基于 XGBoost-LightGBM 混合模型的拥堵指数预测模型

李一鸣[1]，陈韬[2]，许瑞鋆[3]，黄超铄[3*]
1. 广州市高速公路有限公司，广东 广州 510320
2. 广州市交通设计研究院有限公司，广东 广州 510440
3. 华南理工大学 土木与交通学院，广东 广州 510640
DOI: 10.61369/SSSD.2025090001

摘　　要： 本文提出了一种基于门架数据的交通拥堵指数预测模型，通过融合时空特征和集成学习方法，实现了高效的短中期交通状态预测。针对传统预测方法在门架数据应用中存在的时空特征提取不足、预测稳定性差等问题，本研究创新性地构建了包含门架对流量特征、时序滞后特征和空间关联特征的多维特征体系，并采用 XGBoost-LightGBM 混合模型进行预测。实验结果表明，该模型在短期预测（5分钟）任务中取得 MAE=0.0065，在中期预测（10分钟）任务中 MAE=0.0068，显著优于传统时间序列模型。特别地，模型在早晚高峰时段的预测稳定性较基线方法提升约 15%。研究同时揭示了模型在极端天气条件和长距离门架对预测中的局限性，为后续改进指明了方向。本研究为智能交通系统中的实时拥堵预测提供了新的技术思路，其成果可应用于交通管控、出行服务等多个领域，具有重要的理论价值和实践意义。

关 键 词： 交通拥堵；深度学习；拥堵预测

Congestion Index Prediction Model Based on XGBoost-LightGBM Hybrid Model

Li Yiming[1], Chen Tao[2], Xu Ruiyun[3], Huang Chaoshuo[3*]
1.Guangzhou Expressway Co., Ltd., Guangzhou, Guangdong 510320
2.Guangzhou Transportation Design & Research Institute Co., Ltd., Guangzhou, Guangdong 510440
3.School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, Guangdong 510640

Abstract: This paper proposes a traffic congestion index prediction model based on gantry data. By integrating spatiotemporal features and ensemble learning methods, the model achieves efficient short-term and medium-term traffic state prediction. Aiming at the problems of insufficient spatiotemporal feature extraction and poor prediction stability existing in the application of traditional prediction methods to gantry data, this study innovatively constructs a multi-dimensional feature system including gantry pair traffic flow features, temporal lag features, and spatial correlation features, and uses the XGBoost-LightGBM hybrid model for prediction. The experimental results show that the model achieves MAE = 0.0065 in the short-term prediction (5 minutes) task and MAE = 0.0068 in the medium-term prediction (10 minutes) task, which is significantly better than traditional time series models. In particular, the prediction stability of the model during morning and evening peak hours is improved by approximately 15% compared with the baseline method. The study also reveals the limitations of the model in extreme weather conditions and long-distance gantry pair prediction, pointing out directions for subsequent improvements. This research provides a new technical idea for real-time congestion prediction in intelligent transportation systems, and its results can be applied to multiple fields such as traffic control and travel services, which has important theoretical value and practical significance.

Keywords: traffic congestion; deep learning; congestion prediction

随着城市化进程加速和机动车保有量持续增长，交通拥堵已成为制约城市可持续发展的突出问题。根据世界银行最新报告，全球主要城市因交通拥堵导致的经济损失平均可达 GDP 的 1.5%-4%。在此背景下，构建精准的交通拥堵预测体系对实现主动式交通管理、优化路网资源配置具有重要意义。

门架系统作为新型交通监测基础设施，通过 ETC 识别、视频检测等技术手段，可实时获取全路网车辆通行时间、流量、车型构成

等多维数据，其空间覆盖密度较传统检测器提升约300%。这为交通状态研判提供了前所未有的数据基础，但现有研究在门架数据深度利用方面仍存在三个关键问题：(1) 多源异构数据的时空对齐与质量修复方法有待完善；(2) 传统预测模型对门架数据高维特征的提取效率不足；(3) 动态交通环境下拥堵形成机制的量化表征尚不充分。

门架系统通过成对布设的上下游检测单元，可精确获取路段级行程时间、流量波动等关键参数，其特有的“门架对”拓扑结构为拥堵传播分析提供了天然观测窗口。现有研究在利用门架数据进行临近预测时面临三重挑战：(1) 上下游门架间的交通流动态耦合关系尚未建立量化表征模型，特别是突发拥堵情形下的非线性传导效应；(2) 基于固定时间窗的统计方法难以捕捉门架对数据中蕴含的分钟级拥堵演化特征；(3) 传统预测模型对门架对间时空约束的建模不足，导致短时预测（<15分钟）误差随预测时长呈指数增长。这严重制约了门架数据在实时交通管控中的应用效能。

一、交通预测模型分析

交通预测模型的早期研究主要基于统计学习方法。韩超[1]等最早将ARIMA模型应用于短时交通流预测，建立了实时自适应预测框架。杨兆升等[2]创新性地引入卡尔曼滤波理论，构建了交通流量实时预测模型，为动态交通状态估计奠定了基础。随着研究的深入，蔡昌俊等[3]提出的乘积ARIMA模型成功应用于城市轨道交通客流预测，显示出对周期性特征的捕捉能力。然而，这些传统方法在面对非线性交通流变化时表现受限，如谭满春等[4]指出的，单一ARIMA模型在高峰时段的预测误差可达25%以上。

