), f(x_2), \dots, f(x_n))^T$ 为 GP, 则可记作 $F \sim \text{GP}(m(\cdot), k(\cdot, \cdot))$ 。

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right) \quad (1)$$

其中, F 为联合正态分布; $m(\cdot) = (m(x_1), m(x_2), \dots, m(x_n))^T$ 表示均值向量, $m(x) = E(x)$ 表示均值函数; $k(\cdot, \cdot) = (k(x, x'))_{n \times n}$ 表示协方差矩阵。

1.2 GPR 模型建立

将高斯过程应用于回归问题进行建模时,考虑到目标值 y 含有噪声^[13], 即定义为:

$$y = f(x) + \varepsilon \quad (2)$$

其中, $x \in N$, $f(x)$ 为高斯过程; ε 为服从均值为 0、方差为 δ_n^2 的高斯白噪声。

GPR 的关键假设是所有输入的观测值 Y 以及预测值 y^* 都服从联合正态分布, 即给定训练集 $\text{Train} = \{(x_i, y_i)\}, i = 1, 2, \dots, n$, $Y = (y_1, y_2, \dots, y_n)^T$ 服从的先验分布为:

$$Y \sim N(0, k(\cdot, \cdot) + \delta_n^2 I_n) \quad (3)$$

其中, $\cdot = (x_1, x_2, \dots, x_n)^T$; I_n 为 $n \times n$ 阶单位矩阵。

如果给定测试集 $\text{Test} = (x^*, y^*)$, 那么:

$$\begin{bmatrix} Y \\ y^* \end{bmatrix} \sim N \left(0, \begin{bmatrix} k(\cdot, \cdot) + \delta_n^2 I_n & K_* \\ K_*^T & K_{**} \end{bmatrix} \right) \quad (4)$$

其中, $k(\cdot, \cdot)$ 为训练集 X 之间的 $n \times n$ 阶对称的正定协方差矩阵; $K_* = K(X, x^*)$ 为训练集输入 X 与测试集 x^* 的 $n \times 1$ 阶的协方差矩阵; $K_{**} = k(x^*, x^*)$ 为测试点 x^* 自身的协方差。

接下来, 依据贝叶斯方法和联合正态分布理论, 可求得预测值的后验分布为:

$$y^* | X, Y, x^* \sim N(\mu(x^*), \text{var}(x^*)) \quad (5)$$

其中

$$\mu(x^*) = K_*^T (k(\cdot, \cdot) + \delta_n^2 I_n)^{-1} Y \quad (6)$$

$$\text{var}(x^*) = K_{**} + \delta_n^2 + K_*^T (k(\cdot, \cdot) + \delta_n^2 I_n)^{-1} K_* \quad (7)$$

其中, $\mu(x^*)$ 为观测点 x^* 的预测均值; $\text{var}(x^*)$ 为对应的方差。

文中选择平方指数协方差函数(squared exponential covariance function, SE)(见式 8)作为 GPR 的核函数。

$$k(x, x') = \delta_f^2 \exp \left(-\frac{1}{2} (x - x')^T M^{-1} (x - x') \right) \quad (8)$$

其中, δ_f^2 为信号方差; $M = \text{diag}(l^2)$, l 为特征长度尺度参数。把 δ_f^2 , l 与上述的加性噪声的方差 δ_n^2 的集合记作 θ , 则 $\theta = \{\delta_f^2, l, \delta_n^2\}$, 因而得到 GPR 模型的超参数为集合 θ 。

根据贝叶斯理论和最大后验概率估计原理, 转化为求训练样本条件概率的对数似然函数 $L(\theta)$, 即:

$$L(\theta) = \log(p(\mathbf{Y}|\mathbf{X},\theta)) =$$
$$- \left(\frac{1}{2} \mathbf{Y}^T (\mathbf{k}(\cdot, \cdot) + \delta_n^2 \mathbf{I}_n)^{-1} \mathbf{Y} + \right.$$
$$\left. \frac{1}{2} \log |\mathbf{k}(\cdot, \cdot) + \delta_n^2 \mathbf{I}_n| + \frac{n}{2} \log 2\pi \right)$$

