

【交通物流 / Transportation Logistics】

基于GPS数据的公交站运行状态分析

黄洪滔¹, 肖梅¹, 刘倩¹, 明秀玲¹, 边浩毅²

1) 长安大学运输工程学院, 陕西西安710064; 2) 浙江机电职业技术学院, 浙江杭州310053

摘要: 为有效提高公交站点的运行效率, 对公交站的运行状态进行识别、预测及影响因素分析, 以中国西安市公交车全球定位系统轨迹数据为例, 建立平均服务时间和服务车数特征参数反映公交站的运行状态, 并通过分析站点内公交车速度、里程及加速度之间关系计算站台服务时间. 使用Hopkins统计量和轮廓系数分析可聚性和聚类数, 结合高斯混合模型(Gaussian mixture model, GMM)对公交站运行状态进行识别分类. 构建SMOTEENN-XGBoost(synthetic minority oversampling technique edited nearest neighbours - extreme gradient boosting)站点运行状态预测模型, 引入可解释机器学习框架SHAP(Shapley additive explanation)分析站台属性、道路及环境对模型的影响. 结果表明, 公交站运行状态可分为3类, 类型I的平均服务时间最长, 类型II的平均服务时间和服务车数最少, 类型III的服务车数最多; 所建立SMOTEENN-XGBoost模型的准确率为94.68%, 精确率为94.69%, 召回率为91.04%, F_1 分数为92.26%, 与极限梯度提升(extreme gradient boosting, XGBoost)、逻辑回归(logistic regression, LR)、随机森林(random forest, RF)、梯度提升决策树(gradient boosting decision tree, GBDT)和 k 近邻(k -nearest neighbors, KNN)5种模型对比, 本模型能够精准预测站点运行状态; 对站点运行状态具有影响作用的因素按照重要程度由大到小依次为线路数、有无公交专用道、泊位数、站台设置方法、站台几何形状、车道数、站台设置位置、是否工作日、时段及天气类型. 研究结果可为公交站点设计优化提供一定参考依据.

关键词: 城市交通; 公交站点; 运行状态; XGBoost模型; 全球定位系统(GPS)数据; 可解释机器学习
中图分类号: U491.1; TP391 **文献标志码:** A **doi:** 10.3724/SP.J.1249.2023.03326

Analysis of bus-stop operating state based on GPS data

HUANG Hongtao¹, XIAO Mei¹, LIU Qian¹, MING Xiuling¹, and BIAN Haoyi²

1) College of Transportation Engineering, Chang'an University, Xi'an 710064, Shaanxi Province, P. R. China

2) Zhejiang Institute of Mechanical & Electrical Engineering, Hangzhou 310053, Zhejiang Province, P. R. China

Abstract: In order to effectively improve the operational efficiency of bus stops, this study conducts identification, prediction and analysis of the influencing factors of bus stops with different operational states. Taking the global positioning system (GPS) tracing data of bus vehicle in Xi'an as an example, the characteristic parameters of bus stop operating state including average service time and number of service vehicles were established, and the relationship between bus speed, mileage and acceleration within the bus stops was analyzed to calculate the service time. The Hopkins statistics and silhouette coefficient were used to analyze the clustability and determine the number of clustering categories, and three types of bus stop were identified based on Gaussian mixture model (GMM). The

Received: 2022-03-16; **Revised:** 2022-06-14; **Accepted:** 2022-11-15; **Online (CNKI):** 2022-12-15

Foundation: The "Pioneer" and "Leading Goose" R & D Program of Zhejiang Province (2022C01105); Social Science Planning Foundation of Shaanxi Province (2022F021)

Corresponding author: Professor XIAO Mei. E-mail: xiaomei@chd.edu.cn

Citation: HUANG Hongtao, XIAO Mei, LIU Qian, et al. Analysis of bus-stop operating state based on GPS data [J]. Journal of Shenzhen University Science and Engineering, 2023, 40(3): 326-334. (in Chinese)



synthetic minority oversampling technique edited nearest neighbours - extreme gradient boosting (SMOTEENN-XGBoost) model was constructed to predict the bus stop operating state, and interpretable machine learning framework named Shapley additive explanation (SHAP) was employed to analyze the influence of three aspects: station attributes, road factors, and environmental factors on the operating state of bus stop. The results show that type I has the longest average service time, type II has the least average service time and number of service vehicles, and type III has the highest number of service vehicles. Compared with extreme gradient boosting (XGBoost), logistic regression (LR), random forest (RF), gradient boosting decision tree (GBDT) and k -nearest neighbors (KNN), the established SMOTEENN-XGBoost model can accurately predict the operating state of bus stop with an accuracy of 94.68%, precision of 94.69%, recall of 91.04%, and F_1 score of 92.26%. The factors that influence the bus-stop operating state in descending order of importance are: the number of lines, the presence of bus lanes, the number of parking spaces, the method of platform installation, platform geometry, the number of lanes, the location of the platform, working day, the time of day and the type of weather. The results of the study can provide some reference basis for bus stop design optimization.

Key words: urban traffic; bus stop; operating state; XGBoost model; global position system (GPS) data; interpretable machine learning

公交进站停靠是实现乘客和车辆交互以及时空转移的过程, 是公交运营中必不可少的阶段^[1]. 公交站内车辆的运行特征包括公交车停靠特征与到达特征^[2]. 车辆停靠时间与单位时间内的车辆到达数均会影响公交站的运行状态, 进而影响公交系统的运行效率. 因此, 有必要对公交站运行状态进行识别与因素分析, 以提高公交系统的服务质量和吸引力.

