刘佳俊, 唐阳山. 基于 MATLAB 的交通拥堵状态预测研究[J]. 智能计算机与应用,2024,14(8):199-204. DOI:10.20169/j.issn.2095-2163.240931

基于 MATLAB 的交通拥堵状态预测研究

刘佳俊, 唐阳山
(辽宁工业大学 汽车与交通工程学院, 辽宁 锦州 121001)

摘 要: 交通拥堵可分为常发性交通拥堵和偶发性交通拥堵,本文针对常发性交通拥堵,以 TTI 为指标划分城市路段的交通状况等级,并通过采集到的行驶速度数据集转化为 TTI 数据集,并将其划分为训练集和测试集。基于 MATLAB 软件设计出随机森林预测模型、LSTM 预测模型,对算法进行了描述和优缺点对比分析,并将二者模型应用于 TTI 数据集。再次通过 RMSE、MSE、MAE 和 R^2 等指标对设计的预测模型的性能进行对比评估,选取性能最好、拟合优度最高的预测模型为 LSTM。为了验证 LSTM 模型是否产生过拟合,采用公开数据集再次做了对比实验分析,结果表明没有产生明显的过拟合。最后针对 LSTM 模型在 TTI 数据集上预测效果不够理想,分析出 TTI 数据集不够丰富,数据量不多等原因,并提出多路段采集等改善方案,以优化 LSTM 模型的拟合效果。

关键词: 交通拥堵; TTI; MATLAB; 随机森林; LSTM
中图分类号: TP181; U 491.265 **文献标志码**: A **文章编号**: 2095-2163(2024)09-0199-06

Research on traffic congestion state prediction based on MATLAB

LIU Jiajun, TANG Yangshan
(School of Automotive and Transportation Engineering, Liaoning University of Technology, Jinzhou 121001, Liaoning, China)

Abstract: Traffic congestion can be divided into frequent traffic congestion and occasional traffic congestion. This article focuses on frequent traffic congestion and uses TTI as an indicator to classify the traffic conditions of urban road sections. The collected driving speed dataset is converted into TTI dataset and divided into training and testing sets. A random forest prediction model and an LSTM prediction model were designed based on MATLAB software. The algorithms were described and their advantages and disadvantages were compared and analyzed. The two models were applied to the TTI dataset. Once again, the performance of the designed prediction model was compared and evaluated using metrics such as RMSE, MSE, MAE, and R^2. LSTM was selected as the prediction model with the best performance and highest fitting degree. To verify whether the LSTM model produces overfitting, a comparative experimental analysis was conducted again using a publicly available dataset, and the results showed no significant overfitting. Finally, in response to the unsatisfactory prediction performance of the LSTM model on the TTI dataset, it was analyzed that the TTI dataset was not rich enough and had a small amount of data. Improvement plans such as multi segment data collection were proposed to optimize the fitting effect of the LSTM model.

Key words: traffic congestion; TTI; MATLAB; random forest; LSTM

0 引 言

随着中国经济的高速发展和城市化建设进程的不断加快,城市机动车的数量随时间呈快速增加的趋势。机动车数量的增加一方面确实能为城市居民的出行提供便利,另一方面却也是导致城市交通拥堵的重要原因[1]。交通拥堵是一种车多拥挤且车速缓慢的现象,其根本原因是交通需求超过了道路的最大通行能力。根据拥堵的持续时间和成因,可将交通拥堵分为常发性交通拥堵和偶发性交通拥堵。常发性交通拥堵由固定的交通模式引起,如上下班高峰时段的拥堵;偶发性交通拥堵则是由随机事件引起,如交通事故、恶劣天气或道路维修等。

国内外学者自 20 世纪 50 年代起,就对交通拥

基金项目: 辽宁省教育厅 2023 基本科研项目(JYTMS20230842)。
作者简介: 刘佳俊(2000-),男,硕士研究生,主要研究方向:环境感知与汽车智能驾驶,Email:2640439363@qq.com; 唐阳山(1972-),男,博士,教授,硕士生导师,主要研究方向:交通信息工程及控制。
收稿日期: 2024-05-12

