计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(12):177−183 [doi: 10.15888/j.cnki.csa.007147]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于机器学习的公交站间运行时间幂律分布分析①

徐文进[1], 寻晴晴[1], 周　笛[2]

[1](青岛科技大学 信息科学技术学院, 青岛 266061)
[2](解放军北部战区 91049 部队, 青岛 266000)
通讯作者: 寻晴晴, E-mail: 1255820862@qq.com

摘　要: 为了解决城市交通拥挤问题, 国家提倡乘坐公共交通出行, 使用公交智能卡的出行变的普遍了. 目前, 对于城际公交智能卡出行产生的数据, 很少有研究公交站间运行的时间. 因此, 提出了基于机器学习技术公交站间运行时间幂律分布的分析; 运用分站算法对城市公交进行分站, 获得公交车在相邻两站的运行时间; 并且对时间间隔数据进行了线性拟合. 运用南方某城市和北方某城市的两个数据集, 结果表明公交车运行时间间隔符合幂指数分布; 公交车运行的时间间隔符合人类行为动力学.

关键词: 机器学习; 公交智能卡; 分站算法; 人类行为动力学; 幂指数分布

引用格式: 徐文进,寻晴晴,周笛.基于机器学习的公交站间运行时间幂律分布分析.计算机系统应用,2019,28(12):177–183. http://www.c-s-a.org.cn/1003-3254/7147.html

Power Law Distribution Analysis of Running Time Between Bus Stations Based on Machine Learning

XU Wen-Jin[1], XUN Qing-Qing[1], ZHOU Di[2]

[1](School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)
[2](Unit 91049, PLA Northern Theater Command, Qingdao 266000, China)

Abstract: In order to solve the problem of urban traffic congestion, the state advocates travelling by public transportation, and the use of bus smart cards has become more common. At present, for the data generated by intercity bus smart card travel, there is very little research on the time between bus stations. Therefore, the power-law distribution analysis of running time between bus stations based on machine learning technology is proposed. The station algorithm is used to divide the bus stations in the city, and the running time of the bus in the two adjacent stations is obtained. The time interval data were fitted linearly. Using two data sets from a city in South China and a city in North China, the results show that the bus running time interval is in accordance with the power exponential distribution; the time interval of bus operation is in line with human behavior dynamics.

Key words: machine learning; bus smart cards; substation algorithm; human behavior dynamics; power exponential distribution

城际公交是指在城市群或大城市与周边中小城镇间设立的长途公交系统, 它作为行人出行的交通工具, 即方便又价廉; 城际公交具有自身的特点, 与一般的长途和城际快客有所不同, 是在通勤距离在十几到几十公里的公交体系, 在重要的城区之间开行的多班次, 具有固定公交站点和较明确营运计划表的通勤系统, 是

① 基金项目: 2018 年度山东省重点研发计划 (2018GGX105005)
Foundation item: Year 2018, Key Research and Development Program of Shandong Province (2018GGX105005)
收稿时间: 2019-03-28; 修改时间: 2019-04-18, 2019-05-06; 采用时间: 2019-05-16; csa 在线出版时间: 2019-12-10

趋向公交化的短途旅客营运车辆, 是市内公交车的延伸和拓展, 对乘客来讲更加便捷, 票价相比城际长途客运和轨道交通也更低廉, 在我国有较大的发展空间.

但是国内对城际公交的研究不多, 已有的研究主要集中在城际公交建立的政策和规划领域, 如文献[1–6], 较少涉及到城域公交的数据挖掘. 近年来蓬勃发展的人工智能机器学习技术是利用机器学习[7,8]算法对已知数据进行分析、计算得到合乎实际的规律、模型, 并进行合理的预测. 实践证明, 利用人工智能机器学习算法进行数据挖掘是分析城际公交运行规律的有效方法.

本文基于南方某城市和北方某城市公交线路的公交卡的交易记录, 通过对数据的合理分析, 结合机器学习分类算法, 合理设定分站标准, 找出公交站点[9]; 使用 Python 编程, 通过智能公交卡记录, 对乘车的人群行为进行挖掘和预测, 分析得出乘坐公交车的习惯和偏好; 同时在研究中证明城域公交车运行时站间时间间隔遵循人类行为动力学理论[10–13], 满足幂律分布.