对公交或地铁站的分类研究一般具有两个特征: ① 根据站点的设置位置、设置方法、几何形状及周边用地性质等静态指标研究站点分类^[3-4]. 《城市公共汽电车和无轨电车工程项目建设标准》^[3]将公交场站分为首末站、中途站、保养场及停车场. 叶娇^[4]提出根据节点区位特性、换乘需求量及周边用地性质进行城市交通枢纽层次划分的方法. ② 对动态指标使用聚类分析方法, 进而分类识别站点. 朱宏等^[5]使用主成分分析法提取关于站点服务、站点接驳及步行环境满意度3个因子, 在此基础上对19个快速公交站点进行 k -means聚类, 最终划分为5类; 马壮林等^[6]通过分析地铁进出站客流数据对站点进行高斯混合模型聚类, 归纳出不同类型站点的时空分布特征规律. 当考虑如何评价站点的运行状态时, 王学勇等^[7]以每小时平均到达车数与平均停靠时间等指标评价站点服务效率, 并计算出自由停靠和定点停靠时公交站的合理泊位数; 卞张蕾^[8]考虑车辆进出站速度、时间和平均服务时间等因素, 使用 k -means算法对公交站的运行状态进行分类识别; 朱宁等^[9]使用公交负荷到达率衡量站

点的运营情况, 并提出站点延误估算模型.

在站点内公交车辆停靠与到达过程的影响因素分析方面, TAN等^[10]在分析泊位分配与公交站车辆延误过程中, 验证了高峰时段和公交流量变大会增加公交延误; BIAN等^[11]对公交站泊位数、线路数及乘客数量等因素间的关系进行分析, 建立站点服务时间模型; 柳伍生等^[12]研究公交进出站和停站延误, 认为泊位数、公交到达率及上下车乘客数对公交延误具有影响; 张梦洁^[13]从站台、公交停靠及环境方面分析公交停靠时间影响因素, 解释了时段、站型、有无公交专用道及泊位数等因素的显著影响作用.

机器学习模型可以确定以上各因素之间的非线性关系, 并以较高精度预测站点运行状态, 结合可解释机器学习框架SHAP(Shapley additive explanation)可以探究内部变量对预测结果的影响. 罗妍等^[14]使用了可解释机器学习框架SHAP, 提升了机器学习预测模型的透明度. 吕能超等^[15]构建了基于极限梯度提升(extreme gradient boosting, XGBoost)的碰撞风险预测模型, 并引入SHAP对模型变量的影响进行解释.

由以上分析可知, 在公共交通站点分类的现有研究中, 多数研究集中在轨道交通枢纽领域, 公交站的分类识别主要考虑周围用地性质等静态指标, 而综合考虑站点运行特性的研究较少, 难以体现对公交站运行状态的实际评价. 本研究针对公交站运行状态展开研究, 以中国西安市公交车辆全球定位系统(global positioning system, GPS)轨迹数据为基

基础, 将公交车辆的停靠和到达特征作为公交站运行状态特征变量的提取依据, 根据公交站平均服务时间和服务车数特征对公交站运行状态进行识别, 并识别出 3 类公交站. 提出考虑数据不均衡性的 SMOTEENN-XGBoost (synthetic minority oversampling technique edited nearest neighbours - extreme gradient boosting) 站点运行状态预测模型, 利用可解释机器学习框架 SHAP 剖析站台属性、道路及环境对站点运行状态的影响. 研究结果有望进一步协调站点车辆运行, 为乘客提供准确的站点和车辆信息服务, 进而提升公交系统的整体服务质量.

1 数据基础

1.1 公交站运行状态特征变量提取

采用西安市 2017-07-03 至 2017-07-09 的公交车辆 GPS 轨迹数据, 包含公交车辆的身份标识号 (identity document, ID)、经纬度、运行速度、记录时间、当日行驶总里程及行驶方位角等信息. 使用 ArcGIS 对清洗后的 GPS 轨迹数据进行站点匹配, 将所有数据按站点进行归类. 由于公交站的运行状态主要由进站车辆运行特点决定, 因此, 选取公交站内车辆的停靠特征和到达特征来描述站点运行状态特点.

1) 平均服务时间. 平均服务时间为 1 小时内公交站服务时间的平均值, 表示不同运行状态公交站的时间特征. 当车辆经历多次启动停车时, 以车辆驶离站台前的最后一次启动时间与到达站台前的最后一次停车时间之差作为公交站服务时间^[16]. 公交车在站台范围内的运行速度一般不超过 15 km/h, 因此, 可将公交车辆的进出站过程视为匀加速过程. 考虑到 GPS 轨迹数据的采样频率为 10 s, 在采样间隔时间内存在公交车辆停止和启动的现象, 绘制典型公交站公交运行示意图, 如图 1. 其中, A 和 F 为车辆停止前和启动后采集到的第 1 个数据点; B 和 E 为实际的车辆停止和启动数据点; C 和 D 为采集到的车辆停止和启动数据点; t 为时刻; v 为车速; L 为当日当前时刻的行驶总里程. 点 A 到 B 之间车辆加速度 a 、 v 及 L 的关系为

$$\begin{cases} v_B^2 - v_A^2 = 2a(L_B - L_A) \\ v_B = v_A + a(t_B - t_A) \end{cases} \quad (1)$$

由式(1)可推导出公交车辆减速至停止的时刻 t_B . 同理, 可推导出公交车辆停止至加速的时刻 t_E ,

定义站台服务时间 T 为 t_B 和 t_E 之差, 即

$$T = t_E - t_B = t_F - t_A - \frac{2(L_F - L_E)}{v_F} - \frac{2(L_B - L_A)}{v_A} \quad (2)$$

2) 服务车数. 服务车数为站点小时内公交车辆的到达数, 很大程度上反映了公交车辆到达率以及该站点的运营状况, 因此, 选用服务车数作为衡量站点运行状态的特征指标.

