一种基于Gradient Boosting的公交车运行时长预测方法

赖永炫[1,2]，杨旭[3]，曹琦[4]，曹辉彬[1,2]，王田[5]，杨帆[6]
1. 厦门大学信息学院，福建 厦门 361005；2. 厦门大学深圳研究院，广东 深圳 518057；
3. 长春公交（集团）有限责任公司，吉林 长春 130000；4. 龙岩烟草工业有限责任公司，福建 龙岩 364000；
5. 华侨大学计算机科学与技术学院，福建 厦门 361021；6. 厦门大学航空与航天学院，福建 厦门 361005

摘要
目前，我国公交公司主要依靠经验丰富的工作人员估计车辆回场时间，进而进行车辆调度，此方式缺乏辅助的预测方法，常常造成较大的误差与错误的调度决策。从公交公司的实际需求出发，提出了一种基于动态特征选择的预测方法R-GBDT。R-GBDT利用特征选择组件和模型调参组件为预测组件提供符合线路特征的特征组合与参数，由融合组件对其他组件的结果进行融合，形成一个用于预测最终时间间隔的框架。结果表明，相对于其他算法，所提方法能大大提高公交运行时长预测的准确度。

关键词
公交调度；到站预测；GBDT

中图分类号：TP399 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019042

A bus running length prediction method based on Gradient Boosting

LAI Yongxuan[1,2], YANG Xu[3], CAO Qi[4], CAO Huibin[1,2], WANG Tian[5], YANG Fan[6]
1. School of Information, Xiamen University, Xiamen 361005, China
2. Shenzhen Research Institute, Xiamen University, Shenzhen 518057, China
3. Changchun Public Transport (Group) Co., Ltd., Changchun 130000, China
4. Longyan Tobacco Industry Co., Ltd., Longyan 364000, China
5. School of Computer Science and Technology, Huaqiqo University, Xiamen 361021, China
6. College of Aeronautics and Astronautics, Xiamen University, Xiamen 361005, China

Abstract
At present, China's public transport companies rely on experienced staff to estimate the return time of vehicles and then conduct vehicle dispatch. This method often results in large errors and wrang decisions due to the lack of auxiliary prediction methods. Based on the actual needs of bus companies, a prediction method R-GBDT based on dynamic feature selection was proposed. The R-GBDT utilizes feature selection components and model parameter adjustment components to provide predictive components with feature combinations and parameters that conform to the line characteristics, then the fusion component combines the results of other components to form a framework for predicting the final time interval. The experimental results from real bus to off-site data show that compared with other algorithms, the method can greatly improve the accuracy of bus transit time prediction.

Key words
bus dispatch, arrival prediction, GBDT

1 引言

近年来，随着人们对公共交通领域关注度的上升，“智能交通”成为重要的研究领域[1-2]。21世纪将是公路交通智能化的世纪，未来将会出现一种一体化的交通综合管理系统，在这一系统中，人们将借助信息采集技术、数据通信技术、电子传感技术等先进的信息技术，实时、准确、高效地采集交通数据，并将其有效地应用于智能交通管理系统，从而使车辆能够依靠智能交通调整至最佳行驶状态[3]。此外，车辆管理人员也能够利用这个系统对道路、交通状况有更全面、实时、准确的掌握。

而随着“公交优先”观念的日益深入，我国各大城市纷纷研发各自的智能交通系统。公交系统作为智能交通领域的重要组成部分，受众广泛且影响深远。近年来，智能公交系统的研究和应用方向主要集中在公交的排班与调度方面，如利用启发式算法（如遗传算法、模拟退火算法、粒子群算法、蚁群算法等）进行公交车排班的研究，包括：路线的制定[4-6]、发车间隔[7-10]的计算、驾驶员排班[11-12]、车辆排班[13-14]。对于公交调度，主要集中在利用机器学习方法进行研究，如车辆到站时间预测[15-18]、公交动态调度[19-20]。然而已有工作主要集中在合理规划线路、合理安排排班计划等方面，并未从辅助公交公司调度车辆的角度出发，研究车辆从起点站到终点站到站时间的预测方法。

本文从公交公司的实际调度需求出发，提出了一种基于机器学习模型的公交车运行时长的预测方法——修正的梯度提升决策树（revised gradient boosting decision tree，R-GBDT）。该方法利用特征选择组件和模型调参组件选择符合线路特征的特征组合与参数，构建基于Gradient Boosting的预测组件；并由融合组件对预测组件的结果进行融合，形成一个用于预测最终运行时长的框架。对真实公交到站、离站数据进行实验的结果表明，相对于其他算法，本文提出的方法能大大地提高公交运行时长预测的准确度。R-GBDT是公交公司辅助决策工具的关键模块，能够辅助调度人员进行科学合理的车辆调度，使得调度人员做出的决策能更好地减少“串车（公交车遇到一起）”和“大间隔（公交车之间离得太远）”现象的发生，进而提高车辆的运行效率和乘客的满意度。

2 相关工作

王麟珠等人[21]提出了一种基于Elman神经网络的公交车辆到站时间预测方法，并通过福州市的公交数据进行验证，实验结果表明，该模型具有更快的收敛速度、更高的预测精度、不易陷入局部最优解等优势。张强等人[22]提出了一种基于时间分段的动态实时预测算法，将一天分为24个等长的时间段，分时段对公交到站时间进行预测。季彦婕等人[23]基于对公交运行特点的分析，提出了一种结合粒子群算法与神经网络算法的粒子群小波神经网络算法，实验结果表明，结合粒子群算法能有效减少预测误差，且对于工作日与周末都有较高的预测精度。罗频捷等人[24]将遗传算法融入神经网络中，进而提高整个神经网络的寻优能力，作者利用该算法对成都某一公交线路进行预测，实验结果表明，该算法具有较高的精确度。杨奕等人[25]提出了基于遗传算法的反向传播（back propagation，BP）神经网络算法，该算法结合遗传算法与BP神经网络，利用遗传算法改进BP神经网络易陷入局部最优的缺陷，通过对合肥市某一公交线路数据的研究和实验得出，该算法有比

较好的预测效果。张昕等人[26]提出了一种基于遗传算法和支持向量机(support vector machine, SVM)的预测模型,该模型考虑时间因素、天气因素与道路因素等的影响,使模型更适用于客运车辆,其利用遗传算法提升SVM的参数寻优效率。该论文对深圳市某一线路公交数据进行实验模拟,结果表明,该算法能够更好地适应道路交通等的变化,具有较好的预测精度。谢芳等人[27]提出了一种基于MapReduce的聚类和神经网络相结合的公交车到站时间预测模型,其首先分析公交车辆的运行特征,然后结合聚类与神经网络模型对车辆数据进行分段预测,最后,基于MapReduce的并行化框架减少算法的计算时间。实验结果表明,该分段模型优于传统的BP神经网络预测模型,具有较高的预测精度和预测速度。

O' Sullivan A等人[28]认为公交车辆的行驶时长是多种非线性组成因素(如乘客数、交通流量、事故、天气条件、路线特征等)相互作用的结果,这些因素造成了预测到站时间的不确定性。他们使用来自现实世界的数据证明了这种不确定影响存在异方差性,于是开发了一个黑盒解决方案,将预测算法融入黑盒处理中,通过预测和观察到达时间之间的误差来估计与预测相关的数据。Sinn M等人[29]提出了一种用于预测到站时间的基于实时全球定位系统(global positioning system, GPS)数据的非参数算法,关键思想是使用内核回归模型表示位置更新与公交车站到达时间之间的依赖关系。实验表明,对于50 min的时间范围,算法的预测误差平均小于10%,明显优于基于K-最近邻(K-nearest neighbor, KNN)算法的线性回归模型的参数方法。Abidin A F等人[30]提出了一种基于卡尔曼滤波(kalman filtering, KF)的公交车到达时间预测模型,该模型使用交通模拟器城市交通仿真(simulation of urban mobility, SUMO)平台模拟真实的道路情景,利用从社交网络获取的信息预测到达时间。Jeong R等人[31]提出了使用自动车辆定位(automatic vehicle location, AVL)数据,并借助人工神经网络(artificial neural network, ANN)模型预测公交车到达时间的方法,结果发现,ANN模型在预测精度方面优于基于历史数据的模型和回归模型。Maiti S等人[32]提出了一种将车辆轨迹和时间戳视为输入特征的基于历史数据的车辆到达时间的实时预测方法。结果表明,他们提出的基于历史数据(historical data, HD)的模型比ANN模型和SVM模型执行速度更快,同时也具有比较高的预测精度。

同样地,本文针对公交线路进行公交车到站时间的预测。不同的是,本文从公交公司的角度出发,以辅助公交公司调度人员进行调度的预测到站时间为中心,帮助解决某线路多辆公交运行过程中发生的“串车”和“大间隔”问题。目前,较少有针对辅助公交公司调度进行的研究,公交公司往往采用人工调度,调度人员仅依靠个人主观经验估计公交车到站时间,常常错误估计车辆到站时间与晚点情况,导致调度结果缺乏科学性和合理性,使得“串车”和“大间隔”问题不能得到解决。因此,加强该技术的研究具有重要的现实意义。本文提出了一种基于Gradient Boosting的公交车到站时间预测方法。不同于以往使用单一算法预测的方法,该方法将公交车停留时长、运行时长和总时长分为单独的模块,为每个模块选择效果最优的模型、特征以及参数,组合形成一套框架,从而提升预测的准确率。

3 数据处理与分析

3.1 原始数据描述

本文使用的原始数据来自于厦门市22路公交车数据,从动态接口获得的原始

数据包括公交车到离站时上传的GPS数据、车辆信息、计划班次信息、天气数据等。具体见**表1~表4**。

3.2 数据预处理

数据预处理流程如**图1**所示。

3.2.1 线路静态特征处理

(1)数据清洗

原始到站、离站数据文件为一段时间内22路上下行收集到的所有车辆进站及出站信息。进行特征提取前，需要先将数据按日期进行分割，将分割后的每一天的数据根据上下行、车辆ID、获取到离站数据的时间进行排列，将同一趟数据划分到同一组中，剔除组内的脏数据。这些脏数据包括重复的到离站记录、同一站的进站时间在该站出站时间之后的数据、到达后一站的时间在到达前一站之前的数据、同一趟的到离站数据大量丢失的数据等。