1 公交智能卡数据集和数据预处理

1.1 公交智能卡数据集

城际公交作为城市群间重要的交通联系方式, 蕴含着大量的数据. 通过研究城际公交数据可提供大量、丰富的数据, 可以挖掘出助力城际公交发展的信息. 本文主要通过提取公交 IC 卡的数据来进行研究的; 公交 IC 卡现在一般采用射频识别技术进行数据通信, 使用集成电路芯片存储用户数据, 数据的可靠性和真实性有保障. 因此围绕着北方某城市和南方某城市城际公交的数据, 对公交 IC 卡的数据进行分析和应用, 发现公交客流量出行的一个规律. 如表 1 和表 2 描述了原始数据的类型.

表 1 南方某城市数据类型

字段名称	字段类型	样例	描述
地点	字符型	南方某城市市	使用地点
公交线路	字符型	某路城际公交	公交车路线
智能卡类别	字符型	普通卡	卡种类
打卡时间	字符型	20150104083045	打卡的时间
智能卡号	字符型	9999993103329408	公交卡唯一识别号
POS 机编号	字符型	00000055	为刷卡机唯一识别号

1.2 智能数据卡的预处理

智能卡打卡数据首先要进行数据的预处理, 如对打卡数据的时间戳进行转换、合并重复数据, 数据的预处理由以下步骤组成:

(1) 不完整数据清洗; 在智能卡打卡数据中, 一组完整的数据至少应该包括卡类型、卡号、智能终端号和打卡时间; 缺失一个, 就会对后面的算法实验造成影响, 因此对打卡数据中缺失数据项的进行清洗.

(2) 剔除不合理的数据, 在数据处理时发现有的打卡数据是在凌晨, 此时公交车并不运行, 所以应该剔除这些不合理的数据.

(3) 重复数据合并, 在数据预处理中发现, 在短时间内 (1 分钟) 有的卡打卡次数超过 2 次, 这是由于不小心重复打卡或是在数据传输中的错误造成的, 重复的打卡数据其实对应的同一次乘车行为, 所以这些重复的打卡数据要合并成为一项.

表 2 北方某城市数据类型

字段名称	字段类型	数据样例	描述
使用地	字符型	北方某城市公交	公交卡使用地
路线	字符型	21 路	公交车行驶路线
卡类型	字符型	普通卡	卡类型说明
交易日期时间	字符型	2017-03-14 13:27:59	刷卡时间
卡号	字符型	2660002700002350	公交卡的 ID, 具有唯一性
POS 机编号	字符型	370020030321	为公交车刷卡机的 ID, 具有唯一性

1.3 智能数据卡的特征

1.3.1 卡类型分布特征

数据预处理后, 对打卡数据的分布进行分析. 首先根据原始数据刷卡的情况, 卡类型进行分类, 得到的残疾卡, 老人卡, 普通卡, 学生卡, 员工卡, 治安监督卡所占的频数和比例, 如表 3 和表 4 所示.

表 3 南方某城市公交卡刷卡比例

卡类型	次数	百分比	有效的百分比	累计百分比
残疾卡	515	0.5	0.5	0.5
卡类型	1	0.0	0.0	0.5
老人卡	16 134	16.1	16.1	16.7
普通卡	75 966	76.0	76.0	92.6
学生卡	6821	6.8	6.8	99.4
员工卡	547	0.5	0.5	100.0
治安监督卡	16	0.0	0.0	100.0
总计	100 000	100.0	100.0	

通过表 3 和表 4 发现老年卡、学生卡、普通卡的出行所占比例挺高的, 但残疾卡的比例相对较少, 这可能与城际公交车设置无障碍设备有关.

1.3.2 客流量分布特征

把城际公交线路全日客流情况的刷卡的数据累加

起来, 从中可得到该全日累计客流的时间分布的数据. 获取的公交车的数据通过刷卡时间反映出刷卡乘客在全日内的高峰期, 如图 1 选取了某一天的客流情况.

表 4 北方某市公交车刷卡比例

卡类型	次数	百分比	有效的百分比	累计百分比
残疾卡	2664	4.1	4.1	4.1
公交员工卡	791	1.2	1.2	5.3
纪念卡	588	0.9	0.9	6.2
老年卡	18 194	27.9	27.8	34.0
普通卡	33 503	51.2	51.2	85.2
学生卡	3246	5.0	5.1	90.3
异形卡	6187	9.4	9.4	99.7
银行联名卡	133	0.3	0.3	100
总计	65 535	100.0	100.0	

图 1(a) 是南方某城市数据的客流量折线图, 图 1(b) 图是北方某城市数据的客流量折线图, 横轴表示的是刷卡的时间, 从 0 点到 24 点, 纵轴表示的刷卡的次数; 可以从图 1 中得出以下结论:

(1) 由图 1 得出早晨 5 点后至上午 8 点是一个出行高峰, 而且乘车人数几乎是随着时间成正比增加, 符合上班、上学的出行时间规律, 上午 8 点后乘坐公交车出行人数明显下降, 说明公共交通乘坐的主要人流还是上班族和上学族.