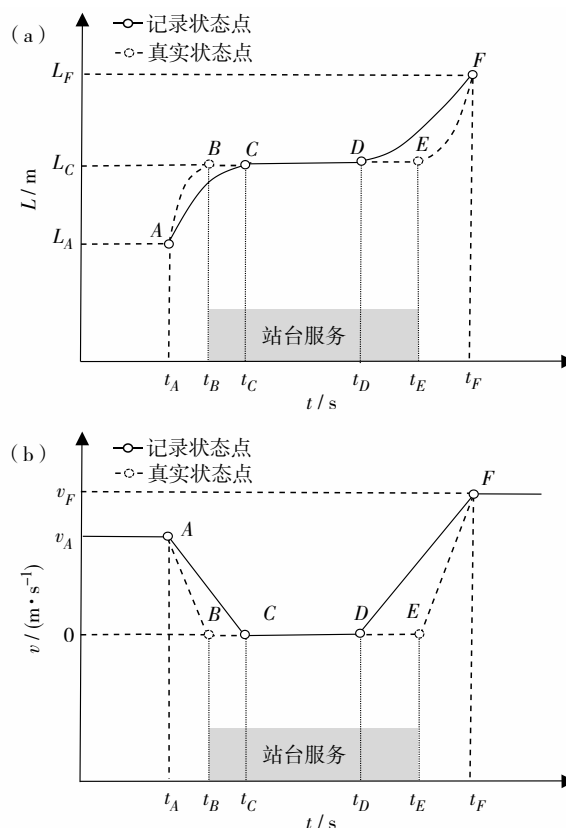


图 1 公交站内公交车辆的(a)行驶里程和(b)速度变化示意

Fig. 1 Schematic of (a) mileage and (b) speed changes of bus vehicles at bus station. Solid line with circle is for the recorded status points. Dashed line with circle is for the real status points.

为保证合理性, 公交站选取的原则^[13]包括:

① 选择位于基本路段的公交站时, 保证数据不受交叉路口信号控制的影响; ② 社会车辆干扰等不确定因素会导致站内车辆在某一瞬间速度骤增或骤减, 避免选取此类公交站; ③ 选取非节假日的公交站运行数据.

1.2 影响因素提取

根据已有的站点车辆运行特性研究结果^[1,10-13,17], 结合实际从站台、道路及环境方面初选 10 个因素, 包括站台设置方法、站台设置位置、线路数、有无公交专用道、车道数、天气类型、是否

工作日、站台几何形状、泊位数及时段, 变量描述如表1所示. 将这10个因素分别与站点的平均服务时间和服务车数进行相关性检验, 结果显示所选因素与站点的平均服务时间或服务车数存在显著性相关关系 ($P < 0.05$).

表1 公交站影响因素的变量描述

Table 1 Variable description of bus stop influencing factors

类别	变量名称	变量类型	变量描述	变量值
站台属性	站台设置方法	分类	沿机非隔离带设置	0
			沿人行道设置	1
	站台几何形状	分类	直线式站台	0
			港湾式站台	1
	泊位数	连续		
道路因素	线路数	连续		
	站台设置位置	分类	基本路段	0
			上下游路段	1
	有无公交专用道	分类	无	0
			有	1
环境因素	车道数	连续		
	天气类型	分类	晴天	0
			雨天	1
	是否工作日	分类	非工作日	0
			工作日	1
	时段	分类	平峰	0
			高峰	1

2 基于高斯混合模型的站点运行状态识别

为了融合公交站的不同运行状态特征, 确定运行状态类型, 使用聚类算法对站点运行状态特征数据进行聚类. 常用的划分聚类算法有针对连续数值的 k -means 算法、针对离散类别的 k -modes 算法、基于密度的带噪声应用空间聚类 (density-based spatial clustering of applications with noise, DBSCAN) 算法及从概率角度出发的高斯混合模型 (Gaussian mixture model, GMM) 等^[18-19]. GMM 融合了参数和非参数估计方法的特点, 对异常值的识别度更高^[19], 因此, 本研究使用 GMM 模型进行数据分类.

2.1 GMM 模型

GMM 模型通过混合多个基于高斯模型的概率分布来表示数据分布, 若将公交站运行状态分为 M 类, 则由 M 个单高斯分布组成的混合模型可表示站

点的运行状态, 高斯混合模型 $P(x)$ 可表示为

$$\begin{cases} P(x) = \sum_{m=1}^M w_m g(x; \mu_m, \sigma_m) \\ \sum_{m=1}^M w_m = 1 \end{cases} \quad (3)$$

其中, M 为高斯分布的个数; w_m 、 μ_m 及 σ_m 为第 m 个高斯分布的权重系数、均值及协方差; $g(x; \mu_m, \sigma_m)$ 为第 m 个高斯分布的概率密度函数.

采用样本轮廓系数 s 确定最佳聚类数 M , s 计算为

$$s = \frac{b - a}{\max(a, b)} \quad (4)$$

其中, a 为与同簇中数据点的平均距离; b 为与下一个最近簇中数据点的平均距离. s 越接近 1 聚类效果越好, 越接近 -1 聚类效果越差.