(9)

上式对 θ 求偏导,可得:

$$\frac{\partial L(\theta)}{\partial \theta_i} = \frac{1}{2} \text{trace}((\boldsymbol{\gamma} \boldsymbol{\gamma}^T - \mathbf{k}(\cdot, \cdot)^{-1}) \frac{\partial \mathbf{k}(\cdot, \cdot)}{\partial \theta_i})$$

(10)

则可得超参数集合 θ 。为了泛化,将均值设为 0;

$$\boldsymbol{\gamma} = \mathbf{k}(\cdot, \cdot)^{-1} \mathbf{Y}。$$

表 1 GPS 数据示例

线路 ID	车辆 ID	时间戳	经度	纬度	下一站编号	路口总数	瞬时速度
902	905273	07:00:03	117.187 99	39.106 3	16	26	0

文中选择从 10 月 9 日至 10 月 20 日之间工作日的 902 线路公交车的 26 377 条 GPS 记录为训练集,选择 10 月 24 日的 999 条 GPS 记录作为测试集。仅对工作日进行讨论。

2.2 数据处理

2.2.1 轨迹数据预处理

GPS 设备在采集数据过程中易受建筑物遮挡、设备故障等因素的影响,使得采集设备在部分时段缺乏位置信息,导致采集的数据存在丢失、异常等问题^[15]。此时,若利用有异常的轨迹数据进行预测,会对预测的精度造成重大影响。为了降低噪声数据对预测模型性能的影响,首先使用阈值过滤法清洗数据;再判断数据是否存在缺失或异常;最后利用前后相邻轨

根据式 10 求得的偏导数,采用共轭梯度法或牛顿法(文献[14]中有详细论述)求得式 9 的最大似然函数和最优超参数 θ 。在进行预测时,根据得到的 GPR 模型,利用式 6 和式 7 式即可得到预测点的预测均值和方差估计值。

2 高斯过程回归的应用

2.1 数据描述

数据来源于 2017 年 10 月 1 日至 10 月 24 日天津市公交车的 GPS 记录。每条记录包含线路 ID、车辆 ID 等。数据示例如表 1 所示。

迹点经纬度的平均值对缺失或异常数据进行修正。

2.2.2 数据整合

整合过程为:根据某条记录的下站编号与前一条记录的下站编号不同,与下一条记录的下站编号相同,判断此记录是否为停靠站,是则标记为 1,否则,标记为 0。然后根据某条记录的下站编号和是否为停靠站点来填充此条记录的下站站点的经纬度以及下站站点的时间戳。

整合之后,原始数据增加四列:是否为停靠站点、下站站点的经纬度以及下站站点的时间戳。

2.3 特征工程

2.3.1 路段距离

构造路段距离特征,计算方法为:

$$\text{Distance} = 6\ 731 * 2 * \text{asin}(\sqrt{[\sin(\frac{\text{lat}_2 - \text{lat}_1}{2})]^2 + \cos(\text{lat}_1) * \cos(\text{lat}_2) * [\sin(\frac{\text{lon}_2 - \text{lon}_1}{2})]^2})$$

(11)

其中, $\text{lon}_1, \text{lat}_1, \text{lon}_2, \text{lat}_2$ 分别为当前记录和下一站站点的经纬度弧度数据;6 371 表示地球半径(单位 km)。