(2) 由图 1(a) 显示中午 12 点, 图 1(b) 显示 13 点乘车有一个小高峰, 这是因为中午是午休时段, 而且有的小学生下午不上课, 这部分乘客利用中午时段乘坐公交车出行.

(3) 由图 1 得出下午 16 点开始至傍晚 18 点乘车人数较多, 刷卡数据一直在增长. 18 点左右达到高峰, 是学生流和下班乘车叠加所致.

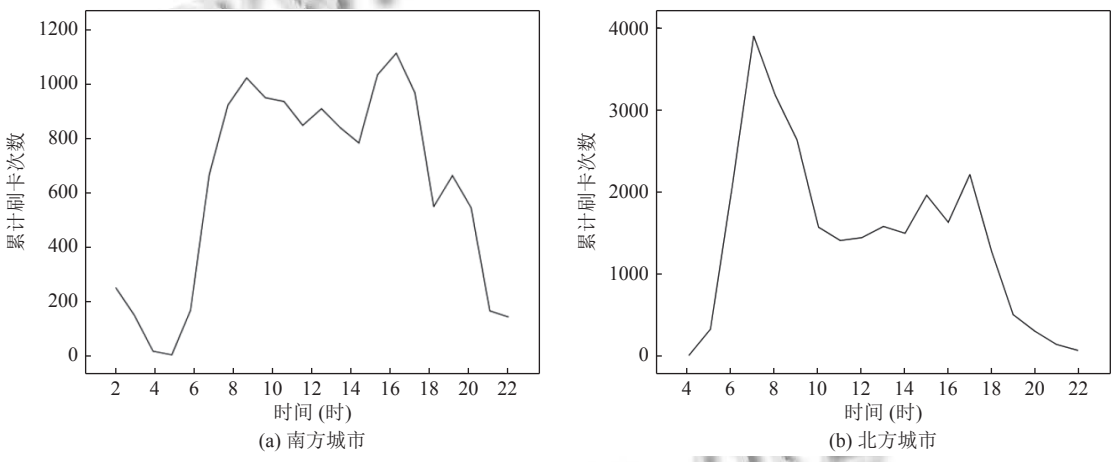


图 1 客流量分布图

2 公交车站间运行规律

在经过简单公交车智能卡数据分析后, 然后从时域角度分析公交车站间运行时间.

2.1 公交车分站算法

本论文选取南方某城市和北方某城市数据的 4 条线路, 对智能卡实际交易记录进行分站算法的研究; 合理的设计的分站算法, 进而找出公交站间的运行时间, 算法 1 是本文设计的公共交通分站算法.

算法 1 主要的思想是确定站间运行时间的阈值, 看查找的阈值是否符合实际的设站数量, 若符合, 算法结束; 否则, 继续查找. 由算法获得分站数据, 可以进一步进行数据拟合, 分析公交车站间运行的时间间隔规律.

算法 1. 公共交通分站算法

1. 输入: *data*,*Time*//*data* 输入智能卡刷卡数据, 按照线路和 POS 终端号进行排列, *Time* 为公交智能卡打卡时间信息;
2. *Stime*=*House*(*Time*)×3600+*Minute*(*Time*)×60+*Second* //将打卡的自然时间转换为可以用于算法的以秒 (s) 为单位的时间;
3. 对打卡时间进行排序;
4. 寻找排序后的打卡数据中时间相差最小阈值的两个数据, 作为站间分割;
5. 找出所有的站间后, 与实际设站数量进行比较, 如果大于时间设站数量, 回到步骤 4 重新设置站间运行时间阈值 (一般是增大阈值, 减少分站数量), 否则说明分站成功;
6. 得到站内打卡数据集并保存;
7. end// 算法结束.

2.2 线性拟合

本文通过分站算法, 得到公交车相邻两站的运行

时间, 对相邻两站的运行时间和相邻两站的运行时间的累计频数做双对数运算, 再通过机器学习线性回归[14–16]算法对其进行拟合.