2.2 公交站运行状态识别

数据的相对大小会影响聚类结果, 因此, 对站点平均服务时间和服务车数进行 z -score 标准化. 对处理后的数据进行 Hopkins 聚类趋势分析, 计算出统计值为 0.908, 表明公交站的运行特征具有可聚性. 根据轮廓系数极值得到 k -means 与 GMM 聚类的最佳聚类数均为 3, 且当聚类数为 3 时 k -means 算法和 GMM 的轮廓系数分别为 0.560 和 0.693, 可见, 高斯混合模型在站点运行状态识别方面的有效性更强.

将站点运行状态根据特征分为 3 类, 分别为类型 I、类型 II 及类型 III, 其基本特征如图 2 和表 2 所示. 可见, I 类公交站的服务车数不多, 平均服务时间最长为 63.261 s, 表明该运行状态的公交站服务车辆少且服务时间最长; II 类公交站的平均服务时间和服务车数在 3 种类型中最低; III 类公交站的平均服务时间短, 服务车数最多为 125 辆, 表明该运行状态的公交站服务车辆最多且服务时间短.

3 模型构建

考虑站台属性、道路和环境等要素, 构建 SMOTEENN-XGBoost 站点运行状态预测模型. 利用 SHAP 可解释性分析框架解释因素对预测结果影响.

3.1 SMOTEENN-XGBoost 预测模型

考虑到不同公交站运行状态样本数不均衡时, 会影响到预测模型的训练, 预测结果会倾向于样本数较多的类型, 因此, 本研究结合 SMOTEENN (synthetic minority oversampling technique edited

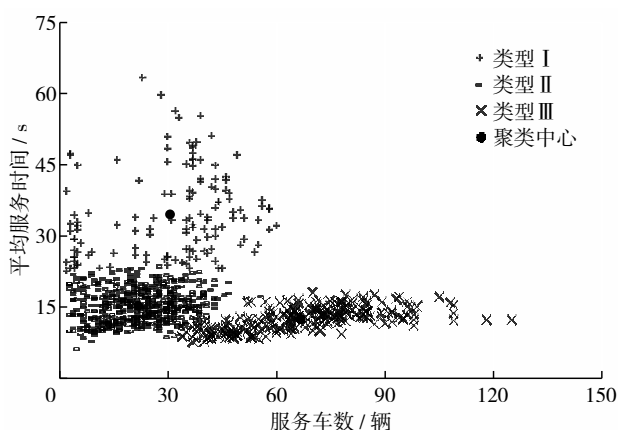


图2 GMM 聚类结果

Fig. 2 Clustering results by GMM. The plus signs are for type I, the minus signs are for type II, the multiplication signs are for type III, and the solid circles are for the clustering centers.

表2 三类公交站运行状态的特征

Table 2 Characteristics of operating states of three types of bus stops

公交站类型	服务车数/辆	平均服务时间/s	聚类中心	样本数
I	2~60	22.750~63.261	30.573, 34.506	134
II	2~55	6.250~23.629	21.698, 15.260	561
III	34~125	7.676~18.100	66.776, 12.491	242

nearest neighbours) 和 XGBoost 算法构建 SMOTEENN-XGBoost 预测模型. 使用 SMOTEENN 算法对训练集数据进行不均衡处理, 对于少数类样本, 使用过采样 SMOTE (synthetic minority oversampling technique) 算法以线性插值方式增加新样本; 对于多数类样本, 使用欠采样 ENN (edited nearest neighbours) 算法对与其 k 近邻点类别不同的数据进行删除, 使得处理后的数据更可靠^[20]. 极限梯度提升算法是在梯度提升决策树算法基础上提出的一种叠加树模型, 其对目标函数进行二阶泰勒展开, 加入模型复杂度的正则项, 使目标值和实际值的误差更小, 进而提高数据预测准确度.

假设模型样本为 (x_i, y_i) , $i = 1, 2, \dots, n$, $f_q(x_i)$ 为第 q 个决策树的预测结果, 经过 Q 轮迭代后模型的预测结果为

$$\hat{y}_i = \sum_{q=1}^Q f_q(x_i) \quad (5)$$

目标函数为

$$O = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^T \Omega(f_i) \quad (6)$$

其中, $l(y_i, \hat{y}_i)$ 为模型的损失函数; Ω 为模型的正

则项. 将目标函数泰勒展开至二次项, 采用贪心算法可以得到最终模型.

3.2 模型参数确定

本研究按照 0.8 的比率随机划分训练集, SMOTEENN 均衡化后的训练集包含 312 个 I 类样本、398 个 II 类样本及 278 个 III 类样本. 为提升模型的性能, 采用网格搜索法和五折交叉验证法寻找预测模型的最优参数组合, 表 3 为 SMOTEENN-XGBoost 模型的最终参数.

表3 SMOTEENN-XGBoost 模型参数

Table 3 SMOTEENN-XGBoost model parameters

参数名	含义	最优值
n_estimators	迭代次数	261
learning_rate	学习速率	0.1
max_depth	树的最大深度	5
min_child_weight	子节点最小样本权重和	1
gamma	子节点分裂所需的最小损失减少值	0
subsample	随机采样比率	0.5
colsample_bytree	特征采样比率	0.6

3.3 模型预测精度评价

使用准确率 A 、精确率 P 、召回率 R 及 F_1 分数 F_1 作为评估指标, 评估指标的值越接近于 1 表明预测效果越好, 由于本模型为多分类预测模型, 因此, 根据各类型数据所占比例计算各类型评估指标的加权平均值, 评估指标的定义为

$$A = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (7)$$

$$P = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \quad (10)$$

其中, TP 表示真实和预测值都是当前类型的样本数; FP 表示真实值不是当前类型且预测值是当前类型的样本数; FN 表示真实和预测值都不是当前类型的样本数.