2.3.2 路段行程时间

根据整合后的数据,由当前时间戳和下一站站点的时间戳,求得当前位置到达下一站点的行程时间。

2.4 数据分析

以行车路段的距离、路口数两个影响因素为例进行讨论。

2.4.1 路段距离

将两个相邻的站点定义为一个行程区间,902 线路的 24 个区间的各区间距离以及各区间行程时间如图 1 所示。

可以看出,行程区间距离的长短与行程时间的大

小大致成正比关系,因此,根据行车距离预测行车时间具有实用价值。

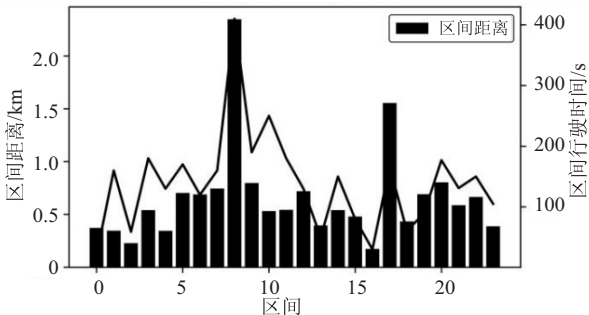


图 1 区间距离与区间行驶时间对比

2.4.2 交通路口数

统计经过不同路口数时所需的行驶时间,如图 2 所示。

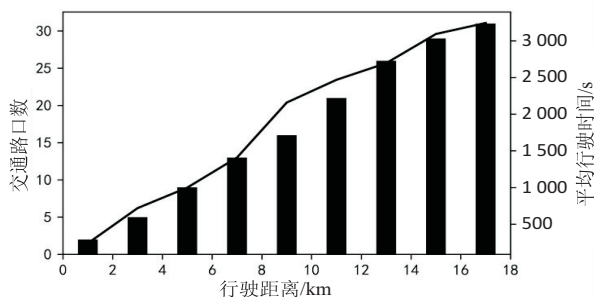


图 2 不同距离不同交通路口数的行驶时间

结果显示公交车的平均行驶时间会随路口数及行驶距离的增加而增加。

2.5 GPR-BATP 模型

GPR-BATP 如下:

$$Y_{i+1} = f(X_i) + \varepsilon \quad (12)$$

其中, Y_{i+1} 表示公交车到站时间的预测值; X_i 表示与 Y_{i+1} 相关的影响因子, $f(\cdot)$ 为预测模型; ε 为车本身性能、交通事故等未统计噪声。

对原始数据集按照 2.2 和 2.3 小节进行处理,得到如 1.2 小节中相同形式的训练集 Train, 定义特征向量 $X_i = (x_i^1, x_i^2, x_i^3, x_i^4, x_i^5, x_i^6, x_i^7, x_i^8, x_i^9)$, x_i^1, x_i^2 分别表示当前 GPS 记录的经纬度; x_i^3 表示当前时间戳; x_i^4 表示路口数; x_i^5 表示下一站编号; x_i^6, x_i^7 分别表示下一站站点的经纬度; x_i^8 表示下一站站点的的时间戳; x_i^9 表示是否为停靠站点。 y_i 表示从当前时间戳到下一站站点的的时间间隔。

按照上述方法,通过原始数据集训练基于 GPR 的路段行程时间预测方法如下所述:

输入:原始数据集;

输出:GPR 预测模型。

Step1:对原始数据集按照 2.2 和 2.3 小节的方法进行处理,得到训练集 Train。

Step2:

Step2.1:由特征向量 $x_i^1, x_i^2, x_i^6, x_i^7$ 按照式 11 计算行程距离,并进行归一化;对特征向量 x_i^4 进行归一化

处理。

Step2.2:基于训练集根据 1.2 小节的方法,按照式 9、式 10 计算协方差函数的最佳超参数集合 θ 。

Step3:确定协方差函数 $k(x, x')$ 。

Step4:得到 GPR-BATP。

对于预测的数据集,处理方法同原始数据集,得到测试集 Test。然后根据得到的 GPR 模型,对 Test 特征向量 X 进行预测。预测过程如下所述:

输入:待预测的数据集;