堵状态进行了系统研究。魏丹[2]提出了一种基于概率神经网络的道路交通状态实时判别方法,并选择交通流量、车速和大车比例作为判别指标。实验证明,该方法在城市道路交通状态判别中可行有效。褚瑞娟[3]采用无监督学习 FCM 聚类方法和建立基于 BPSO-(RSM-DAG-SVM$_S$)选择性集成的交通状态判别模型,对高速公路交通运行状态进行判别与预测,并通过实测验证了该模型的有效性。易术等[4]提出了一种基于多源数据结合元胞传输模型(Multi-Source Data Cell Transmission Model, MD-CTM)的交通状态预估方法,以满足针对高速公路精准、可靠地交通状态估计。该方法能灵活调整元胞的长度和数量,弥补了传统 CTM 模型要求元胞长度必须一致的缺陷,并选取了成都市绕城高速路段的实际数据集进行仿真训练,最后验证了该方法的有效性和准确性。田润泽[5]提出了基于 SVR-LightGBM 模型的高速公路交通拥堵预测方法,并通过实验数据验证了该方法的有效性和实用性。陈悦[6]通过在 CNN 处理空间信息层和 LSTM 处理时间信息层中间添加 Softmax 层,并改进 C-BiLSTM 预测框架,提出了一种 CS-BiLSTM 网络预测框架,提高了空间信息特征提取的准确性,进而提高了短时交通拥堵状态预测模型的准确性。

城市交通拥堵不仅会浪费大量资源,还会污染环境且影响交通参与者心态。智能交通系统(Intelligent Transport System, ITS)作为一种大范围,全方位覆盖、实时、准确、高效的综合运输和管理系统[7],将物联网技术与先进的控制、传感、通讯、信息技术与计算机技术高效结合,综合应用于整个交通管理体系。由于其能极大地缓解城市交通拥堵问题,有效改善交通状况,因此受到越来越多城市的重视,并在许多国家和地区应用与发展[8]。ITS 中比较重要的环节就是不断地对车流量进行准确预测,对可能发生的拥堵事件做出提前预判。因此,本文基于 MATLAB,并设计包括随机森林、LSTM 等机器学习算法对 TTI 进行数值预测,以实现对路段交通状况的预测,采用均方根误差和均方误差评价模型对比分析,以选取最佳算法模型。

1 交通运行状况分析

行程时间比(Travel Time Index, *TTI*)是指机动车通过某路段的实际行驶时间与自由状态下行驶时间的比值(该路段自由流速度与通过该路段平均速度的比值)。*TTI* 可以反映交通运行状况,且 *TTI* 数值越大,该路段越拥堵[9]。因此,本文主要通过行程时间比对城市路段交通运行状况进行等级划分。

2016 年交通运输部公布的《城市交通运行状况评价规范》,依据速度可划分路段交通运行状况等级;再依据行程时间比的概念,可计算得到路段交通各运行状态等级下 *TTI* 的取值范围,等级划分见表 1。

表 1 以 *TTI* 为指标的路段交通状况等级划分表

Table 1 Classification table of traffic conditions on road sections based on *TTI* as an indicator

运行状态等级	畅通	基本畅通	轻度畅通	中度畅通	严重拥堵
取值范围	$1 \leqslant TTI < 1.43$	$1.43 \leqslant TTI < 2$	$2 \leqslant TTI < 2.5$	$2.5 \leqslant TTI < 3.33$	$TTI \geqslant 3.33$

2 数值预测模型

MATLAB 作为一个强大的机器学习实现工具,即可设计出复杂的机器学习模型,并快速实现训练,因此可应用于数值预测领域。

2.1 随机森林模型

随机森林(Random Forest)是一种经典的集成学习(Bagging)模型,由多个决策树组成,每个决策树都是通过在不同样本和特征子集上进行训练得到的,能有效避免过拟合问题。其通过投票或平均等方式整合每个决策树的预测结果,从而得到最终的预测结果。该模型具有较高的准确性和稳定性[10],其实现的关键步骤如下:

(1)对数据集进行预处理后,在数据集中随机选择一定数量的样本,基于此建构一颗决策树。

(2)在每个节点上,随机选择一部分特征子集,并基于其特征子集进行最优划分。最小子节点数量可通过设定参数进行控制。

(3)重复上述步骤,构建多棵决策树。生成的决策树数量可通过设定参数进行控制。

(4)通过投票或平均等方式,对每棵决策树的预测结果进行综合,得到最终的预测结果。对于分类问题,采用投票的方式,即多数决定。对于回归问题,对所有决策树的预测结果取平均值[11]。本文采