为了验证模型的预测性能, 选取 XGBoost、逻辑回归 (logistic regression, LR)、随机森林 (random forest, RF)、梯度提升决策树 (gradient boosting decision tree, GBDT) 和 k 近邻 (k -nearest neighbors, KNN) 5 种机器学习算法进行对比, 在相同数据集基础上进行模型的训练和测试. 表 4 是各个模型的对比结果. 可见, 与其他 5 种模型相比,

SMOTEENN-XGBoost模型的准确率、精确率、召回率及F1分数均为最高, 预测性能最好.

表4 六种模型的预测性能比较

Table 4 Prediction performance comparison of the six models				
模型名称	A	P	R	F_1
SMOTEENN-XGBoost	94.68	94.69	91.04	92.26
XGBoost	92.88	93.40	86.98	88.87
LR	89.68	88.57	86.02	86.50
RF	91.81	88.61	86.46	86.93
GBDT	92.88	93.40	86.98	88.87
KNN	86.83	85.07	77.11	80.35

3.4 基于可解释机器学习框架SHAP的模型解释

SHAP是一种事后解释黑箱机器学习模型的框架. 对于每个样本的不同影响因素会生成对应的SHAP值, SHAP值的主要作用是量化各因素对模型的贡献. 假设第*i*个样本的第*j*个因素为 x_{ij} , x_{ij} 的SHAP值为 $f(x_{ij})$, 模型对第*i*个样本的预测值为 y_i , 整个模型的基线(通常是所有样本目标变量的均值)为 y_{base} , 则SHAP值有如下关系

$$y_i = y_{base} + f(x_{i1}) + \dots + f(x_{ik}) \quad (11)$$

其中, $f(x_{ij})$ 即为第*i*个样本中第*j*个因素对预测值 y_i 的贡献值. 当 $f(x_{ij}) > 0$ 时, 该因素提升了预测值; 反之, 该因素使预测值降低.

根据SHAP值得到各个因素对模型分类结果的重要性排序, 如图3. 可见, 各因素按照重要程度由大到小的排序为线路数、有无公交专用道、泊位数、站台设置方法、站台几何形状、车道数、站台设置位置、是否工作日、时段和天气类型.

图4为各因素对不同公交站运行状态的影响, 横坐标为因素的SHAP值, 每个点代表1个样本,

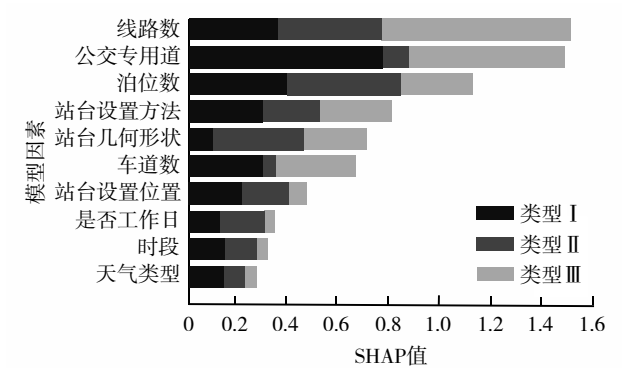


图3 因素重要度排序
Fig. 3 The importance order of factors.

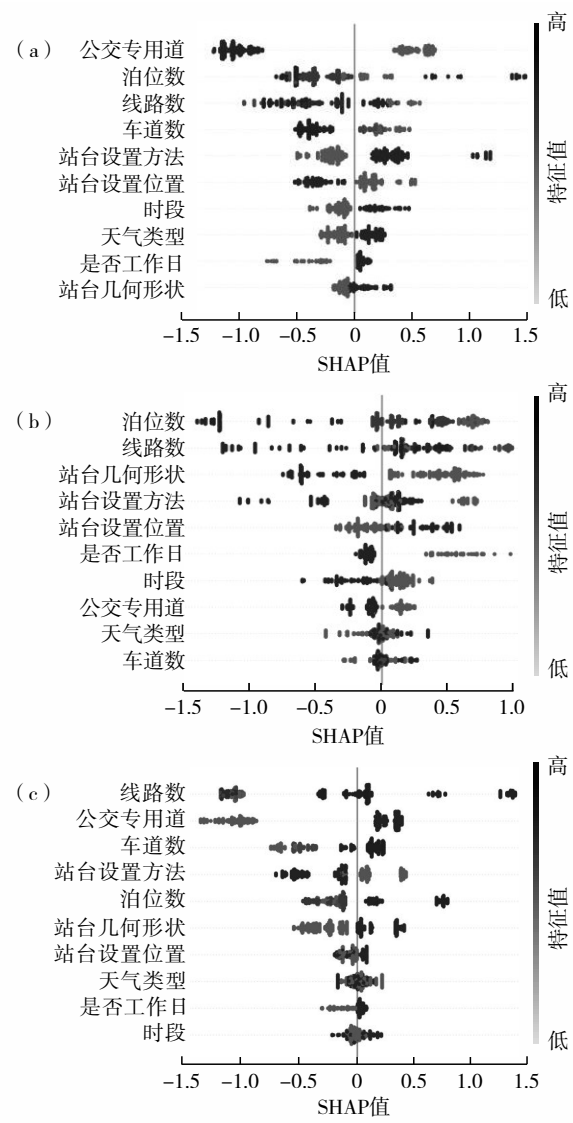


图4 因素对公交站运行状态(a)类型Ⅰ、(b)类型Ⅱ及(c)类型Ⅲ的影响分析

Fig. 4 Analysis of influencing factors of bus stop operating state on (a) type I, (b) type II, and (c) type III.