2.2.1 最佳拟合函数

本文用到线性模型是一元一次的方程, 所以预测本文预测线性函数如式 (1):

$$h_\theta(x) = \theta_0 + \theta_1 x \tag{1}$$

式 (1) 得到的模型为数据预测的线性回归模型, x表示的是站与站时间间隔进行对数运算的结果; θ_0和θ_1表示的是两个变量建立联系的相关系数. $h_\theta(x)$用参数θ和x预测出来的y值.

对于式 (1), 为了找到最佳的θ_0和θ_1, 使拟合线的预测值更加接近y, 使用式 (2):

$$J(\theta_0,\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h(x^i)-y^i)^2 \tag{2}$$

在式 (2) 中, m表示训练样本的个数; y表示原训练样本中的y值, 也就是标准答案.

式 (2) 函数被称为平方误差函数, 这个平方误差函数结果越小, 即各个数据点更加接近拟合的函数, 这时找到θ_0和θ_1也是最佳的值. 得到的线性回归方程也是最佳的拟合函数.

2.2.2 梯度下降法

对于损失函数, 想要得到最优解, 找出最佳的θ_0和θ_1, 这里用到了梯度下降法. 梯度下降法, 就是沿着梯度下降最快的地方求偏导, 得到损失函数最小值时的θ_0和θ_1. 所以计算$J(\theta)$关于θ^{T}的偏导数, 也就得到了向量中每一个θ的梯度. 即式 (3):

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=0}^{m}(y^i - h_\theta(x^i))x_j^i \tag{3}$$

再沿着梯度的反方向更新参数θ的值, 即式 (4):

$$\theta_j = \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)x_j^i \tag{4}$$

一直迭代下去, 直到收敛某一个值, 就是最终要找的θ值.

3 实验结果和模型评价

3.1 分站算法仿真结果

对于本文提出的分站算法, 用了 4 条线路上的数据进行仿真实验. 仿真结果如图 2 所示.

图 2 中, 横坐标表示刷卡的时间间隔, 纵坐标表示分站的个数; 图 2(a) 为南方某城市地区的某两条公交路线的仿真结果, 两条路线的实际站数为 32 和 33 站, 得到的实验结果与实际结果一致, 对应的站与站的时间间隔最小值分别为 60 秒和 88 秒; 图 2(b) 为北方某城市地区的某两条公交路线的仿真结果; 两条路线的实际站数为 52 和 31 站, 得到的实验结果与实际结果也一致, 对应的站与站的时间间隔最小值分别为 56 秒和 81 秒. 由图可知随着刷卡时间间隔的增大, 分站的个数也相应的减少; 所以对于用分站算法找出相应的分站个数, 是可行的. 如图 2 中的特殊标记就是得出的分站个数和最小的站与站的时间间隔.

3.2 拟合结果

本文对南方某城市北方某城市的城际公交刷卡的时间做了计算, 所有的刷卡时间都是以秒为单位计算的, 得出来刷卡的时间间隔, 并对时间间隔和时间间隔累计的频数做双对数运算, 得出了数据的散点图, 如图 3 所示, 该散点图是符合人类行为动力学.

图 3 横坐标表示每站与每站的时间间隔的对数运算结果, 纵坐标表示是公交车每站与每站时间间隔的累计频数的对数运算结果. 由图 3 中可以看出数据是出现重尾特征, 是符合人类行为动力学的, 即大部分公交车站间的时间间隔较短, 而有少数部分公交车站间的时间间隔较长. 这种分布可能与出行的高峰期有关; 在出行高峰期, 由于乘车人数多、交通拥挤等状况, 可能出现公交车的站间时间间隔比较长; 在其他时间, 公交车站间的时间间隔相对较短.

3.3 拟合线路

本文通过线性回归方程对公交车的时间间隔的幂指数运算与累计的频数的散点图进行拟合. 如图 4 所示.