(2)线路特征处理

线路特征处理包括对线路、方向、车辆及驾驶员信息的处理。其中，线路和方向数据可从到离站数据中直接获取，到离站信息里车辆数据需要先经过一定规则的转换，转换成计划班次信息中可识别的车辆ID，根据此车辆的ID查询驾驶员的ID，将车辆ID和驾驶员ID分别作为区分车辆和驾驶员的特征。

(3)时间特征处理

公交车辆的运行时长通常具有一定的时间规律，比如工作日与周末的区别、节假日与平时的区别、上下班高峰期与平峰期的区别等。为了研究以上时间因素对研究结果的影响，需要根据到离站数据里的时间字段提取车辆的发车日期、发车时间、是否为节假日、是否为工作日等信息。

(4)空间特征处理

空间特征包括站点的经纬度信息，该数据可从到离站数据中直接获得。

表1 公交车到离站的GPS数据

字段名	含义
carID	车辆ID
alarmSign	报警标志
carstate	状态
latitude	纬度
longtitude	经度
altitude	高程
speed	速度
direction	方向
time	时间
odometer	累积里程
routeID	线路ID
upanddown	上下行
siteorder	站点序号
inandout	进出站标志
upnum	上车人数
downnum	下车人数

表2 车辆信息

字段名	含义
busline_id	线路ID
car_id	车辆ID
car_no	车牌号
dev_id	设备ID

表3 计划班次信息

字段名	含义
busline_id	线路ID
stime	发车时间
car_id	车辆ID
car_no	车牌号
driver_id	驾驶员ID
driver_name	驾驶员姓名
updown	上下行

表4 天气数据

字段名	含义
city_id	城市
date	日期
temperature	温度
humidity	湿度
visibility	能见度
weather	天气描述

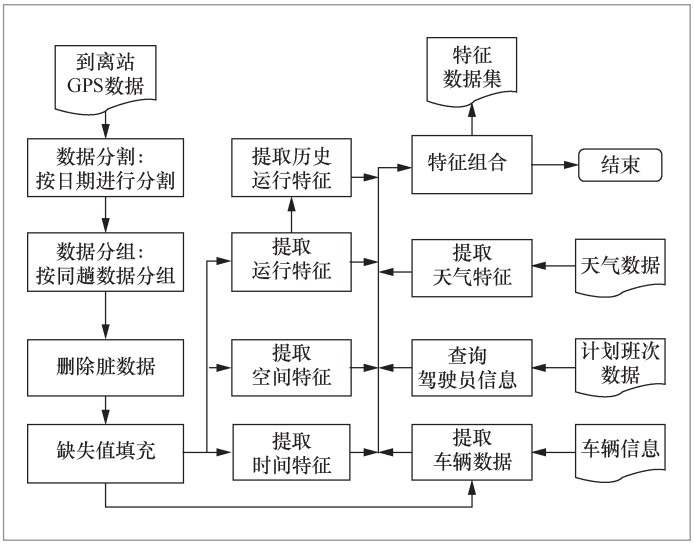


图1 数据预处理流程

3.2.2 线路动态特征处理

线路动态特征主要是线路运行时产生的一些数据特征，比如站间行驶时长和站点停留时长。根据线路运行过程上传的站点、时间等数据，得出车辆运行中在各个站点的停留时长、各个站间的行驶时长。此外，计算经过拼接的最新各个站点的停留时长与行驶时长；近3天、近1周相同时段下的各站点的平均停留时长与平均行驶时长。拼接方式有直接加和、赋予权重再求和等。

其中，最新各个站点的停留时长与行驶时长通常是从起点站开始计算，对于每个站点，把最近一班的车辆在该站的运行时长和停留时长作为该站最新运行时长和停留时长，直至查询到终点站。将拼接形成的一整条数据作为该趟车最新的站点停留和站间行驶特征值。近3天、近1周相同时间段的数据则可通过查询该时间段内相同时间段下各车辆实际运行时的对应数据，并求平均值。

3.2.3 天气特征处理

在获得的原始天气数据中，利用的主要字段为时间、能见度与天气描述。其中，时间需要处理成日期与时间段，天气描述需要将晴、多云、大雾等转换为可利用的数字描述方式。

3.2.4 缺失数据填充

数据收集段的不完整以及部分公交车到离站GPS数据的丢失使得当前车辆、最近时间内、3天内、1周内的站点停留和站间行驶时长可能缺失，需要进行填充。常见的缺失数据处理方式包括删除缺失大量数据的样本、删除缺失大量数据的特征、数据填充。数据填充的方式包括均值填充、中位数填充、众数填充、相同条件下的数据填充等。在缺失数据填充中，对于站点停留和站间行驶时长的缺失，主要利用历史数据相同条件下的均值进行填充，当不存在相同条件下的历史数据时，则用临近班次的数据进行填充。对于天气数据的缺失，主要利用临近小时内的天气状况进行填充。

3.3 公交运行特征及影响因素

通过对各路公交车辆的运行数据进行分析，发现公交车辆的运行特点呈现以下规律。

（1）周期影响

通常，公交线路的运行时间具有长期趋势与周期趋势。从长期角度看，运行时长具有一个稳定变化的趋势，即运行时间在一个较长的时间内，同一线路的总时长在某一个范围内浮动（如**图2**所示）。在一段时间内，每周运行时长的变化趋势相同，具有周期趋势，即周末的运行时长略高于工作日的运行时长。

（2）工作日影响

在一周内，周一至周五的运行时长通常略低于周六至周日的运行时长。这是因

为，周末私家车出行的数量高于工作日，使得道路交通通行状况变差，进而间接地影响了公交车辆的路段行驶速度。此外，周末出行的人数较多，这也在一定程度上增加了站点停靠的时间。

（3）时段影响

如图3所示，即使在同一天，不同发车时刻的公交车辆运行时长也有所不同，主要体现在高峰期的运行时长大于平峰期与低峰期的时长。造成这一现象的主要原因是，高峰期是人们上下班出行的时间段，在这一时间段内，部分主要路段的车流量变大，常常会出现交通拥堵的现象，进而影响公交车辆通行时间以及邻近路段的运行速度。

（4）节假日影响

与工作日和周末在数据表现上的差异类似，节假日的运行时长通常也高于其他时间，这是因为在节假日出行的人数大幅增加，特别是热门旅游景点附近的公交线路，这些线路在站点停留的时间会有所增加。

（5）路段影响

不同线路的行驶路线不同，进而表现出不同线路的总行驶时长的差异。这是因为不同道路的交通运行状况有差异，部分路段的交通流量大，会增加行驶时长。部分路段的上下车乘客较多，也会增加停留时长。

4 基于动态特征选择的时间预测方法R-GBDT

4.1 整体设计

本文提出了一种基于动态特征选择的公交到站时间预测方法。其动态主要体现在对于不同线路、同一线路不同方向的公交车经过特征选择组件（分别选择对该线路该方向站点停留和站间行驶影响较大的特征）选出的特征可能有所差异。这种基于动态特征选择的公交到站时间预测框架主要包括以下模块。

（1）特征选择组件

利用递归特征消除法，根据特征对预

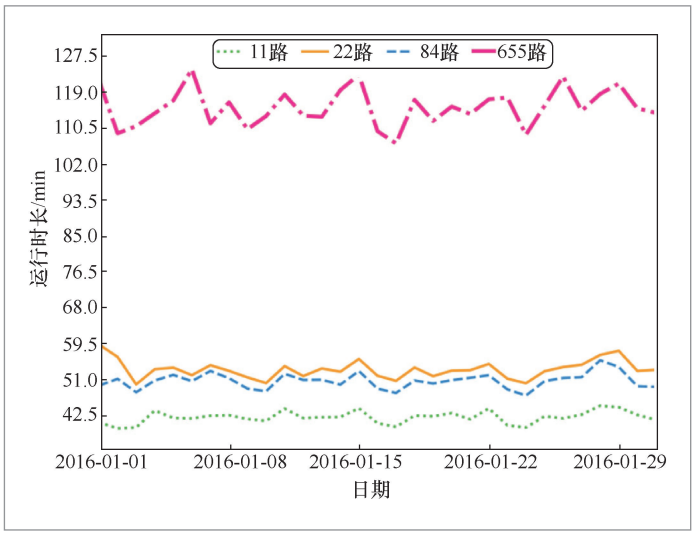


图 2 公交运行时长长期变化

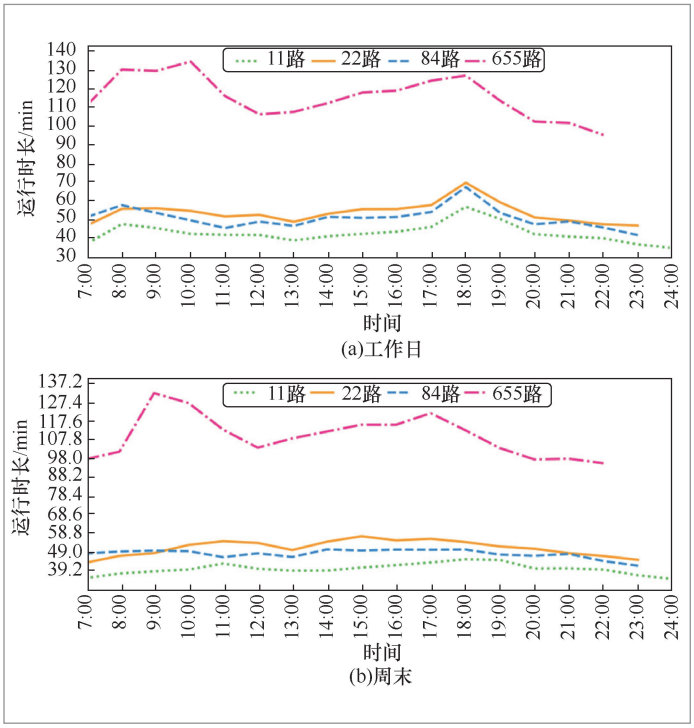


图 3 公交运行时长时段变化

测结果的影响程度，提供特征排名，尝试不同特征数量对预测结果的影响，选择最佳特征数量与特征组合，将其作为预测站点停留和站间行驶的子模型的训练特征。

（2）停留时长预测组件

通过对不同模型（Gradient Boosting、AdaBoost、决策树、岭回归等）的实验尝试，选择采用预测效果最优的GradientBoosting模型对特征选择组件选出的用于停留时长预测的特征进行预测，预测该趟车在所有站点停留的总时长。