灰度值越低即颜色越深表示因素值越大.

在站台属性方面, 线路数是重要度最高的因素, 随着线路数的增加, 站点运行状态属于类型Ⅲ的概率越大, 这与线路数越多时到达公交站的车辆数越多, 站点的服务车数越多有关; 泊位数对于类型Ⅱ的重要程度最高; 在站台设置方法中, 沿人行道设置的公交站倾向属于类型Ⅰ; 站台几何形状对类型Ⅱ和类型Ⅲ具有有较高贡献度, 相比于直线式公交站, 港湾式公交站运行状态倾向属于类型Ⅲ, 这是由于港湾式公交站通常设置在机动车饱和度较大的路段, 对应客流量一般比较大, 导致更多的公车辆经过^[13].

在道路因素方面, 有无公交专用道是重要度最高的因素, 公交专用道提供的路权可以减少车辆在站点的服务时间, 对类型 I 和类型 III 的影响最为显著; 同理, 随着车道数的减少, 公交车辆进出站容易受其他车辆的影响而导致停留时间较长; 位于上下游的公交站不倾向属于类型 I, 倾向属于类型 II.

在环境因素方面, 是否工作日是影响站点运行状态的主要因素, 非工作日时站点运行状态属于类型 II 的概率大, 容易出现服务时间和服务车数少的现象; 平峰时段车辆少, 不易出现由于机动车和乘客流量大所导致的车辆停留时间长的现象, 站点运行状态不倾向属于类型 I, 倾向属于类型 II; 天气类型难以直接表征对类型 II 和类型 III 的影响, 但对类型 I 具有一定影响.

3.5 基于 SHAP 的因素交互分析

在实际情况中, 公交站的运行状态会受到多种因素的综合影响, 单一因素分析难以全面解释站台属性、道路和环境因素对模型的影响, 因此, 有必要对不同因素间的交互作用及其对模型的影响做进一步探究. 以聚类特征比较明显的类型 I 和类型 III 中的不同因素组合为研究对象, 图 5 和图 6 分别为站台设置方法、有无公交专用道、天气类型及线路数等因素组合的 SHAP 交互图.

当不考虑图形灰度差别时, 图形 SHAP 值的变化反映了主因素对预测结果的影响. 由图 5(a) 可见, 沿人行道设置的公交站运行状态属于类型 I 的可能性增大, 这与公交车辆进出站时行人与公交车容易产生冲突, 进而增加停留时间相关; 样本点灰度值的变化体现了站台设置方法与公交专用道的交互作用, 当站台设置方法取值从 0 变成 1 时, 灰度值低的点转至灰度值高的点下方, 即 SHAP 值为正时, 无公交专用道的样本点 SHAP 值更高. 该结果表明沿人行道设置的站点, 以及无公交专用道的站点, 其运行状态属于类型 I 的可能性更大. 由图 5(b) 可见, 当天气类型由晴天变成雨天时, SHAP 值由负值变成正值, 这与雨天时乘客的上下车时间增加, 车辆由于视线干扰、路况变差等原因在进出站耗时更多有关; 在 SHAP 值为正值时, 沿人行道设置的公交站样本点位于沿机非隔离带设置的公交站的样本点上方, 这与图 5(a) 的结论一致.

类似地, 图 6 显示了运行状态类型 III 中有无公交专用道与线路数、站台几何形状的 SHAP 交互

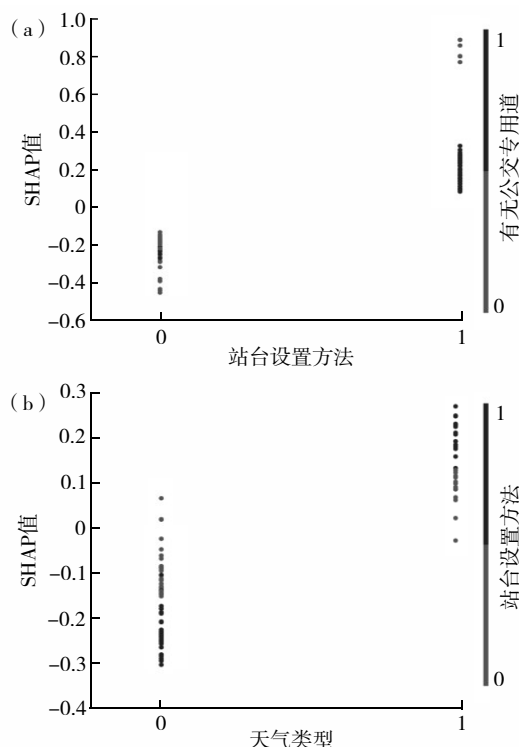


图 5 考虑(a)站台设置方法和公交专用道与(b)天气类型和站台设置方法因素的类型 I 公交站 SHAP 交互性分析

Fig. 5 SHAP interaction analysis of type I considering (a) method of platform installation and bus lane and (b) weather of type and method of platform installation factors.

图. 可见, 随着线路数变多, SHAP 值增大, 当站台几何形状为港湾式时, 站点运行状态属于类型 III 的概率更大; 当 SHAP 值 > 0 时, 有公交专用道的样本点位于无公交专用道的样本点的上方, 这是因为有公交专用道时, 公交车辆驶入和驶出公交站时具有专用路权, 与其他车辆冲突的影响较小, 在站点内的停留时间较短.