用平均的方法。

2.2　LSTM 模型

长短期记忆神经网络[12]（Long Short Term Memory，LSTM）是一种特殊的递归神经网络（RNN），被设计用来处理和预测时间序列数据。LSTM 通过其特殊的结构和内部机制，如记忆单元（Cell State）、状态单元（Hidden State）、输入门（Input Gate）、输出门（Output Gate）和遗忘门（Forget Gate），有效地解决了传统神经网络存在的梯度爆炸和长期依赖问题[13]。LSTM 的基本工作原理如图 1 所示。

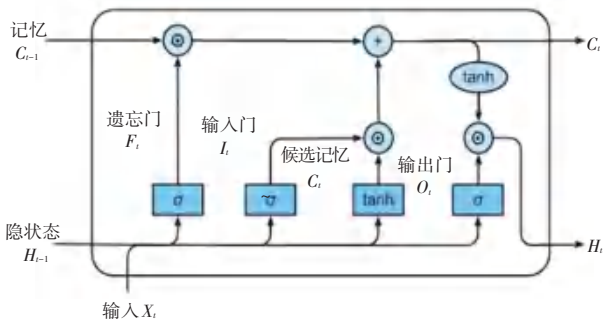

图 1　LSTM 网络结构图

Fig. 1　LSTM network architecture diagram

（1）遗忘门

从图 1 可以看出，记忆单元首先进行的操作是遗忘部分历史信息，该操作通过遗忘门实现。而遗忘门接受的输入包括 C_{t-1}、X_t 和 H_{t-1}，并通过下式在遗忘门实现部分历史信息的遗忘。

$$F_t \times C_{t-1} = sigmoid(W_{x_f} \times X_t + W_{h_f} \times H_{t-1} + b_f) \times C_{t-1} \tag{1}$$

其中，W_{xf} 和 W_{h_f} 表示输入数据和状态单元到遗忘门的权重参数，该参数需要学习得到；H_{t-1} 表示 $t-1$ 时刻隐藏层的状态；X_t 表示 t 时刻的输入数据；b_f 表示偏置常数；C_{t-1} 表示 $t-1$ 时刻的记忆单元；*sigmoid* 函数能够将任何输入值映射到 0 ~ 1 之间。

（2）输入门

记忆单元的第二步更新是增加当前时刻的信息，因此需要定义一个候选记忆，用于暂时存储当前时刻的信息，其次定义 I_t，用于保留当前时刻的信息。而输入门就是将 t 时刻的重要信息输入至记忆细胞，并通过下式实现。

$$I_t = sigmoid(W_{x_i} \times X_t + W_{h_i} \times H_{t-1} + b_i) \tag{2}$$

$$C_t^{'} = \tanh(W_{x_c} \times X_t + W_{h_c} \times H_{t-1} + b_c) \tag{3}$$

式中的各系数与式（1）中的系数含义相近，tanh 为双曲正切函数，与 *sigmoid* 函数同属于激活函数。$I_t \times C_t^{'}$ 为保留当前时刻的重要信息。

（3）记忆单元的更新

所谓记忆单元的更新，就是将遗忘了部分历史信息的长期记忆与保留当前时刻重要信息的短期记忆进行叠加。记忆单元的更新表示如下：

$$C_t = F_t \times C_{t-1} + I_t \times C_t^{'} \tag{4}$$

（4）输出门

通过以上步骤，实现了对记忆单元的更新，C_t 即为更新后的记忆单元。但是需要将记忆单元的信息有选择的通过输出门输出给下一时刻。使用 H_t 来表示输出，其计算如下[14]：

$$H_t = sigmoid(W_{x_o} \times X_t + W_{h_o} \times H_{t-1} + b_o) \times \tanh(C_t) \tag{5}$$

至此，完成了 LSTM 算法的全过程。

2.3　随机森林与 LSTM 算法对比分析

随机森林算法在处理大规模复杂数据和高维数据时，具有高准确性和强鲁棒性，不容易导致过拟合；而对于小样本数据集，随机森林可能会因设定的参数过大导致模型的拟合效果较差，因此在小样本处理方面存在一定限制。LSTM 算法擅长处理长期依赖关系和梯度爆炸问题，在进行训练时不需要过多数据就能有较好的拟合效果；但 LSTM 算法模型相较于随机森林算法模型更复杂，需要更多的参数，且训练时间较长[15]。