（3）行驶时长预测组件

通过对不同模型（Gradient Boosting、AdaBoost、决策树、岭回归等）的实验尝试，选择采用预测效果最优的岭回归模型对特征选择组件选出的用于行驶时长预测的特征进行预测，预测该趟车在所有站间行驶的总时长。

（4）融合组件（总时长预测组件）

通过对不同模型（Gradient Boosting、AdaBoost、决策树、岭回归等）的实验尝试，选择采用预测效果最优的Gradient Boosting模型将预测的站点停留时长和站间行驶时长融合其他特征，预测该趟车从起点站到终点站所需的总时长。

（5）模型调参组件

用于模型训练的特征是根据特征选择组件选择的，不同线路、不同方向选出的特征也有所不同，且即使不同线路或方向选择的特征是相同的，它们对结果的影响程度可能也有差异。这就使得用于模型训练的最佳参数有所差异。基于这样的情况，模型调参组件的主要工作是针对模型与特征选择合适的参数。

图4展示了框架的整体流程。首先将数据处理得到的待选特征集输入特征选择组件中，特征组件根据目标任务，选择停留时长特征或行驶时长特征，将选出的最优特征组合传入模型调参组件中。模型调参组件根据目标与特征组件选择的特征，选择合适的参数组合，并将特征与参数选择结果传入停留时长预测组件或者行驶时长预测组件中。停留时长预测组件根据特征与参数训练模型，并在后续预测时为融合组件提供预测停留时长。行驶时长预测组件根据特征与参数训练模型，并在后续预测时为融合组件提供预测行驶时长。到这一阶段，将获得两个新特征：模型组件预测出的停留时长与行驶时长，用于后续融合组件预测总时长。融合组件在进行预测时，还需要特征选择组件和模型调参组件为融合组件提供用于预测总时长的特征与参数，然后进行总时长的预测。接下来，本文将进一步对上述几个模型的组件进行阐述。

4.2 详细设计

（1）特征选择组件

特征选择组件对公交车运行时长（停留时长和行驶时长）的影响因素进行分析，待选择的特征因素见**表5**。

（2）停留时长预测组件

停留时长预测组件主要根据特征选择组件选出的用于预测站点停留时长的特征，利用Gradient Boosting模型对站点停留时长进行预测。

Gradient Boosting实际上是一种Boosting算法，其通过多次迭代，根据一定的规则，不断改进模型误差，提高模型的准确率。常见的Boosting算法包括AdaBoost与Gradient Boosting。使用集成算法的优势在于：第一，由于模型之间的差异性，往往模型在决策时会出现不同的错误，将多个模型结合能够得到更好的预测结果，做出更合理的决策；第二，当数据集较小时，采用有放回的方式随机抽取N个集合用于训练，再把多个模型进行集

成，可以得到更好的结果；第三，当模型的决策边界过于复杂时，往往多个模型的集成结果会优于单一模型的集成结果。

（3）行驶时长预测组件

行驶时长预测组件主要根据特征选择组件选出的用于预测站间行驶时长的特征，利用岭回归模型对站间行驶时长进行预测。

进行回归预测的方法有很多种，如线性回归与岭回归。线性回归假设样本满足式（1），其中将预测结果表示为n个变量（x_1~x_n）的线性相关（其中，x_0=1，θ_0~θ_n为待定参数），使用普通最小二乘法使其损失函数（式（2））最小，其中，$f_\theta(x^i)-\hat{y}$是第i个预测值与真实值的差值，共有m次测试，对每次预测的差值平方求和。由于差值平方后被放大，这使得在线性回归模型中，异常值对误差的影响较大，即模型对异常值较为敏感。在现实生活中收集到的数据往往会因为各种因素导致异常值的存在，故利用线性回归进行预测时效果往往不尽人意。

$$f_\theta(x)=\theta_0x_0+\theta_1x_1+\cdots+\theta_nx_n \qquad (1)$$

$$L=\frac{1}{2m}\sum_{i=1}^{m}(f_\theta(x^i)-\hat{y})^2 \qquad (2)$$

为了减少异常值对模型的影响，岭回

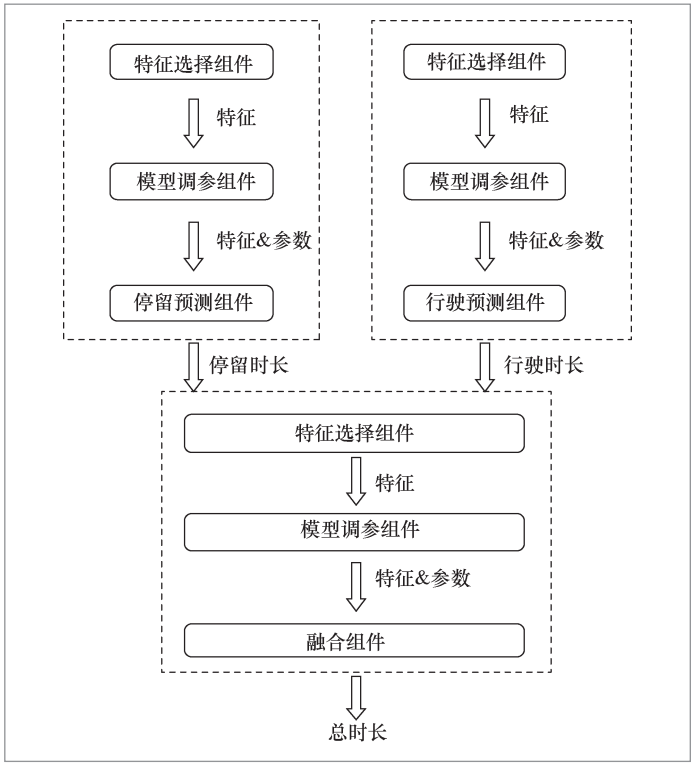


图 4 框架总体流程

表 5 待选特征集

类型	种类	特征名	维数
静态特征	线路特征	线路	1
		方向	1
		预定的行驶距离	1
		计划行驶的站数	1
	时间特征	星期几	1
		时段	1
		段内分组	1
		是否为节假日	1
		是否为工作日	1
	驾驶员特征	驾驶员编号	1
	周期运行特征	该时段下发车的车辆的各个站点的1周内平均运行时长、停留时长	2×站点数量
	近期运行特征	该时段下发车的车辆的各个站点的3天内平均运行时长、停留时长	2×站点数量
动态特征	路段特征	最新的各个站点的行驶时长	2×站点数量
		最新的各个站点的停留时长	2×站点数量

归引入正则化项作为阈值，见式（3）。其中，λ是正则化参数，作用是控制$\theta_j(j=1,\cdots,n)$的取值平衡，以减弱异常值对模型的影响程度。此外，由于自变量之间往往存在多重共现性，即使这种多重共现性并不影响最小二乘估计量的无偏性和最小方差性，但是，利用这种方式求得的方差并不一定小。事实上，可以找一个有偏估计量，虽然这种有偏估计量有微小的偏差，但其精度却比无偏估计量高。此外，岭回归通常用于自变量间存在多重共线性的情况，由于各时刻和各站点之间的停留时长可能存在时间与空间上的相关关系，因此选用岭回归模型作为站间行驶时长的预测模型。

$$L=\frac{1}{2m}\sum_{i=1}^{m}(f_\theta(x^i)-\hat{y})^2+\lambda\sum_{j=1}^{n}\theta_j^2 \quad (3)$$

（4）融合组件

实际上，由于各特征因素间的复杂影响，通过行驶时长预测组件和停留时长预测组件预测得到的站间行驶和站点停留时长并不能直接加和作为最终的预测结果，还需要融合组件结合其他特征进行结果融合，以求得最终从起点到终点总时长的预测结果。由于不同线路之间的差异，各个站点对最终结果的影响程度也有所不同，所以融合组件在进行融合时，同样需要特征选择组件在已有特征和两个预测特征之间进行特征组合的选择，以求得体现该线路特点的特征组合。

（5）模型调参组件

动态特征选择的公交到站时间预测方法，其动态性不仅仅体现在特征选择随线路和方向的变化而动态变化，而且，对于同一个训练模型来说，由于不同的特征组合与训练数据表现出的差异，其最优参数也相应体现出差异，故模型调参组件主要是根据输入的特征数据选择最优的参数组合。

5 实验与分析

5.1 实验过程与结果

实验基于厦门市所有22路公交车自2018年12月至2019年2月的到离站数据、计划班次信息、车辆信息以及对应时间的厦门天气状况等数据进行，使用Python3.7进行数据处理与算法编写，实验中使用sklearn0.18.1库中的模型接口进行部分算法实现。

实验对于公交站点停留时长、站间行驶时长以及总时长3个模块均使用不同模型（Gradient Boosting、AdaBoost、决策树、岭回归等）进行实验，在每种模型基础上使用不同特征、不同参数组合，最后为每个模块选择出预测效果最优的模型（包括模型的最佳特征与参数组合）进行组合，形成最终框架。