4 结 论

本研究以西安市公交车辆的 GPS 轨迹数据为研究对象, 将平均服务时间和服务车数作为公交站运行状态的特征变量, 通过聚类融合后识别出 3 类公交站. 考虑到样本的不均衡性构建 SMOTEENN-XGBoost 预测模型, 引入 SHAP 值增强模型的可解释性, 进而研究站台属性、道路因素及环境因素对公交站运行状态的影响, 得到的主要结论如下:

1) Hopkins 统计值为 0.908 说明站点的运行状态特征具有可聚性, 使用 GMM 方法确定 3 类公交站运行状态, 其中, 类型 I 的平均服务时间最长为

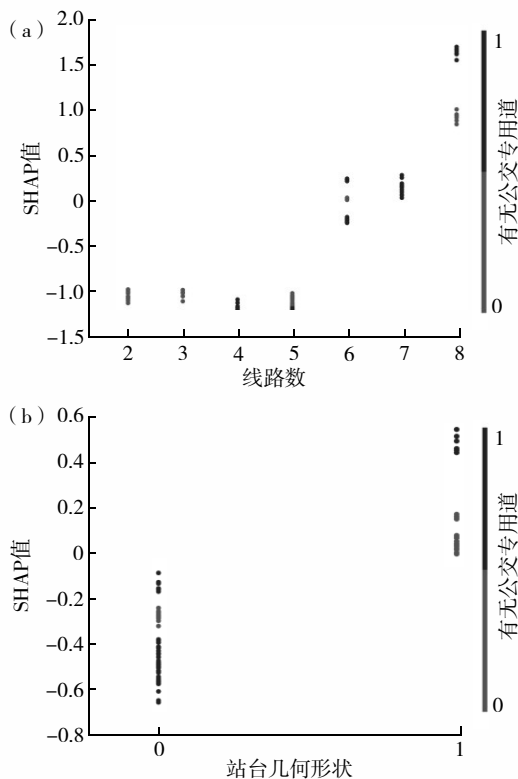


图6 考虑(a)线路数和公交专用道与(b)站台几何形状和公交专用道因素的类型Ⅲ公交站SHAP交互性分析

Fig. 6 SHAP interaction analysis of type III considering (a) number of lines and bus lanes and (b) platform geometry and bus lanes factors.

63.261 s; 类型Ⅱ的平均服务时间和服务车数最少; 类型Ⅲ的服务车数最多为125辆。

2) 与XGBoost、LR、RF、GBDT及KNN模型相比, 所构建SMOTEENN-XGBoost模型的准确率、精确率、召回率及 F_1 分数均为最高, 可实现公交站运行状态的精准预测。

3) 线路数是预测模型最重要的因素, 线路数的增加提高了公交站运行状态属于类型Ⅲ的概率; 有无公交专用道对类型Ⅰ预测模型的影响最大; 泊位数、站台设置方法、站台几何形状、车道数及站台设置位置等因素对预测模型的重要度逐渐减小; 天气类型对服务时间比较长的类型Ⅰ具有一定重要度, 相比于晴天, 雨天时公交站运行状态属于类型Ⅰ的概率更大。

由于数据种类的限制, 本研究内容仍存在一定局限性, 如公交车辆的站点到达与停靠行为会同时受到道路交通流、客流及发车间隔等因素影响, 如何综合各影响因素的作用是下一步研究的重点。

基金项目: 浙江省“尖兵”“领雁”研发攻关计划资助项目(2022C01105); 陕西省社会科学基金资助项目(2022F021)

作者简介: 黄洪滔(1998—), 长安大学硕士研究生。研究方向: 交通运输规划与管理。E-mail: hhtao2022@163.com

引文: 黄洪滔, 肖梅, 刘倩, 等. 基于GPS数据的公交站运行状态分析[J]. 深圳大学学报理工版, 2023, 40(3): 326-334.

参考文献 / References:

- [1] 祁昊, 刘好德, 杨宇航, 等. 基于全样本大数据的公交停靠站时间规律分析[J]. 交通信息与安全, 2020, 38(6): 113-121.
QI Hao, LIU Haode, YANG Yuhang, et al. An analysis of time regularity of urban bus stop dwell based on full sample big data [J]. Journal of Transport Information and Safety, 2020, 38(6): 113-121. (in Chinese)
- [2] BIE Yiming, WANG Yunhao, ZHANG Le. Impact of carriage crowding level on bus dwell time: modelling and analysis [J]. Journal of Advanced Transportation, 2020, 2020: 6530530.
- [3] 中华人民共和国建设部, 中华人民共和国国家计划委员会. 城市公共汽车和无轨电车工程项目建设标准: JB99-104-1996[S].
Ministry of Construction of the People's Republic of China, State Planning Commission of the People's Republic of China. Construction standards for urban bus and trolley bus projects: JB99-104-1996 [S]. (in Chinese)
- [4] 叶娇. 面向城市交通枢纽的多模式组合出行效用及网络承载力评估[D]. 南京: 东南大学, 2020.
YE Jiao. Evaluation of intermodal utility and multi-modal network capacity considering the urban transport hubs [D]. Nanjing: Southeast University, 2020. (in Chinese)
- [5] 朱宏, 谢云侠, 张汝华, 等. 快速公交站点类别及步行吸引范围研究[J]. 交通运输研究, 2015, 1(4): 22-29.
ZHU Hong, XIE Yunxia, ZHANG Ruhua, et al. Classification and walking attractive distance of BRT stations [J]. Transport Research, 2015, 1(4): 22-29. (in Chinese)
- [6] 马壮林, 杨兴, 谭晓伟, 等. 基于客流时间序列的城市轨道交通车站分类[J]. 长安大学学报自然科学版, 2021, 41(6): 113-126.
MA Zhuanglin, YANG Xing, TAN Xiaowei, et al. Classification of urban rail transit stations based on passenger flow time series [J]. Journal of Chang'an University Natural Science Edition, 2021, 41(6): 113-126. (in Chinese)
- [7] 王学勇, 邵勇, 王玉石. 定点停靠公交站的服务效率与停靠能力研究[J]. 交通标准化, 2010(11):