3　评价模型

通常情况下，可使用均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对误差（*MAE*）等指标，对数值预测模型预测效果进行评价。*MSE*、*RMSE*、*MAE* 的值在 0 到正无穷之间，数值越小表示模型的预测误差越小，其预测能力越强。其值在 0 到 1 之间是可接受的[16]。

均方误差（*MSE*）是指预测值与真实值差值的平方再求和，最后求平均[17]。

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i^{'} - y_i) \times (y_i^{'} - y_i) \tag{6}$$

均方根误差（*RMSE*）的数值是均方误差数值的算术平方根[18]。

$$RMSE = \sqrt{\frac{1}{m}} \cdot \sqrt{\sum_{i=1}^{m}(y_i^{'} - y_i) \times (y_i^{'} - y_i)} \tag{7}$$

平均绝对误差（*MAE*）是预测值与真实值差值的绝对值求和平均[19]。

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | y_i^{'} - y_i | \tag{8}$$

4 实例分析

本文于 2022 年 3 月 1 日至 2022 年 3 月 3 日，将通过测速仪连续采集到成都市高新区红星路三段机动车的行驶速度作为机动车的实际速度，并将该时间段分成 103 个时间段，每个时间段为半小时。依据每辆机动车实际行驶速度，可确定该路段上机动车在每个时间段的平均行驶速度。已知自由流速度 60 km/h，最后依据行程时间比的概念确定每个时间段 *TTI* 数值，以此组成 *TTI* 数据集。

4.1 预测模型

4.1.1 随机森林预测模型

在基于 MATLAB 设计的随机森林预测模型中，首先以年、月、日作为该模型的特征输入；以 *TTI* 数值作为数据集，代表标签，并将该数据集的前 80% 的数据划分为训练集，后 20% 的数据划分为测试集；使用 30 棵决策树，并设置叶子节点的最小样本数为 5。训练集和测试集的预测结果与真实值之间的对比如图 2、图 3 所示：

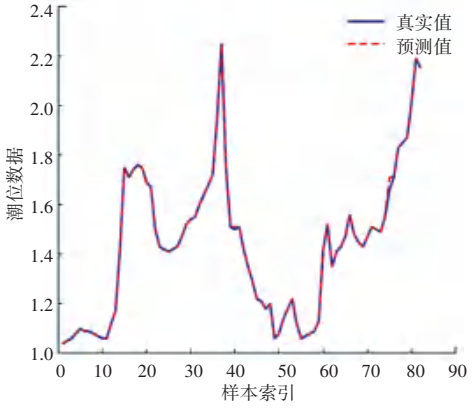

图 2 随机森林模型训练集预测结果对比图

Fig. 2 Comparison of prediction results for random forest model training set

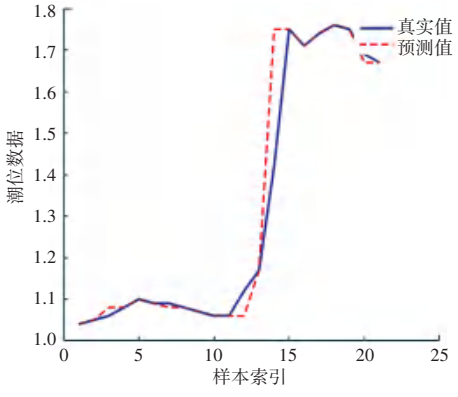

图 3 随机森林模型测试集预测结果对比图

Fig. 3 Comparison of prediction results for random forest model test set

4.1.2 LSTM 预测模型

在基于 MATLAB 设计的 LSTM 预测模型中，同样以年、月、日 3 个维度作为 LSTM 模型的特征输入；其次以 *TTI* 数值作为数据集，同样将该数据集前 80% 的数据划分为训练集，后 20% 的数据划分为测试集。设置了 4 个隐藏单元，最大训练次数为 1 200 次，初始学习率为 0.01，学习率下降因子为 0.1，经过 800 次训练后，学习率为 0.001，每次训练打乱数据集。训练集和测试集的预测结果和真实值的对比结果如图 4、图 5 所示：