5.1.1 停留时长预测

（1）特征选择

在特征选择阶段，笔者发现，当组件选取**表6**所示的18个特征时，在Gradient Boosting、AdaBoost、决策树、岭回归上表现相对较优。

从选择的特征可以看出，是否为工作日、是否为节假日、天气是影响停留时长的因素之一；此外，线路特征的表现较为复杂，停留时长的变化主要体现在个别站点上，且不同站点停留时长的变化具有不同的时间特征，比如第6、7、8、13站点等的客流量具有周期趋势，第15、22站点受3天内变化的影响较大，第11、16站点受最新情况的影响较大。

接着在实验中，对比了不同模型对停留时长预测的准确率和误差率随特征数

量的变化（如**图5**、**图6**所示），其中，准确率表示预测值与真实值的差值的绝对值在一定范围内（本文设定为5 min）的比率，误差率=（预测值与真实值的差值的绝对值）/真实值，并对比了随着特征数量的增加，同一训练集训练时长的变化（如**图7**所示）。笔者发现，利用同一训练集，在不同的模型下，各模型的准确率与误差率呈现一个波动变化的过程，并且，随着特征数量的增多，训练模型消耗的时间增加。此外，由于影响公交车辆在站点停留时长的因素很多，各种因素间相互作用的复杂程度不同，且不同线路具有不同的线路特征，因此，在进行训练前，根据待选特征集选择出体现线路特征的特征集合是非常有必要的。一方面，可以针对不同线路选取更有效的线路特征；另一方面，通过降低特征数量，可以缩减训练时间。

（2）模型调参

由于不同的线路经过特征选择组件会筛选出不同的特征，因此同一个模型的最优参数组合也不同，故需要经过模型调参组件找到最优的参数。

首先对比了4种情况下模型的误差率，**表7**中列出了GBDT算法中的各项参数。其中，loss表示采用的损失函数（huber表示Huber损失，ls表示均方差，lad表示绝对损失，quantile表示分位数损失），learning_rate表示模型的学习率，n_estimators表示学习器最大的迭代次数，subsample表示控制模型子采样的比例，alpha表示需要指定分位数的值。

对于Gradient Boosting模型，其主要参数包括loss、learning_rate、n_estimators、subsample、alpha。通过模型调参组件，发现对于停留时长预测组件来说，最佳的参数组合为：loss设置为huber、learning_rate设置为0.1、n_estimators设置为600、subsample设置为0.7、alpha设置为0.2。在这样的参数组合下，模型的误差率为7.98%（如**图8**所示），5 min的误差范围内准确率为98.45%。下面在上述基础上，分别采用不同的损失函数，固定其他变量，调整某一变量，进行误差百分比的对比，结果如下。

表 6　停留时长特征

字段名	含义
weatherid	天气
weekdayid	星期几
isworkday	是否是工作日
isholiday	是否是节假日
hourid	小时
groupid	段内分组
time_group	时间分组
week_staytime_6	第6站1周内同时段平均停留
week_staytime_7	第7站1周内同时段平均停留
week_staytime_8	第8站1周内同时段平均停留
week_staytime_13	第13站1周内同时段平均停留
week_staytime_19	第19站1周内同时段平均停留
week_staytime_22	第22站1周内同时段平均停留
thrday_staytime_15	第15站3天内同时段平均停留
thrday_staytime_22	第22站3天内同时段平均停留
current_every_staytime_11	第11站最新的停留时长
current_every_staytime_16	第16站最新的停留时长
sum_week_staytime	1周内平均总停留时长

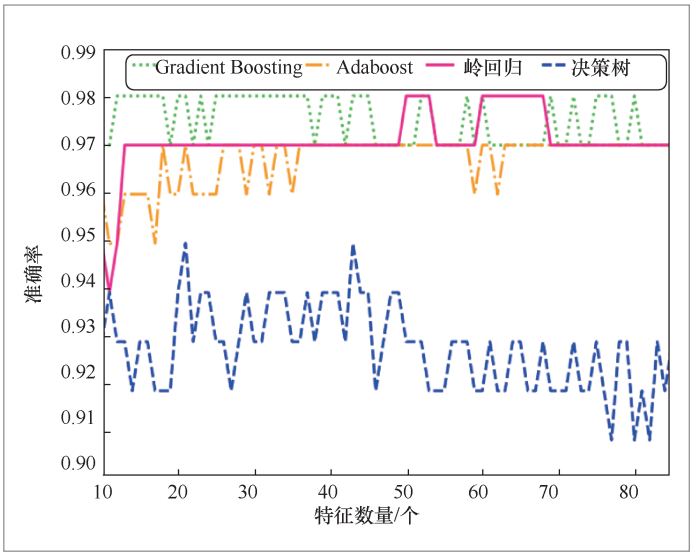


图 5　不同模型对停留时长预测的准确率随特征数量的变化

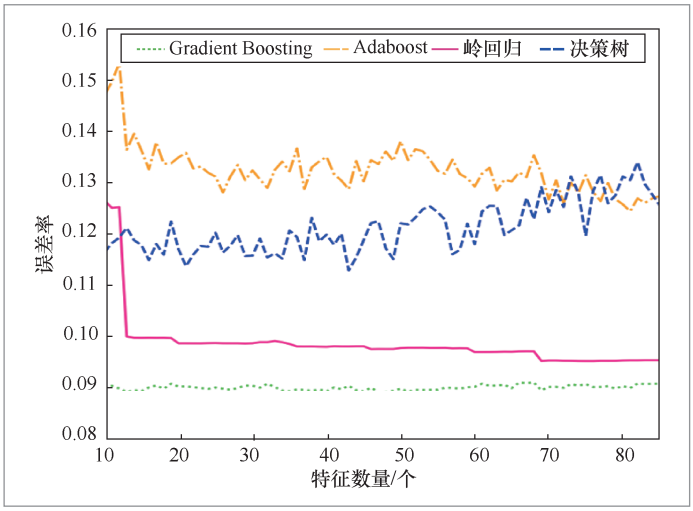


图 6 不同模型对停留时间预测的误差率随特征数量的变化

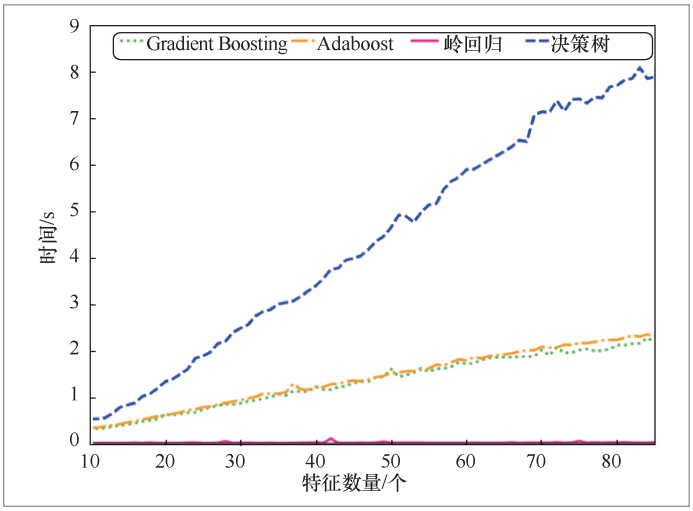


图 7 不同模型对停留时间预测的训练时长随特征数量的变化

表 7 参数组合情况

方法	loss	learning_rate	n_estimators	subsample	alpha
1	huber	0.1	600	0.7	0.2
2	ls	0.4	550	0.5	0.4
3	lad	0.5	550	0.1	0.1
4	quantile	0.45	400	0.1	0.2

如**图9**所示，alpha对loss为quantile时的误差率影响较大，而对其他几个损失函数的影响较小。通常，alpha只在loss为quantile时需要进行调整。

如**图10**所示，随着subsample的变化，模型的误差率在不同的loss下，都会有所浮动。subsample是控制模型子采样的比例，此处的子采样不同于随机森林中有放回的采样。通常，当该值为1时，表示不进行采样，使用全部样本；当该值小于1时，表示要进行采样，使用部分样本。当该值小于1时，可以在一定程度上防止过拟合，但是若该值过小，也会增加样本拟合的偏差，导致欠拟合，因此，该值最优的取值范围为[0.5,0.8]。

n_estimators指的是学习器最大的迭代次数，当该值过小时，容易导致模型欠拟合现象的发生，当该值过大时容易导致模型过拟合现象的发生，模型的预测误差率随着n_estimators变化的情况如**图11**所示。

learning_rate指的是模型的学习率，即模型学习的快慢，当该值较大时，可能导致在训练过程中跳过最优点，该值较小时，模型学习得较慢。如**图12**所示，随着学习率的增加，模型的误差率呈上升趋势。

(3) 小结

特征选择组件为线路选择合适的特征，模型调参组件为线路选择最佳的模型参数组合，停留时长预测组件将根据选出的特征和参数组合进行模型训练，给出停留时长预测结果，该结果作为融合组件中所用的停留时长预测特征。通过前面的实验对比发现，Gradient Boosting在准确率与误差率上均优于其他模型，故选用Gradient Boosting对公交站点停留时长进行预测。

5.1.2 行驶时长预测

(1) 特征选择

在特征选择阶段，笔者发现当组件选取**表8**所示的16个特征时，在Gradient Boosting、AdaBoost、决策树、岭回归上表现相对较优。

从上述实验结果选择的的特征可以看

出，与停留时长模块中被选出的停留时长特征不同，是否为节假日、天气等外部特征并没有被选出，这是因为对于22路公交车来说，其行驶时长不受这些因素的影响。厦门市作为热门旅游城市，部分路段在某一时段内交通状况基本稳定在某一范围，不受节假日和天气等因素的影响。对于22路公交车的行驶时长来说，其部分站点（如第6站点、第9站点等）间的行驶时长受周期因素影响，第1、4、17站点等受近3天的数据表现影响，第8、21站点等受最近时段因素影响，第2站点同时受1周和近3天时段表现的影响。