- 202-205.
WANG Xueyong, SHAO Yong, WANG Yushi. Service efficiency and parking capacity of pointing parking bus stop [J]. Transport Standardization, 2010 (11): 202-205. (in Chinese)
- [8] 卞张蕾. 基于车辆GPS数据的公交站运行状态识别及通行能力研究[D]. 扬州: 扬州大学, 2021.
BIAN Zhanglei. Bus-stop operating state recognition and capacity study based on bus GPS data [D]. Yangzhou: Yangzhou University, 2021. (in Chinese)
- [9] 朱宁, 马健霄, 钱思文, 等. 公交站点延误估算方法研究[J]. 森林工程, 2021, 37(1): 87-94.
ZHU Ning, MA Jianxiao, QIAN Siwen, et al. Research on delay estimation methods for bus station [J]. Forest Engineering, 2021, 37(1): 87-94. (in Chinese)
- [10] TAN Jiyuan, LI Zhiheng, LI Li, et al. Berth assignment planning for multi-line bus stops [J]. Journal of Advanced Transportation, 2014, 48(7): 750-765.
- [11] BIAN Bomin, ZHU Ning, LING Shuai, et al. Bus service time estimation model for a curbside bus stop [J]. Transportation Research Part C: Emerging Technologies, 2015, 57: 103-121.
- [12] 柳伍生, 潘自翔, 周骞, 等. 基于运行特性的港湾式公交站延误时间研究[J]. 交通运输系统工程与信息, 2020, 20(4): 150-157.
LIU Wusheng, PAN Zixiang, ZHOU Qian, et al. Delay time model of pull-out bus stops considering bus operational characteristics [J]. Journal of Transportation Systems Engineering and Information Technology, 2020, 20(4): 150-157. (in Chinese)
- [13] 张梦洁. 基于生存分析的公交停靠耗时间影响因素研究[D]. 成都: 西南交通大学, 2019.
ZHANG Mengjie. Research on the influencing factors of the time-consuming of bus stop based on survival analysis [D]. Chengdu: Southwest Jiaotong University, 2019. (in Chinese)
- [14] 罗妍, 王枏, 叶文玲. 基于XGBoost和SHAP的急性肾损伤可解释预测模型[J]. 电子与信息学报, 2022, 44(1): 27-38.
LUO Yan, WANG Cong, YE Wenling. An interpretable prediction model for acute kidney injury based on XGBoost and SHAP [J]. Journal of Electronics & Information Technology, 2022, 44(1): 27-38. (in Chinese)
- [15] 吕能超, 彭凌枫, 吴超仲, 等. 区分冲突类型的路段实时碰撞风险预测模型[J]. 中国公路学报, 2022, 35(1): 93-108.
LÜ Nengchao, PENG Lingfeng, WU Chaozhong, et al. Real-time crash-risk prediction model that distinguishes collision types [J]. China Journal of Highway and Transport, 2022, 35(1): 93-108. (in Chinese)
- [16] 肖梅, 王海明, 边浩毅, 等. 交叉路口公交等待时长分布及影响因素分析[J]. 公路交通科技, 2022, 39(1): 96-105.
XIAO Mei, WANG Haiming, BIAN Haoyi, et al. Analysis on distribution of bus waiting time at intersections and its influencing factors [J]. Journal of Highway and Transportation Research and Development, 2022, 39(1): 96-105. (in Chinese)
- [17] ARHIN S, NOEL E, ANDERSON M F, et al. Optimization of transit total bus stop time models [J]. Journal of Traffic and Transportation Engineering (English Edition), 2016, 3(2): 146-153.
- [18] 何玉林, 黄哲学. 大规模数据集聚类算法的研究进展[J]. 深圳大学学报理工版, 2019, 36(1): 4-17.
HE Yulin, HUANG Zhexue. A review on clustering algorithms for large-scale data sets [J]. Journal of Shenzhen University Science and Engineering, 2019, 36(1): 4-17. (in Chinese)
- [19] 刘通, 付锐, 张名芳, 等. 融合K-means与高斯混合模型的驾驶风格聚类研究[J]. 中国安全科学学报, 2019, 29(12): 40-45.
LIU Tong, FU Rui, ZHANG Mingfang, et al. Study on driving style clustering based on K-means and Gaussian mixture model [J]. China Safety Science Journal, 2019, 29(12): 40-45. (in Chinese)
- [20] 束鹏. 基于可解释机器学习的城市道路交通事故严重程度预测[D]. 西安: 长安大学, 2021.
SHU Kun. Analysis of factors contributing to crash severity on urban road based on explainable machine learning [D]. Xi'an: Chang'an University, 2021. (in Chinese)

【中文责编: 方圆; 英文责编: 淡紫】