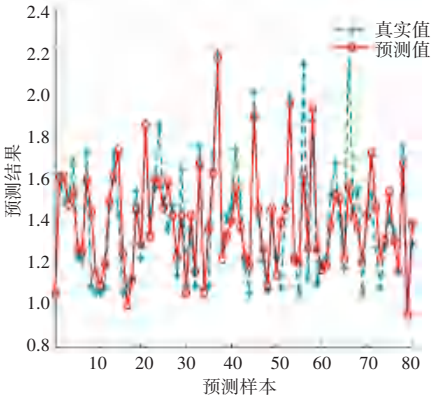

图 4 LSTM 模型训练集预测结果对比图（*RMSE* = 0.155）

Fig. 4 Comparison of prediction results for LSTM model training set

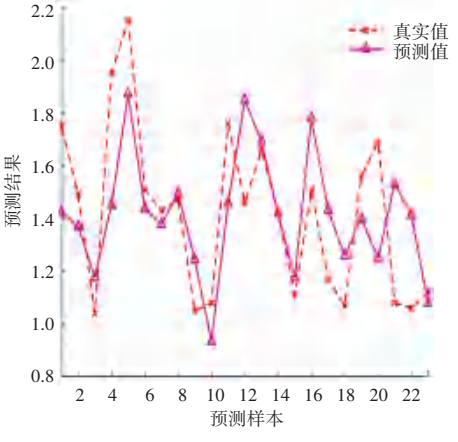

图 5 LSTM 模型测试集预测结果对比图（*RMSE* = 1.114 8）

Fig. 5 Comparison of prediction results for LSTM model test set

4.2 评价模型对比分析

由于模型在训练集上进行的多次迭代训练，通过优化算法对模型进行参数更新和调整，以逐步学习数据的规律和模式。而测试集是模型从未见过的数据，用于模拟模型在实际应用场景中的表现。因此，测试集可用于评估模型的性能[20]。

本文选用的相关指标对各时序预测模型在训练

集中的评估结果见表 2。

表 2　各时序预测模型对比表

Table 2　Comparison table of various time series prediction models

评价指标	随机森林	LSTM
RMSE	1.56	1.11
MSE	2.43	1.23
MAE	0.38	0.21

由表 2 可知，测试集中 LSTM 算法预测的 *RMSE*、*MSE*、*MAE* 值均小于随机森林算法预测的预测结果，说明 LSTM 算法的预测效果较好。由于相关指数 R^2 可表示算法估测的可靠性程度，其取值范围为[0,1]，R^2 越接近 1，说明预测值与真实值越接近，拟合优度越高[21]。而随机森林算法与 LSTM 算法在测试集上的 R^2 值分别为 0.70、0.75，LSTM 算法的 R^2 值比随机森林算法更接近 1，说明 LSTM 算法的预测值更接近真实值，其预测的拟合优度更高。

因此，本文在基于 MATLAB 对 *TTI* 数值进行预测的模型中，设计的 LSTM 时序预测模型的拟合效果最好，其性能总体上均优于随机森林算法。

值得注意的是，由于本文将总时间划分为 103 个时间段，也就只产生了 103 行的 *TTI* 数据集，因此其不算大规模复杂数据集和高维度数据。而由随机森林算法与 LSTM 算法的优缺点对比分析可知，随机森林算法在处理小样本数据时具有一定局限性[22]，而 LSTM 算法不需要过多的数据量就能有较好的预测效果。所以针对该 *TTI* 数据集，可理论上分析推测 LSTM 算法的预测效果优于随机森林算法的预测效果。通过 *RMSE*、*MSE*、*MAE* 等评价指标以及相关指数 R^2 对 LSTM 算法和随机森林算法的实际预测效果进行对比分析，证实了 LSTM 算法的预测效果较好。

4.3　*TTI* 数据集与公开数据集对比分析

本文的数据集来自于成都市 3 天的数据，其时间跨度小、周期性不明显。为了验证 LSTM 模型是否产生过拟合，收集了百度出行的公开数据集，与本文的 *TTI* 数据集进行对比分析[23]。得到 LSTM 算法在公开数据集上预测的 *RMSE*、*MAE* 以及 R^2 的值分别是 0.86、0.15、0.85。由于其 *RMSE*、*MAE* 和 R^2 值均接近于其在 *TTI* 数据集上的 *RMSE*、*MAE* 和 R^2 值，说明该算法没有产生明显的过拟合，且 R^2 值大于 *TTI* 数据集上的 R^2 值，说明了 LSTM 算法在公开数据集上的预测效果优于其在 *TTI* 数据集上的预测效果，进而说明了本文采集的数据集不够丰富，数据量不多。