实验中对比了不同特征数量在各个模型上预测行驶时长的准确率与误差率（如**图13**、**图14**所示），并对比了随着特征数量的增加，同一训练集训练时长的变化（如**图15**所示）。笔者发现，利用同一训练集，在不同的模型（包括Gradient Boosting、AdaBoost、岭回归、决策树）下，各模型的准确率与误差率依旧呈现一个波动变化的过程，也就是说，特征数量并不是越多越好，相反，随着特征数量的增多，模型的训练时长会增多。因此，进行特征数量的选择对于行驶时长预测来说，同样是有必要的。

（2）模型调参

由于不同的线路通过特征选择组件可能会筛选出不同的特征，因此同一个模型的最优参数组合可能也有所不同，故需要通过模型调参组件找到最优的参数。经过多次实验发现，当alpha为90时，岭回归模型的表现较优，误差率为6.69%，5 min的误差范围内准确率为86.17%。

（3）小结

通过特征选择组件和模型调参组件为线路选择合适的特征和模型参数组合，行驶时长预测组件将根据特征和参数进行模型训练，给出行驶时长预测结果，作为融合组件中所用的行驶时长预测特征。

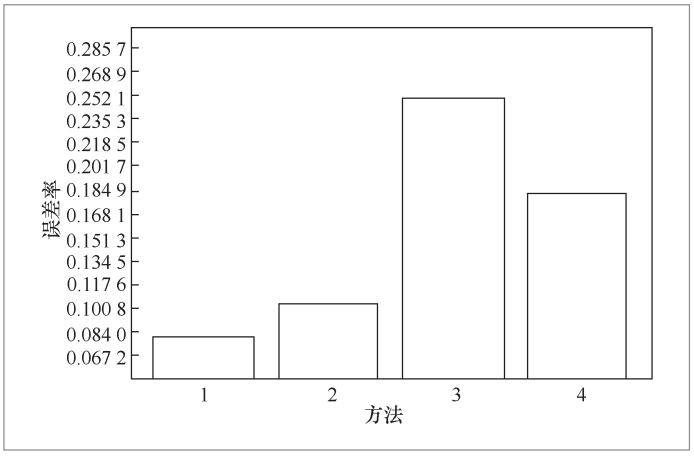


图 8　不同参数组合对停留时长预测的误差率

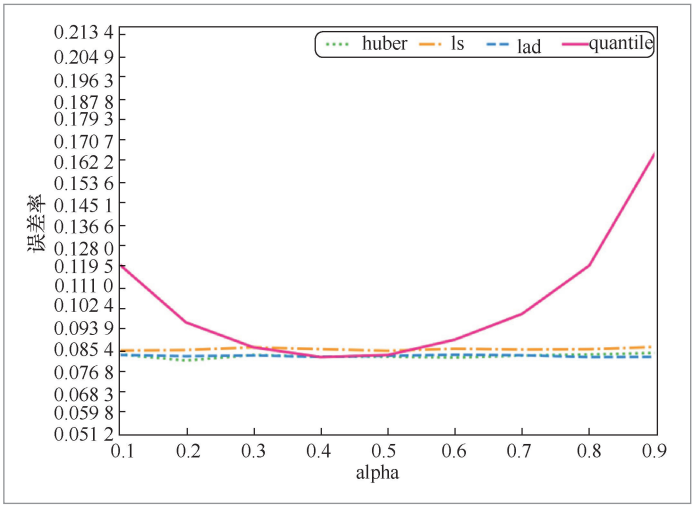


图 9　模型误差率随 alpha 的变化

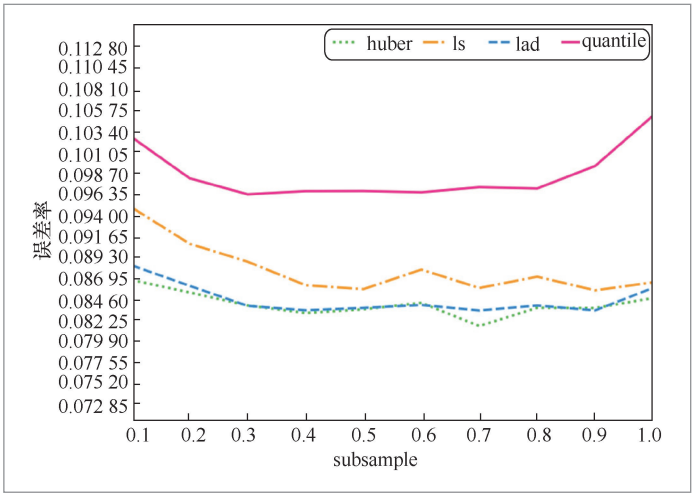


图 10　模型误差率随 subsample 的变化

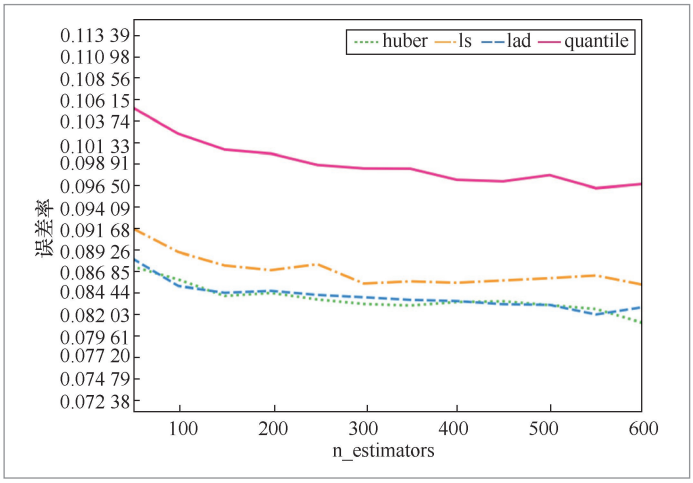


图 11 模型误差率随 n_estimators 的变化

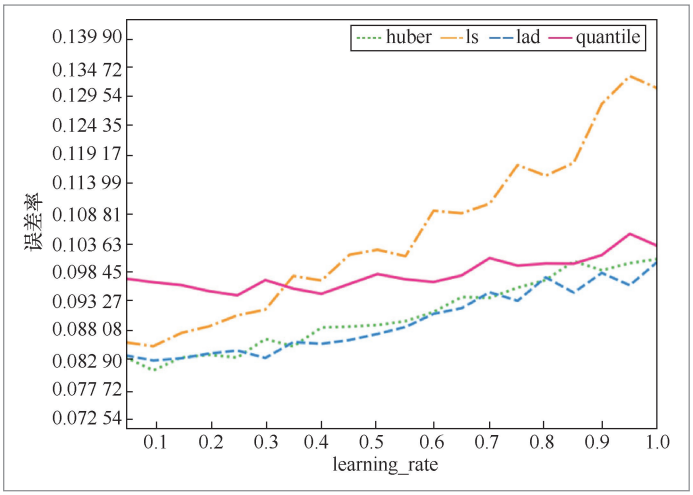


图 12 模型对停留时长预测的误差率随 learning_rate 的变化

表 8 行驶时长特征

字段名	含义
weekdayid	星期几
time_group	时间分组
week_runtime_2	第2站1周内同时段平均站间行驶
week_runtime_6	第6站1周内同时段平均站间行驶
week_runtime_9	第9站1周内同时段平均站间行驶
week_runtime_20	第20站1周内同时段平均站间行驶
thrday_runtime_1	第1站3天内同时段平均站间行驶
thrday_runtime_2	第2站3天内同时段平均站间行驶
thrday_runtime_4	第4站3天内同时段平均站间行驶
thrday_runtime_17	第17站3天内同时段平均站间行驶
thrday_runtime_23	第23站3天内同时段平均站间行驶
current_every_runtime_8	第8站最新的行驶时长
current_every_runtime_21	第21站最新的行驶时长
sum_week_runtime	1周同时段平均行驶时长
sum_thrday_runtime	3天同时段平均行驶时长
sum_current_runtime	最新行驶时长

经过多种模型的对比发现，岭回归模型在性能上（准确率与误差百分比）的表现与Gradient Boosting基本相同，但随着特征数量的增长，其时间的增长比Gradient Boosting慢，因此，选择综合表现较好的岭回归模型作为预测行驶时长的主要模型。

5.1.3 总时长预测

（1）特征选择

在总时长预测模块中，特征选择组件将在影响站间行驶和站点停留的所有特征字段中进行特征筛选。笔者发现，当组件选取**表9**所示的12个特征时，在Gradient Boosting、AdaBoost、决策树、岭回归上表现相对较优。

从上述实验结果选择的特征可以看出，停留时长预测组件和行驶时长预测组件产生的预测结果对预测总时长是有帮助的，因为在不同特征数量下，特征组件都选出了预测的两个特征。此外，预测组件还为融合组件选择了一些与运行或停留有关的特征，这可能是因为影响总时长的因素的复杂性，使得部分站点的情况是影响最终目标值的重要因素。

在实验中还对比了不同特征数量在各个模型上预测总时长的准确率与误差率（如**图16**、**图17**所示），并对比了随着特征数量的增加，在同一训练集训练时长的变化（如**图18**所示）。利用同一训练集，在不同的模型（包括Gradient Boosting、AdaBoost、决策树、岭回归）下，各模型的准确率与误差率呈现波动变化的过程，且随着特征数量的增多，模型的训练时长会增多，因此，进行特征数量的选择对于总时长预测来说，同样是有必要的。

（2）模型调参

由于不同的线路经过特征选择组件可

能会筛选出不同的特征，因此同一个模型的最优参数组合也不同，故需要经过模型调参组件找到最优的参数。对于总时长的预测，选择Gradient Boosting作为预测模型，该模型的参数组合见**表10**。