图 4 是绘制出使用经过函数优化得出的最优参数θ值所做预测的图形. 得到的曲线拟合方程是一元一次方程, 即式 (5) 和式 (6):

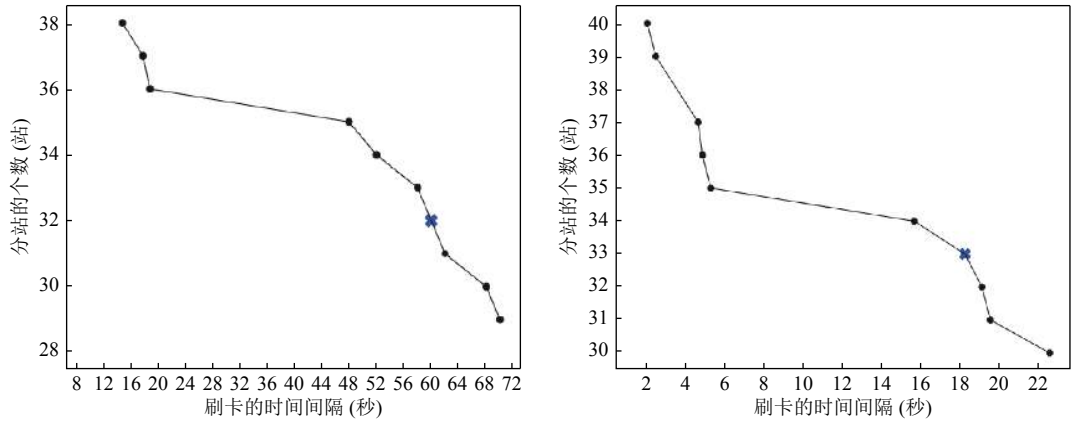
$$y = 7.868 - 1.018x \tag{5}$$

$$y = 12.251 - 1.793x \tag{6}$$

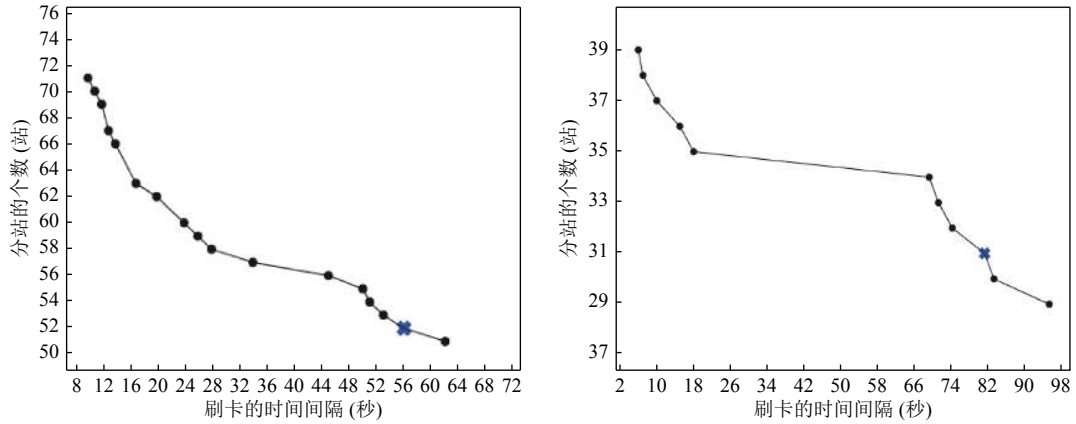
3.4 模型评价

对于本文得到的模型式 (5) 和式 (6) 进行检验, 判定模型是否符合, 得到结果如表 5.

如表 5 所示, R 方为判定系数, 一般认为需要大于 60%, 用于判定线性方程拟合优度的重要指标, 体现了回归模型解释因变量变异的能力, 越接近 1 越好. 模型得到的 R 方分别为 0.74 和 0.83, 判断模型拟合效果良好. 显著性为方差分析的显著性, 值都为 0.000<0.01<0.05, 表明由自变量时间间隔和因变量频数建立的线性关系回归模型具有极显著的统计学意义, 所建立的模型符合预期的规定.