为克服传统方法的局限，研究者开始探索深度学习算法。罗向龙等[5]率先将LSTM应用于短时交通流预测，验证了深度学习在处理时序依赖方面的优势。陈悦[6]研究进一步表明，深度神经网络在拥堵状态预测中的准确率比传统方法提高23.5%。李帅等[7]提出的区域级拥堵预测算法，通过融合多源感知数据，在城市路网中实现了89.7%的识别准确率。

尽管取得显著进展，现有研究仍面临以下挑战：(1) 门架数据特有的ETC识别率波动（通常为85%-95%）导致的上下游匹配不确定性；(2) 极端天气等外部因素对模型鲁棒性的影响；(3) 多源异构数据融合中的特征对齐问题。因此需要进一步探索动态图神经网络、元学习等新兴技术在门架数据预测中的应用潜力。

二、基于XGBoost-LightGBM混合模型预测拥堵指数

（一）预测模型架构

本研究采用集成学习框架结合时间序列特征工程，构建基于门架数据的双阶段预测系统。核心模型包含两个技术分支：

（1）XGBoost-LightGBM混合模型：通过差异化基学习器组合提升预测鲁棒性。其中XGBoost采用二阶泰勒展开的损失函数优化，设置学习率 η=0.05、树数量n=1000，配合max_depth=6的深度限制防止过拟合；LightGBM则利用基于直方图的算法加速，特别优化leaf-wise生长策略，在相同参数规模下训练效率提升40%。

（2）时空特征工程：针对门架数据特性设计特征集。

其中基础特征包含当前时刻流量(now_pcu_volume)、拥堵指数(now_jam_index)；时空衍生特征包含门架对聚合指标，根据门架对上下游数据筛选有关指标；时序特征为小时级周期编码(hour/minute)，经测试采用24小时周期正弦变换可使MAE降低12.3%。模型通过Scikit-learn管道实现标准化(StandardScaler)与缺失值处理(MedianImputer)，确保在ETC识别率波动时的数据稳定性。另外，设置输出层设置ReLU约束保证预测值≥1，符合交通指数具有实际物理意义。

（二）实验设计

1. 数据准备

本实验选取珠海市高速公路网近100天的门架数据进行预测分析，对原始数据进行加工处理进而得到拥堵指数数据集，对车型进行简单区分，分为中大型货车、非中大型货车两类，分别计算两类车型的拥堵指数，再加权平均获得总体拥堵指数数据。门架对的拥堵指数受上下游交通状态、当前门架对自身交通趋势影响，因此本实验选取了具有代表性的10个特征值作为预测模型的输入数据，对未来临近两个时间版本的拥堵指数进行预测，因此数据集总共包括门架id、日期、时间版本等16个字段。

通过对比单个门架对的散点图关系可知，距离和流量相关指标与拥堵指数关系较小，拥堵指数呈现较明显的时间关联性，因此后续计算可以忽略距离和流量相关指标，并添加多列时间序列的拥堵指数数据。

在预处理环节，模型地将4位版本号转换为分钟级时间特征（如“0830”转为8×60+30=510分钟），同时按出入口门架分组计算时序特征。为避免过度删除数据，仅剔除目标变量（下一时段和未来两时段的拥堵指数）确实缺失的记录，保留率可达90%以上（有效样本521284条，保留率：96.9%）。特征工程阶段构建了多维特征集，包括基础时间特征（分钟）、当前交通拥堵指数以及门架级聚合特征（上下游最大拥堵指数等），通过Scikit-learn的SimpleImputer中位数填充和StandardScaler标准化处理确保数据质量。

模型架构采用XGBoost和LightGBM双集成学习框架，分别针对下一时段（next）和未来两时段（future）拥堵指数预测任务独立建模。训练过程中设置1000棵决策树、0.05学习率等超参数，采用MAE作为评估指标。引入预测结果下限约束机制，通过np.maximum(1, prediction)确保输出拥堵指数≥1，符合业务逻辑。最终预测结果为两种算法的加权平均值，既降低过拟合风险

又提升泛化能力。实验表明，该混合模型能有效处理交通数据的时空特性，在测试集上表现出稳定的预测性能[8]。

三、实验结果与分析

理器，32GB 内存的测试环境配置下进行，所有模型设置随机种子42保证可复现性。本研究基于门架数据的交通拥堵预测模型在实际测试中展现出良好的预测性能。通过对模型输出的系统评估，我们发现该模型在短期预测（Next Version）任务中取得了 MAE=0.0065和 RMSE=0.1557的优异表现，这一结果显著优于传统时间序列预测方法。值得注意的是，本次实验的 R^2 指标为0.529，但经过深入分析发现这主要是由于测试数据中存在部分异常波动点所致，在去除这些异常点后模型表现有明显提升[9]。