经过多次实验发现，当loss设置为lad、learning_rate设置为0.1、n_estimators设置为250、subsample设置为0.8、alpha设置为0.6时，数据表现最优，其中误差率为5.55%（如**图19**所示），5 min的误差范围内准确率为81.07%。

下面针对最优的组合，变化其中的值，观察模型准确率的变化。

对于learning_rate，如**图20**所示，模型的准确率整体上随着learning_rate的增加而下降。

对于n_estimators，如**图21**所示，模型的准确率在n_estimators取值为200~300时得到最优。

对于subsample，如**图22**所示，模型的准确率在subsample取值为0.8左右时达到最优。

对于alpha，如**图23**所示，模型的准确率在alpha取值为0.6左右时达到最优。

（3）小结

通过特征选择组件和模型调参组件为线路选择合适的特征和模型参数组合，融合组件根据特征和参数进行模型训练，将行驶时长预测组件和停留时长预测组件的预测结果与其他特征进行融合，给出最终预测的行驶时长。经过多种模型的对比发现，Gradient Boosting在特征数量较少的时候，其模型在准确率和误差率上的表现较好，因此，在融合组件中，选择Gradient Boosting作为预测总时长的算法。**图24**为模型拟合总时长的曲线，其中，横坐标表示预测的第x条记录，纵坐标表示总时长的预测值（红色）和真实值（数色），**表11**为对3个模块组件分别各自统计

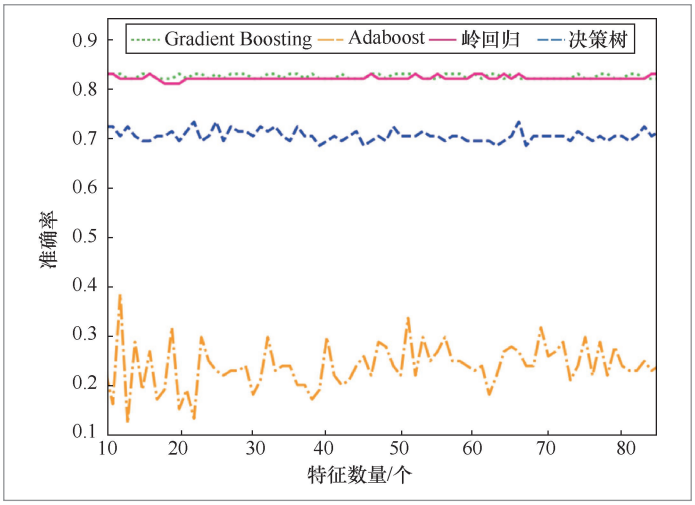


图 13　不同模型对行驶时长预测的准确率随特征数量的变化

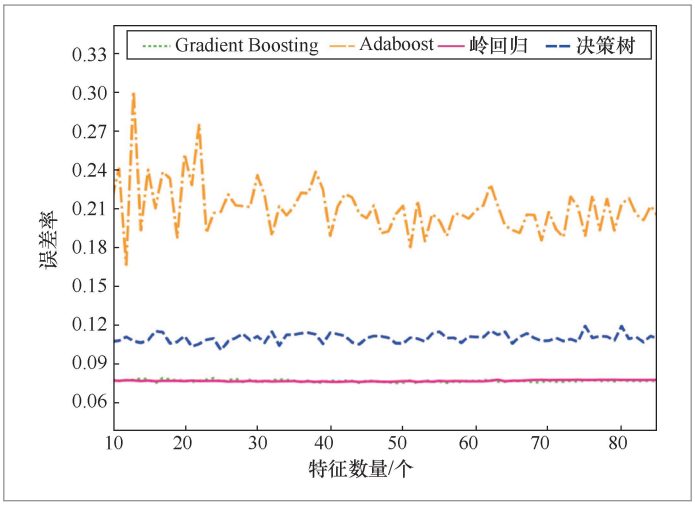


图 14　不同模型对行驶时长预测的误差率随特征数量的变化

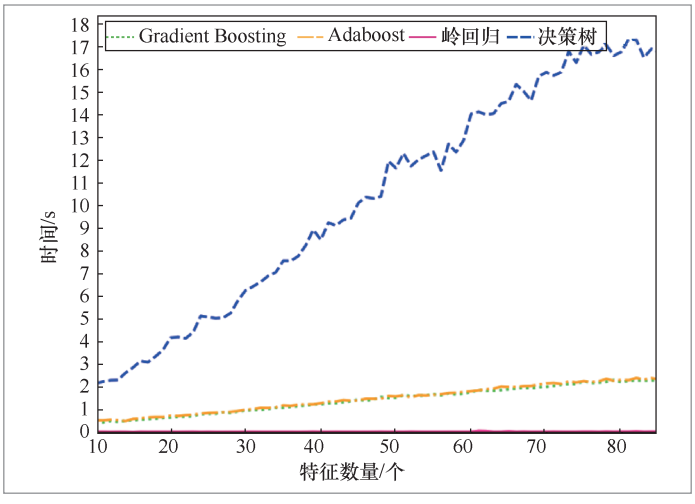


图 15　不同模型对行驶时长预测的训练时长随特征数量的变化

表 9 总时长特征

字段名	含义
predict_runtime	预测的总站间行驶时长
predict_staytime	预测的总站点停留时长
time_group	时段分组
current_every_staytime_7	最新的第7站站点停留时长
current_every_staytime_10	最新的第10站站点停留时长
thrday_runtime_17	3天内第7站的平均站间运行时长
week_staytime_4	1周内第4站的平均站点停留时长
week_staytime_11	1周内第11站的平均站点停留时长
week_staytime_14	1周内第14站的平均站点停留时长
thrday_staytime_1	3天内第1站的平均站点停留时长
thrday_staytime_18	3天内第18站的平均站点停留时长
sum_current_time	最新的总运行时长

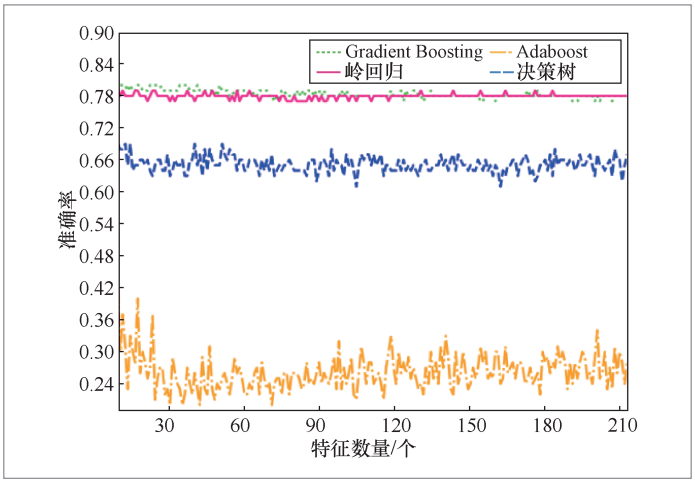


图 16 不同模型对总时长预测的准确率随特征数量的变化

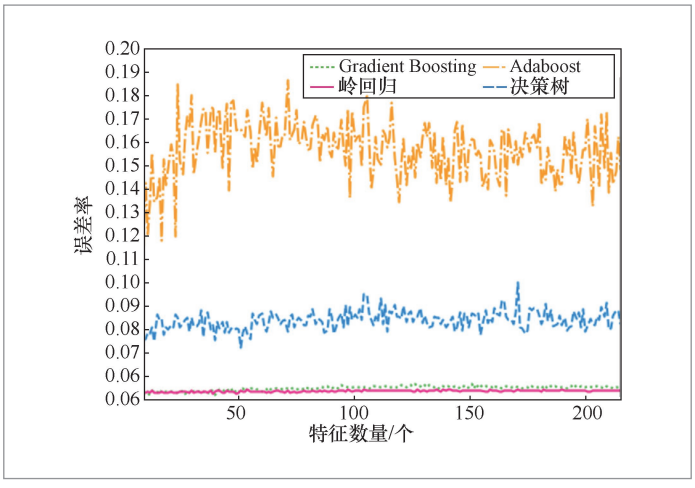


图 17 不同模型对总时长预测的误差率随特征数量的变化

的准确率与误差率。

5.2 对比实验

在已有的预测算法中，关于公交预测车辆到达终点站的研究较少，故将本文提出的R-GBDT和典型的几个传统机器学习算法进行对比。实验结果见**表12**。

其中，各个算法如下。

- R-GBDT：本论文提出的方法，即利用特征选择组件与模型调参组件选择合理的特征组合与参数。由停留时长预测组件与行驶时长预测组件预测停留时间与行驶时间。由融合组件将上述两个组件的预测结果与其他特征进行融合，形成一个框架。
- AdaBoost[33]：机器学习中的一种集成的方法，使用的特征为R-GBDT中未经特征组件选择的特征。
- 决策树[34]：使用的特征为R-GBDT中未经特征组件选择的特征。
- SVM[35]：支持向量机，使用的特征为R-GBDT中未经特征组件选择的特征。
- TTP（total predict）：不引入行驶和停留运行特征时利用GBDT进行总时长预测。
- DS（direct sum）：直接加和R-GBDT中停留时长预测组件与行驶时长预测组件的预测结果。

从实验结果看，本文提出的基于动态特征选择的预测方法R-GBDT的误差率相对于其他算法更低，实验结果分析如下。

- 对比R-GBDT与其他机器学习模型（AdaBoost、决策树、SVM）发现，其误差率与准确率均优于其他模型。这是由于R-GBDT使用的GBDT方法是一种集成学习的方法，类似于一种投票的机制，能够很好地纠正单一模型的错误。从另一个角度看，其他机器学习模型均是采用直接预

测最终结果的方式，R-GBDT采用的是类似于分类预测再累加的方式，对于实验结果，融合结果并没有出现分段预测模型中常出现的误差累加现象，这也说明该方法在降低分段预测的误差累加上有一定优势。