输出:公交车到达待预测站点的精准时刻 T , 预测值 95% 的置信区间 $[T - \text{std}, T + \text{std}]$, MAPE, RMSE, MAE。

Step1:对待预测的数据集按照 2.2、2.3 小节的方法进行处理,得到训练集 Test。

Step2:

while:

输入新的特征向量 X , 由特征向量 $x_i^1, x_i^2, x_i^6, x_i^7$ 按照式 11 计算距离,并进行归一化;对特征向量 x_i^4 进行归一化处理

按照式 8, 计算 K_* , K_*^T , K_{**}

计算预测值 $\mu(x^*)$, $\text{var}(x^*)$

end while

对预测值 $\text{var}(x^*)$ 进行反归一化得到预测结果 y_{pre} , 根据 y_{pre} 和当前时间戳 x_i^4 得到将要到达的站点时刻 T , 由 $\text{var}(x^*)$ 得到 std (偏差), 得到置信区间为 95% 的到站时间 $[T - \text{std}, T + \text{std}]$

Step3: 根据 y_{pre} 和 y 计算得到 MAPE, RMSE, MAE。

3 仿真实验

3.1 实验结果对比

分别采用 GPR-BATP 和 SVM 对 902 线路公交车 24 个运行区间的行程时间进行预测, 结果如图 3 所示。

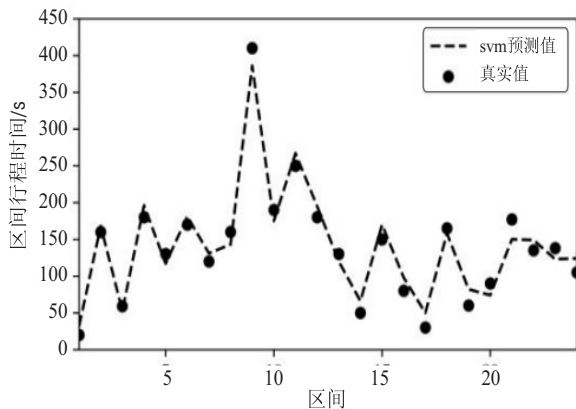
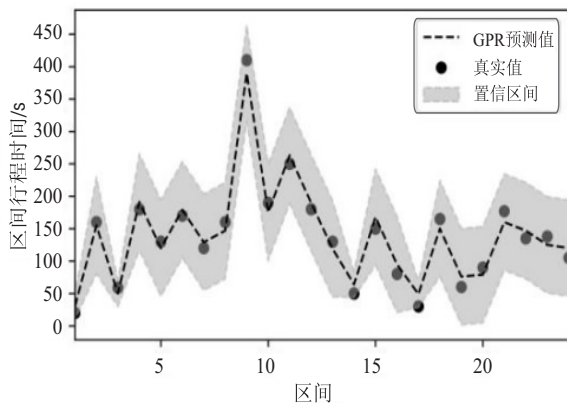


图 3 GPR、SVM 到站时间预测结果

3.2 预测方法性能对比

算法质量的衡量标准采用平均绝对百分比误差 (mean absolute percentage error, MAPE)、标准误差 (root mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE):

MAPE = \frac{1}{2} \sum_{i=1}^m \frac{|y_{true} - y_{pre}|}{y_{true}} \tag{13}

RMSE = \sqrt{\frac{\sum_{i=1}^m (y_{true} - y_{pre})^2}{m}} \tag{14}

MAE = \frac{\sum_{i=1}^m |y_{true} - y_{pre}|}{m} \tag{15}

其中, m 为数据样本集的样本数量; y_{pre} 为预测值; y_{true} 为真实值。

分别采用 GPR 和 SVM 公交车到站时间预测方法获得的预测性能指标如表 2 所示。

表 2 两种方法的预测性能指标

方法	MAPE	MAE	RMSE
GPR	0.14	12.91	13.39
SVM	0.17	15.35	16.19

可见,无论是 MAPE、RMSE 还是 MAE,GPR 模型可以获得与 SVM 相近的预测性能。相对于 SVM,GPR-BATP 具有如下优势^[16]:

