基于多源数据的景区客流量分析与预测

数科2101班：丁国豪 指导教师：卢胜男 论文类型：毕业设计

摘要：本文针对景区客流量预测难题，提出融合多源数据与深度学习的方法，并以四川四姑娘山景区为研究对象进行实证分析。本文利用爬虫技术采集景区近十年的日客流量数据，通过数据预处理提取关键影响因素。其次，分别构建基于 LSTM 和 Transformer 的单一预测模型，并通过网格搜索优化参数和引入特殊时期标记提升模型性能。最后，创新性地设计并实现了 Transformer-LSTM 和 PI-LSTM 集成模型。实验结果表明，集成模型整体预测效果显著优于单一模型，其中 Transformer-LSTM 模型在 RMSE、MAE 和 R^2等指标上均取得最佳表现。

关键词：景区客流量预测；多源数据融合；集成学习；时间序列分析

1 研究背景

随着旅游业快速发展，景区客流量呈现显著季节性波动和突发性变化，给资源配置、服务调度和安全保障带来巨大挑战。传统预测方法依赖线性统计模型或单一数据源，难以应对天气、节假日、疫情等多因素叠加的复杂波动。大数据与人工智能技术的进步，使得基于多源数据融合的客流量精准预测成为提升景区智慧化管理水平的关键。

2 发展现状

多源数据融合技术：国内外研究普遍认同融合气象、节假日、社交媒体、交通等多源数据能显著提升预测精度。常用技术包括数据预处理、特征提取和模型优化，如 D-S 证据理论、加权模糊融合、时空注意力网络等。

预测技术：预测模型从传统统计方法（如 ARIMA、SARIMA）向机器学习与深度学习演进。LSTM 因其优秀的时序建模能力被广泛应用于客流预测；Transformer 凭借自注意力机制在处理长序列依赖上展现优势；集成学习通过组合不同模型提升整体性能与鲁棒性成为趋势。

3 相关分析

3.1 数据探索与预处理

本研究采用 Python 爬虫技术构建了自动化数据采集系统，通过模拟浏览器请求头参数，有效突破反爬限制，从四姑娘山景区官网稳定获取 2015-2024 年共 3164 条日客流量记录。同时从百度指数平台采集了 6 个相关关键词的 PC 端和移动端综合搜索量数据。图 3.1 是百度指数可视化的结果。

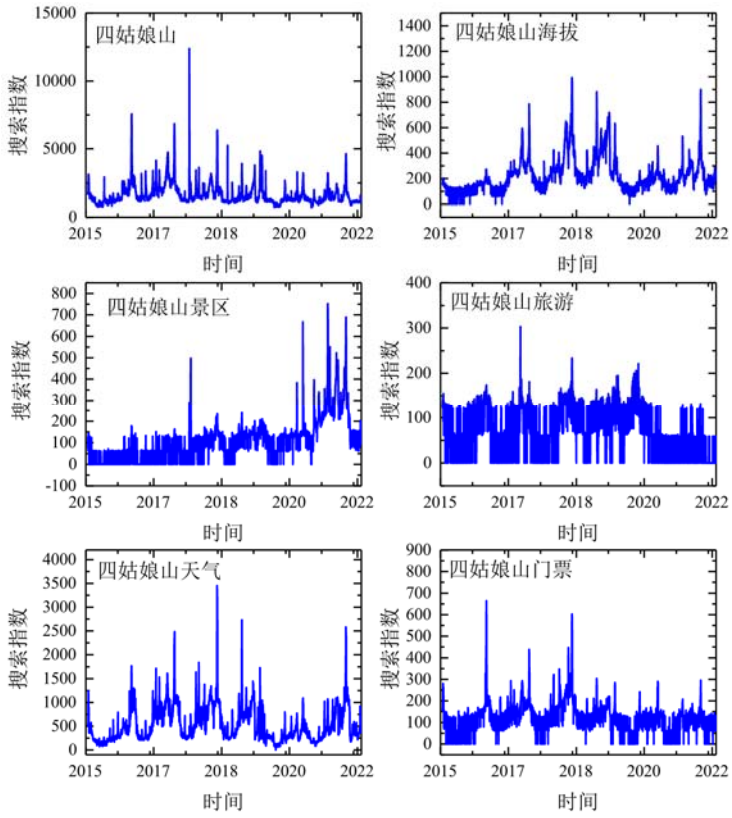


图3.1 六个关键词的百度指数可视化

在数据预处理方面，针对 143 条缺失记录采用差异化填补策略。异常值检测采用 IQR 和 Z-score 方法，发现节假日高峰期客流量达到景区上限 20000 人次属于合理现象，予以保留。特征工程方面，创新性地将"四姑娘山天气"搜索指数作为天气关注度特征，并添加节假日和特殊时期标记，构建了多维特征空间。