- 对比R-GBDT与DSR发现，二者都是在传统的机器学习模型上进行的，不同之处在于是否使用细化的停留和行驶的相关运行特征。从实验结果上看，引入细化的停留和行驶相关特征对预测最终结果是有一定帮助的。
- 对比R-GBDT与DS发现，融合组件的结果比直接加和停留时长预测组件与行驶时长预测组件的结果略微好一些。这在一定程度上也反映了对于总时长，其可能存在除停留和行驶外的其他影响因素。

6 结束语

本文提出了一种利用机器学习模型的动态特征选择预测方法R-GBDT，该方法将特征选择组件、停留时长预测组件、行驶时长预测组件、融合组件以及模型调参组件连接组合形成一个框架，将Gradient Boosting和岭回归预测出的站点停留时长和站间行驶时长作为特征，融合其他线路特征，预测从起点站到终点站的总时间。通过对真实公交到离站数据进行实验得到的结果表明，相对于其他算法，本文提出的方法能大大提高公交运行时长的预测准确度。

对于未来工作，由于不同线路、不同方向、不同站点和路段对最终结果的影响程度不同，且存在多种复杂因素的影响，故需要根据线路和方向进行具有线路特点的特征选择，从而使模型更合理且更具有解释性。

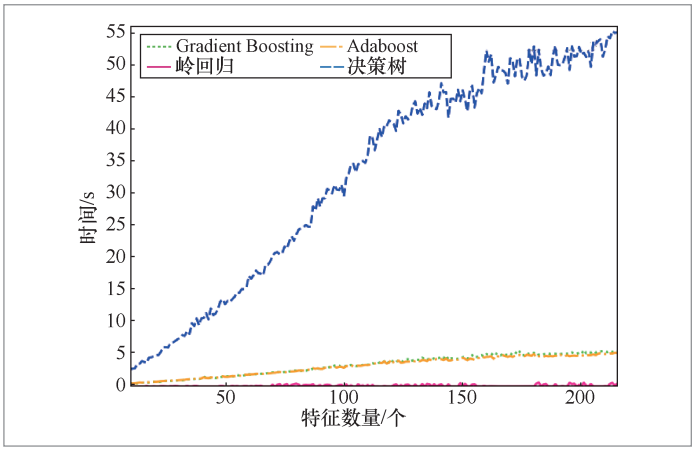


图 18 不同模型对总时长预测的训练时长随特征数量的变化

表 10 参数组合情况

方法	loss	learning_rate	n_estimators	subsample	alpha
1	huber	0.35	550	0.1	0.6
2	ls	0.35	600	0.1	0.4
3	lad	0.1	250	0.8	0.6
4	quantile	0.15	250	0.2	0.9

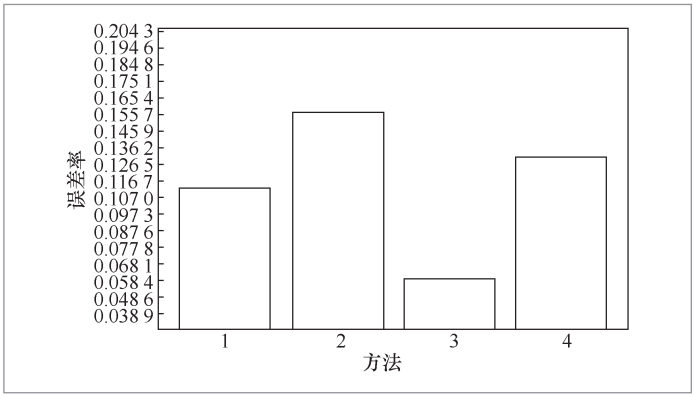


图 19 不同参数组合对总时长预测的误差率

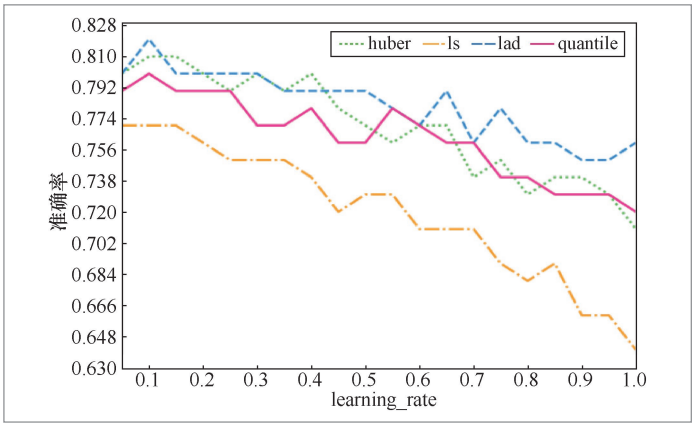


图 20 不同模型对总时长预测的准确率随 learning_rate 的变化

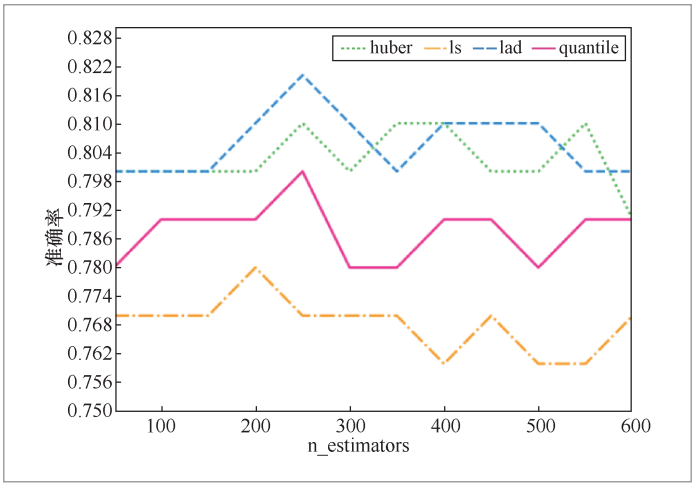


图 21 模型对总时长预测的准确率随 n_estimators 的变化

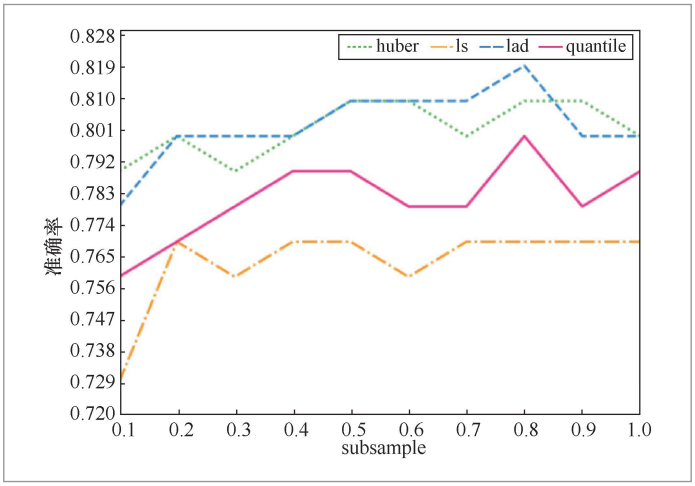


图 22 模型对总时长预测的准确率随 subsample 的变化

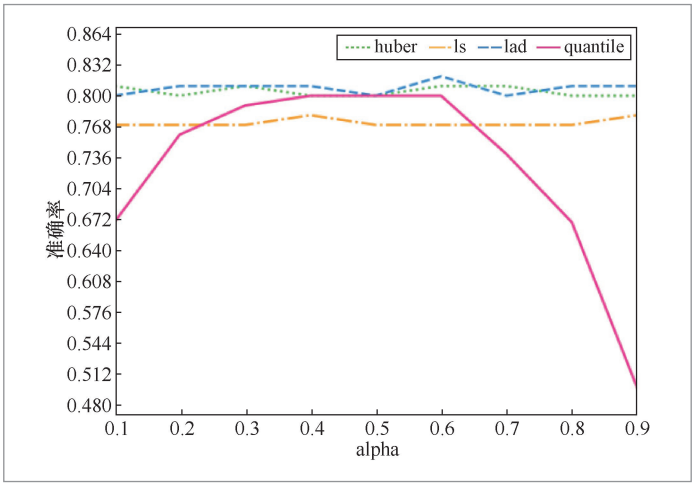


图 23 模型对总时长预测的准确率随 alpha 的变化

参考文献:

[1] 陆化普, 李瑞敏. 城市智能交通系统的发展现状与趋势[J]. 工程研究-跨学科视野中的工程, 2014(1): 6-19.
LU H P, LI R M. Developing trend of ITS and strategy suggestions[J]. Journal of Engineering Studies, 2014(1): 6-19.

[2] LI Y G, FU K, WANG Z, et al. Multi-task representation learning for travel time estimation[C]//The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19-23, 2018, London, UK. New York: ACM Press, 2018: 1695-1704.

[3] 曹宏, 何凤娟. 深化城市轨道交通智能化综合管理研究[J]. 城市轨道交通研究, 2014, 17(1): 74-77.
CAO H, HE F J. Development of smart integrated information management in urban mass transit[J]. Urban Mass Transit, 2014, 17(1): 74-77.

[4] CHIEN S, YANG Z. Optimal feeder bus routes on irregular street networks[J]. Journal of Advanced Transportation, 2010, 34(2): 213-248.

[5] WU J Q, SONG R, XU W T, et al. Transit network optimization for feeder bus of BRT based on genetic algorithm[C]// 2013 4th International Conference on Digital Manufacturing & Automation, June 29-30, 2013, Qingdao, China. Piscataway: IEEE Press, 2013.

[6] NEWELL G F. Some issues relating to the optimal design of bus routes[J]. Transportation Science, 1979, 13(1): 20-35.

[7] GAO L, HU X, WANG W, et al. An optimal bus departure plan under the condition of partially overlapping routes[C]//American Society of Civil Engineers 14th COTA International Conference of Transportation Professionals, July 4-7, 2014, Changsha, China. [S.l.:s.n.], 2014: 1153-1163.