(1)最优超参数求取过程的系统化。SVM 中参数求取只能通过交叉验证进行,而 GPR 预测模型利用贝叶斯理论通过最大后验似然估计方法来求取最优超参数集合。

(2)可以得到预测结果的误差带。在实现精准预测的同时,能够确定预测结果的置信区间,因此可以有效地对预测值的可信度进行把握。

4 结束语

文中提出一种基于 GPR 的公交车到站时间预测方法。实验结果表明,该方法不仅与 SVM 方法具有相近的预测精度,还能确定预测结果的 95% 置信区间,从而可以从概率意义上对到站时间实现准确预测,具有较高的实用价值和理论参考意义。

相对于 SVM 方法,该方法需要对实验数据进行归一化处理,这会影响实际预测过程的实时性。因此,如何提高 GPR 法的计算效率将是下一步的研究重点。另外,市民的出行规律(分为高峰期、平峰期、低峰期),也是影响到站时间的要素,也需要进一步研究。

参考文献:

[1] ZHANG M,XIAO F,CHEN D. Bus arrival time prediction based on GPS data[C]//Fourth international conference on transportation engineering. Chengdu, China; [s. n.], 2013: 1470-1475.

[2] CHEN G,YANG X,ZHANG D,et al. Historical travel time based bus-arrival-time prediction model[C]//American society of civil engineers 11th international conference of Chinese transportation professionals. Nanjing, China; [s. n.], 2011:1493-1504.

[3] SUN Dihua,LUO Hong,FU Liping,et al. Predicting bus arrival time on the basis of global positioning system data[J]. Transportation Research Record,2007,2034(1):62-72.

[4] 段颖超,张健钦,李明轩,等. 一种公交到站时间预测方法[J]. 测绘通报,2016(5):50-53.

[5] 任 远,吕永波,马继辉,等. 基于粒子滤波的公交车到站时间预测研究[J]. 交通运输系统工程与信息,2016,16(6):142-146.

[6] CHIEN I J,DING Y,WEI C. Dynamic bus arrival time prediction with artificial neural networks[J]. Journal of Transportation Engineering,2014,128(5):429-438.

[7] PAN Jian,DAI Xiuting,XU Xiaoqi,et al. A self-learning algorithm for predicting bus arrival time based on historical data model[C]//International conference on cloud computing & intelligent systems. Hangzhou, China; IEEE, 2012: 1112-1116.

[8] 陈 鹏. 基于 BP 神经网络的公交智能实时调度模型研究及系统实现[D]. 北京:北京交通大学,2008.

[9] 季彦婕,陆佳炜,陈晓实,等. 基于粒子群小波神经网络的公交到站时间预测[J]. 交通运输系统工程与信息,2016,16(3):60-66.

[10] 柏 丛,彭仲仁. 基于动态模型的公交车行程时间预测[J]. 计算机工程与应用,2016,52(3):103-107.

[11] 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.

[12] CARL E R,CHRISTOPHER K I W. Gaussian processes for machine learning[M]. Cambridge,MA:MIT Press,2006.

[13] 李 军,张友鹏. 基于高斯过程的混沌时间序列单步与多步预测[J]. 物理学报,2011,60(7):143-152.

[14] RASMUSSEN C E,WILLIAMS C K. Gaussian processes for machine learning[M]. Cambridge,MA:MIT Press,2006.

[15] 毛嘉莉,金澈清,章志刚,等. 轨迹大数据异常检测:研究进展及系统框架[J]. 软件学报,2017,28(1):17-34.

[16] 康 军,段宗涛,唐 蕾,等. 高斯过程回归短时交通流预测方法[J]. 交通运输系统工程与信息,2015,15(4):51-56.