数据标准化处理采用 Min-Max 归一化方法，将客流量数据线性变换至[0,1]区间，显著提高了模型训练的稳定性。通过时间序列分解，采用加法模型将原始数据分解为趋势项、季节项和残差项，清晰展现了客流量的长期增长趋势、周期性波动特征以及突发事件的影响效应。分析发现，景区客流量呈现明显的季节性特征，旺季客流量显著高于淡季，节假日效应突出，国庆、春节等假期客流量通常达到峰值。

3.2 预测模型构建与结果分析

在预测模型构建方面，本研究系统性地比较了单一模型和集成模型的性能表现。基础 LSTM 模型通过网格搜索优化确定了最佳参数组合，引入特殊时期标记后模型性能显著提升。表 3.1 是 LTSM 模型优化前后对比结果。

表3.1 LSTM优化性能对比

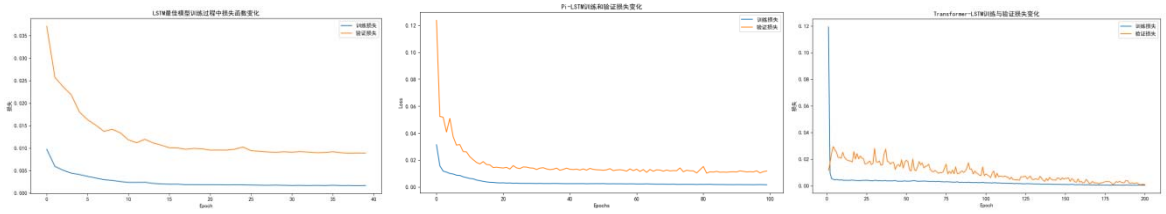
模型	RMSE	MAE	R
原始模型	1909.01	1229.29	0.8725
加入标记后	1123.00	582.61	0.9150
网格搜索寻优（最优结果）	1113.06	599.70	0.9165

创新性地提出的 Transformer-LSTM 集成模型通过并行架构结合了两种模型的优势，采用加权平均策略融合预测结果。实验结果表明，该模型在各项评估指标上均表现最优，预测曲线与实际值拟合度最高。PI-LSTM 模型通过引入物理信息约束，将数据驱动与基于规则的预测相结合，在特殊事件解释性方面展现出独特优势。Transformer 模型虽然理论上更适合处理长序列依赖，但在本数据集上表现相对欠佳，分析可能是由于数据规模不足或模型结构需要调整。评价指标对比如表 3.2 所示：

表3.2 各模型评价结果

模型	RMSE	MAE	R^2
LSTM	1113.06	599.70	0.9165
Transformer	1422.78	691.32	0.8644
Transformer-LSTM	373.04	255.74	0.9906
PI-LSTM	965.07	507.83	0.9373

模型对比分析发现，集成方法能有效提升预测性能。图 3.2，图 3.3 和图 3.4 是三种模型的损失函数变化。Transformer-LSTM 训练损失和验证损失均能快速下降并保持稳定，表明该架构具有优秀的泛化能力。



(a)LSTM最佳模型　　(b)PI-LSTM模型　　(c)Transformer-LSTM模型

图3.2 各模型训练过程损失函数变化

4 研究结论及对策建议

4.1 研究结论

融合景区客流历史数据、天气搜索指数、节假日及特殊时期标记等多源信息，能有效捕捉影响客流的关键因素。深度学习模型（LSTM, Transformer）在景区客流量预测中表现优异，其中经过参数优化和特征增强的 LSTM 模型已具备较高精度。提出的两种集成模型显著提升了预测性能：Transformer-LSTM 模型通过结合长短期依赖建模优势，实现了最高精度的预测；PI-LSTM 模型通过融入物理信息约束，增强了对异常事件的可解释性和适应能力。Transformer 模型在本研究数据集上表现相对欠佳，可能需更大规模数据或结构调整。

4.2 对策建议

在数据融合与模型优化方面，建议进一步整合多维度数据源，包括社交媒体舆情数据、游客消费数据以及周边交通流量数据，构建更为全面的游客行为特征体系。

同时需要持续推进算法创新，通过采用自监督学习技术解决标注数据不足的问题，开发基于迁移学习的跨景区通用预测框架，并探索时序Transformer变体等新型算法以提升长序列预测效率。针对实时预测系统建设，应当重点优化计算架构设计，通过部署边缘计算节点实现景区本地化实时预测，并应用联邦学习技术保障多景区数据隐私下的协同训练。此外还需深化物理模型研究，将景区最大承载量、游客移动轨迹等物理规则编码到PI-LSTM中，建立包括极端天气在内的突发事件动态约束机制。