5　结束语

本文以行程时间比（*TTI*）数值为指标，对城市路段交通运行状况进行等级划分；然后通过 MATLAB 设计出随机森林模型和 LSTM 模型对 *TTI* 数值进行预测分析，以实现基于 MATLAB 的交通拥堵状况预测研究；最后通过 *RMSE*、*MAE* 和相关指数 R^2 对模型的性能进行了对比分析，得出最佳预测模型为 LSTM 模型。并通过与公开数据集对比分析，表明 LSTM 模型没有产生过拟合，具有一定可靠性。

然而，LSTM 模型的预测值与真实值之间的误差值还不够小，拟合效果不够理想，其重要原因就是采集的数据不够丰富，数据量不多，时间跨度小，周期性不明显。下一步将通过采集多个路段的机动车速度并适当增加采集天数来扩充数据量，丰富数据集，并与长周期的公开数据集做对比，以优化预测模型的拟合效果，加强预测模型的有效性和可靠性。

参考文献

[1] 雷晓. 基于LSTM的短时车流量预测模型研究[D]. 重庆:重庆邮电大学,2019.

[2] 魏丹. 基于机器学习的交通状态判别与预测[D]. 长春:吉林大学,2020.

[3] 褚瑞娟. 高速公路交通运行状态判别与预测方法研究[D]. 长春:吉林大学,2021

[4] 易术,黄丹阳. 融合多源数据与元胞传输模型的高速公路交通状态估计方法[J]. 交通运输工程与信息学报,2023,4(9):103-114.

[5] 田润泽. 高速公路交通拥堵识别和预判方法研究[D]. 桂林:桂林电子科技大学,2023.

[6] 陈悦. 基于深度学习的短时交通拥堵状态预测模型研究[D]. 成都:西南交通大学,2021.

[7] 施炎峰,王心莹,朱相如,等. 智能交通系统隐私保护方案综述[J]. 电脑知识与技术,2022,16(1):2-5.

[8] 钱磊,赵长海,孙明. 关于我国智能交通系统发展的思考[J]. 内蒙古科技与技术,2021,103(2):5-6.

[9] 郭浩宇. 基于多源数据融合的交通拥堵预测方法研究[D]. 北京:中国人民公安大学,2022.

[10] BREIMAB L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[11] 吕红燕,冯倩. 随机森林算法研究综述[J]. 河北省科学院学报,2019,37(3):3-5.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[13] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE

Transactions on Neural Networks, 1994, 5(2): 157-166.
[14] LI K, HUANG W, HU G Y. Ultra-short term power load forecasting based on LSTM neural network [J]. Energy and Buildings, 2023, 4(1): 4-5.
[15] 黄衍,查伟雄. 随机森林与支持向量机分类性能比较[J]. 软件,2012,33(6):107-110.
[16] 孙士宏. 城市交通拥堵指数的长短期预测模型研究[D]. 北京:北京交通大学,2023.
[17] 刘铭,何利力,郑军红. 融合多源异构数据的图卷积神经网络混合推荐模型[J]. 智能计算机与应用, 2024,14(2):1-8.
[18] 张威特,李俊松,刘雁飞. 基于随机森林和 BP 神经网络的船舶驾驶员疲劳检测算法[J]. 智能计算机与应用,2024,14(2):142-144.
[19] 李晓宇,张功学,何凯,等. 改进 YOLOv5s 的室内喷涂机器人的窗户检测算法[J]. 智能计算机与应用,2024,14(1):23-25.
[20] 刘伟,Bagui S C,贾宏恩,等. 基于 XGBoost 算法的短期交通流预测[J]. 应用数学进展,2020,9(9):13-15.
[21] 姚亮宇. 基于 LSTM 模型改进的实际交通流预测模型的研究与实现[D]. 南京:南京邮电大学,2021.
[22] 赵锦阳,卢会国,蒋娟萍,等. 一种非平衡数据分类的过采样随机森林算法[J]. 计算机应用与软件,2019,36(4):255-261.
[23] 彭璐. 基于长短时记忆网络的时间序列预测与应用研究[D]. 武汉:华中科技大学,2021.