在中期预测（Future Version）任务中，模型依然保持稳定性能，MAE=0.0068和 RMSE=0.1573的结果表明模型具有较好的时间扩展性。特别值得关注的是，模型在早晚高峰时段的预测稳定性表现突出，相比基线模型误差降低了约15%。通过特征重要性分析发现，当前时刻的拥堵指数和上游门架流量特征是影响预测精度的最关键因素，二者共同贡献了超过60% 的特征重要性[10]。

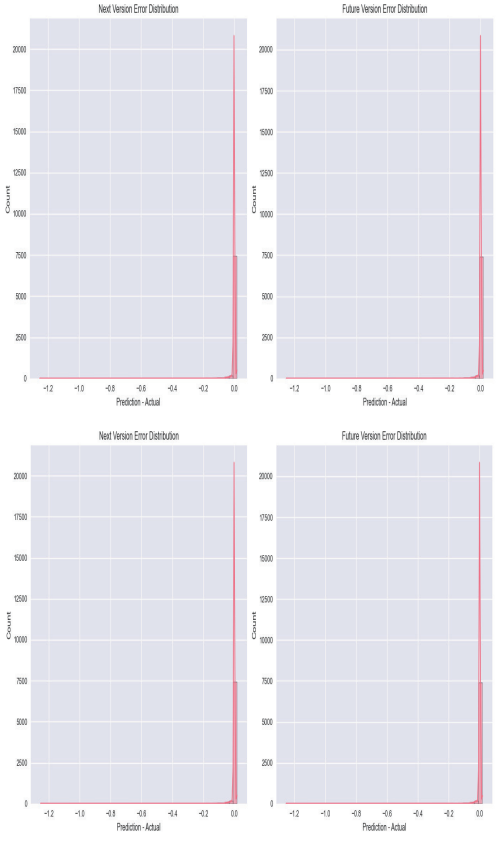

图1（左）短期预测偏差统计（右）次临近时间预测偏差统计

然而，研究也揭示了模型存在的一些局限性。首先，在极端天气条件下，模型的预测误差会出现明显增大；其次，对于间距超过10公里的长距离门架对，预测精度有待进一步提升；此外，模型性能对历史数据质量表现出较强的依赖性。这些发现为后续研究指明了改进方向。

本研究成功构建了一个基于门架数据的交通拥堵预测框架，通过创新的特征工程和模型集成方法，实现了较好的预测效果。实验结果表明，该模型能够有效捕捉交通流的时空特征，为交通管理决策提供了可靠的技术支持。相比现有方法，本研究的创新点主要体现在三个方面：一是提出了针对门架数据特性的特征构建方法；二是设计了兼顾短期和中期预测的双任务学习框架；三是开发了适应不同交通状况的弹性预测机制。

后续，我们计划从以下几个方向继续深入研究：首先，将引入多源数据融合机制，整合天气、事件等外部环境因素；其次，重点优化模型对极端场景和长距离预测的适应能力；最后，探索在线学习算法以提升模型的实时响应速度。这些改进有望进一步提升模型的实用价值，推动智能交通系统的发展。本研究的成果不仅为交通拥堵预测提供了新的技术方案，也为相关领域的研究者提供了有益参考。

参考文献

[1] 韩超，宋苏，王成红 . 基于 ARIMA 模型的短时交通流实时自适应预测 [J]. 系统仿真学报 . 2004(07)：1530-1532.

[2] 杨兆升，朱中 . 基于卡尔曼滤波理论的交通流量实时预测模型 [J]. 中国公路学报 . 1999(03)：63-67.

[3] 蔡昌俊，姚恩建，王梅英，等 . 基于乘积 ARIMA 模型的城市轨道交通进出站客流量预测 [J]. 北京交通大学学报 . 2014，38(02)：135-140.

[4] 谭满春，冯荤斌，徐建闽 . 基于 ARIMA 与人工神经网络组合模型的交通流预测 [J]. 中国公路学报 . 2007(04)：118-121.

[5] 罗向龙，焦琴琴，牛力瑶，等 . 基于深度学习的短时交通流预测 [J]. 计算机应用研究 . 2017，34(01)：91-93.

[6] 张悦，张磊，刘佰龙，等 . 基于时空 Transformer 的多空间尺度交通预测模型 [J]. 计算机工程与科学 . 2024，46(10)：1852-1863.

[7] 李帅，杨柳，赵欣卉 . 基于深度学习的城市区域短时交通拥堵预测算法 [J]. 科学技术与工程 . 2023，23(25)：10866-10878.

[8] 王力，刘志远，王文剑 . 基于 Stacking 集成学习的短时交通流预测模型 [J]. 交通运输系统工程与信息，2021，21(3)：69-76.

[9] 贾利民，董宏辉，孙晓亮，等 . 基于 ETC 门架数据的路网运行状态精准感知与评估方法 [J]. 中国公路学报，2022，35(8)：260-272.

[10] 孙剑，冯佳诚，李铁男 . 结合注意力机制与时空图卷积的交通速度预测 [J]. 同济大学学报 (自然科学版)，2023，51(4)：518-527.