[8] CEDER A. Methods for creating bus timetables[J]. Transportation Research, 1987, 21(1): 59-83.
[9] CEDER A, GOLANY B, TAL O. Creating bus timetables with maximal synchronization[J]. Transportation Research, 2001, 35(10): 913-928.
[10] CEDER A. Bus frequency determination using passenger count data[J]. Transportation Research, 1984, 18(5-6): 439-453.
[11] CHEN Q, LI C. An approach to bus-driver scheduling problem[C]// 2010 2nd WRI Global Congress on Intelligent Systems, December 16-17, 2010, Wuhan, China. Piscataway: IEEE Press, 2010.
[12] KWAN R S K, WREN A, KWAN A S K. Hybrid genetic algorithms for scheduling bus and train drivers[C]//The 2000 Congress on Evolutionary Computation, July 16-19, 2000, La Jolla, USA. Piscataway: IEEE Press, 2000.
[13] KLIEWER N, MELLOULI T, SUHL L. A time-space network based exact optimization model for multi-depot bus scheduling[J]. European Journal of Operational Research, 2006, 175(3): 1616-1627.
[14] BALL M, BODIN L, DIAL R. Matching based heuristic for scheduling mass transit crews and vehicles[J]. Transportation Science, 1983, 17(1): 4-31.
[15] YU H, XIAO R, DU Y, et al. A bus-arrival time prediction model based on historical traffic patterns[C]// The 2013 International Conference on Computer Sciences and Applications, December 14-15, 2013, Wuhan, China. Piscataway: IEEE Press, 2013.
[16] YU B, YANG Z Z, ZENG Q C. Bus arrival time prediction model based on support vector machine and Kalman filter[J]. China Journal of Highway and Transport, 2008, 21(2): 89-92.
[17] LIN Y, YANG X, ZOU N, et al. Real-time bus arrival time prediction: case study for Jinan, China[J]. Journal of Transportation Engineering, 2013, 139(11): 1133-1140.
[18] PADMANABAN R P S, DIVAKAR K, VANAJAKSHI L, et al. Development of a real-time bus arrival prediction system for Indian traffic conditions[J]. IET Intelligent Transport Systems, 2010, 4(3): 189-200.
[19] BERREBI S J, WATKINS K E, LAVAL J A. A real-time bus dispatching policy to minimize passenger wait on a high frequency route[J]. Transportation

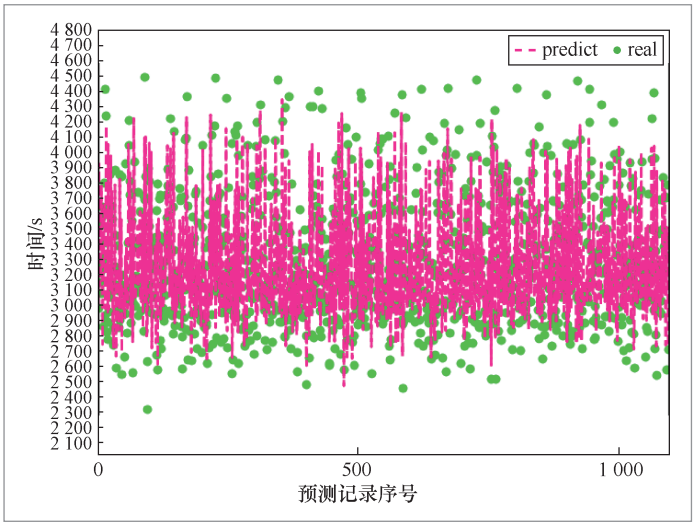


图 24 融合组件预测结果

表 11 各组件实验结果

组件名	准确率	误差率
停留时长组件	98.45%	7.98%
行驶时长组件	86.17%	6.69%
融合组件	81.07%	5.55%

表 12 对比实验结果

模型名	准确率	误差率
R-GBDT	81.16%	5.55%
AdaBoost	66.97%	7.32%
决策树	64.97%	8.49%
SVM	57.05%	9.69%
TTP	67.79%	8.26%
DS	79.07%	5.87%

Research Part B Methodological, 2015, 81(Part 2): 377-389.

[20] SHEU J B. A fuzzy clustering approach to real-time demand-responsive bus dispatching control[J]. Fuzzy Sets and Systems, 2005, 150(3): 437-455.

[21] 王麟珠, 苏庆列, 郑日博. 基于Elman动态神经网络的公交到站时间预测[J]. 机电技术, 2012, 35(1): 135-137.
WANG L Z, SU Q L, ZHENG R B. Bus arrival time prediction based on Elman dynamic neural network[J]. Mechanical and Electrical Technology, 2012, 35(1): 135-137.

[22] 张强, 张艳艳. 基于时间分段的公交车到站时间预测模型研究[J]. 数字技术与应用, 2014(11): 60, 62.
ZHANG Q, ZHANG Y Y. Study on prediction model of bus arrival time based on time segmentation[J]. Digital Technology and Application, 2014(11): 60, 62.

[23] 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.
JI Y J, LU J W, CHEN X S, et al. Prediction model of bus arrival time based on particle swarm optimization and wavelet neural network[J]. Journal of Transportation Systems Engineering and Information Technology, 2016, 16(3): 60-66.

[24] 罗频捷, 温荷, 万里. 基于遗传算法的模糊神经网络公交到站时间预测模型研究[J]. 计算机科学, 2016, 43(S1): 87-89, 108.
LUO P J, WEN H, WAN L. Research on bus arrival time prediction model based on fuzzy neural network with genetic algorithm[J]. Computer Science, 2016, 43(S1): 87-89, 108.

[25] 杨奕, 张雯蕊, 张灿. 基于遗传算法的BP神经网络在公交车到站时间预测中的应用[J]. 现代商业, 2017(16): 38-40.
YANG Y, ZHANG W R, ZHANG C. Application of BP neural network based on genetic algorithms in bus arrival time prediction[J]. Modern Business, 2017(16): 38-40.

[26] 张昕, 姜佳佳, 刘进. 基于SVM+GA的客运车辆到站时间预测[J]. 计算机与数字工程, 2017, 45(6): 1062-1066, 1085.
ZHANG X, JIANG J J, LIU J. Coach bus arrival time prediction based on SVM and GA[J]. Computer and Digital Engineering, 2017, 45(6): 1062-1066, 1085.

[27] 谢芳, 顾军华, 张素琪, 等. 基于MapReduce聚类和神经网络的公交车到站时间预测模型[J]. 计算机应用, 2017, 37(S1): 118-122.
XIE F, GU J H, ZHANG S Q, et al. Predicting model of bus arrival time based on MapReduce clustering and neural network[J]. Journal of Computer Applications, 2017, 37(S1): 118-122.

[28] O'SULLIVAN A, PEREIRA F C, ZHAO J, et al. Uncertainty in bus arrival time predictions: treating heteroscedasticity with a metamodel approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(11): 1-11.

[29] SINN M, YOON W J, CALABRESE F, et al. Predicting arrival times of buses using real-time GPS measurements[C]// International IEEE Conference on Intelligent Transportation Systems, September 16-19, 2012, Anchorage, USA. Piscataway: IEEE Press, 2012.

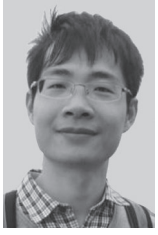
[30] ABIDIN A F, KOLBERG M. Towards improved vehicle arrival time prediction in public transportation: integrating SUMO and Kalman filter models[C]//The 17th Uksim-amss International Conference on Modelling & Simulation, March 25-27, 2015, Cambridge, UK. Piscataway: IEEE Press, 2016.

[31] JEONG R, RILETT L R. Bus arrival time prediction using artificial neural network model[C]//The 7th International IEEE Conference on Intelligent Transportation Systems, October 3-6, 2004, Washington, USA. Piscataway: IEEE Press, 2004.

[32] MAITI S, PAL A, PAL A, et al. Historical data based real time prediction of vehicle arrival time[C]//The 17th IEEE International Conference on Intelligent Transportation Systems, October 8-11,

2014, Qingdao, China. Piscataway: IEEE Press, 2014.
[33] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions[C]//The 11th Annual Conference on Computational Learning Theory, July 24-26, 1998, Madison, USA. New York: ACM Press, 1998: 80-91.
[34] QUINLAN J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1): 81-106.
[35] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features[C]//The 10th European Conference on Machine Learning, April 21-23, 1998, Chemnitz, Germany. Heidelberg: Springer, 1998: 137-142.

作者简介



赖永炫(1981-),男,博士,厦门大学副教授,主要研究方向为交通大数据分析、车载网络。

杨旭（1975- ），男，长春公交（集团）有限责任公司助理工程师，主要研究方向为公交排班管理和公交系统信息化。

曹琦（1982- ），男，龙岩烟草工业有限责任公司工程师，主要从事生产制造信息化与自动化及质量控制领域方面的研究与应用工作。

曹辉彬（1997- ），男，厦门大学信息学院硕士生，主要研究方向为交通大数据分析、车载网络。

崔磊（1982- ），男，博士，华侨大学计算机科学与技术学院教授，主要研究方向为智能感知网络、交通大数据分析、车载网络。

杨帆（1982- ），男，博士，厦门大学航空与航天学院副教授，主要研究方向为数据挖掘、聚类研究。

收稿日期： 2019-07-22

基金项目： 国家自然科学基金资助项目（No.61672441，No.61872154）；深圳市基础研究计划基金资助项目（No. JCYJ20170818141325209）；福建省自然科学基金资助项目（No.2018J01097）

Foundation Items: The National Natural Science Foundation of China(No.61672441, No.61872154),The Shenzhen Basic Research Proyram(No.JCYJ20170818141325209), The Natural Science Foundation of Fujian(No.2018J01097)