(a) 南方城市



(b) 北方城市

图 2 分站算法仿真结果

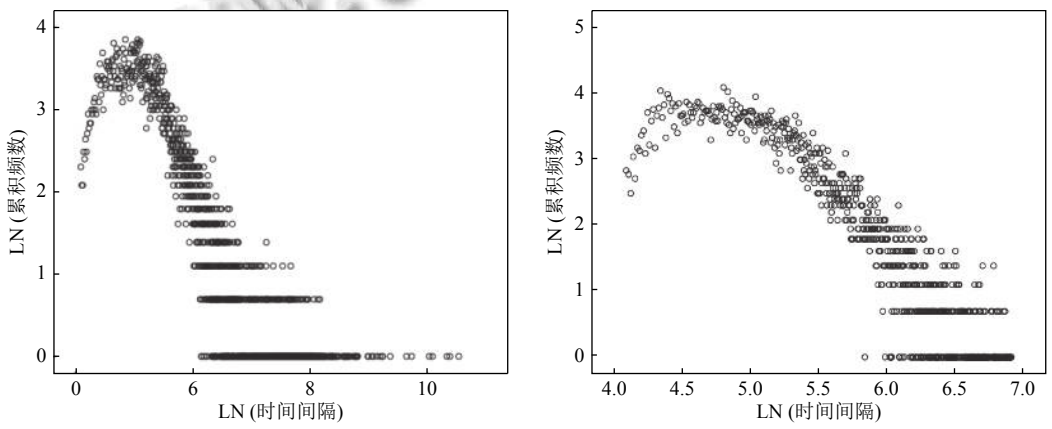


图 3 散点图

$$Ln(N) = 7.868 - 1.018Ln(T) \quad (7)$$

$$Ln(N) = 12.251 - 1.793Ln(T) \quad (8)$$

对于式 (7) 和式 (8) 为原始数据得到的公式, N 表示的是公交站间运行时间出现的次数, T 表示公交站间的运行时间. 由于验证了模型式 (5) 和式 (6) 的模型拟合效果良好, 式 (7) 和式 (8) 的模型拟合效果也良好. 所以式 (7) 和式 (8) 满足了幂指函数的判定准则, 由模型可以得出城市公交车站与站之间的运行时间符合幂律分布.

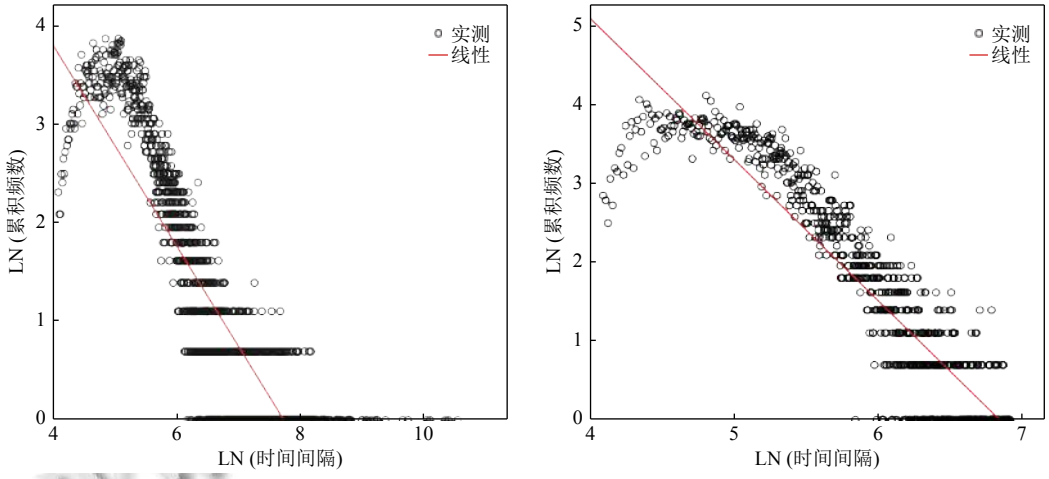


图 4 拟合图

表 5 模型结果验证

验证数据	R 方	显著性
南方某城市数据	0.74	0.000
北方某城市数据	0.83	0.000

4 结论

对于本文设计公交车分站算法, 通过刷卡的时间间隔去找分站的个数, 这使得分站的结果可靠准确; 并且运用 2 个数据集进行实验, 都能够准确的找到分站的个数和相应的最小的站与站的时间间隔, 该算法具有可行性.

对于用分站算法得到的站与站的时间间隔, 实验结果证明了城际公交车站与站之间的运行时间符合幂律分布; 站与站的时间间隔符合人类行为动力学的. 可以得出, 公交车站与站时间间隔序列具有强记忆和较弱的突发性. 这意味着大多数站与站的时间间隔相对均匀, 这个时间序列具有一定的记忆性, 并且长时间间隔可能遵循较长的时间间隔, 短时间间隔可能遵循较短的时间间隔, 有比较弱的突发性. 这与日常时间表一致, 在高峰时段, 有许多乘客在等车, 刷卡的人较多, 站与站的时间间隔相对长点. 但是在大多数人呆在家里的夜晚, 乘客很少, 刷卡的人相对较少, 站与站的时间间隔相对较短.

参考文献

1 李耀鼎, 朱洪, 程杰. 国内城际公交发展案例分析. 交通与运输, 2012, (2): 26–28.
2 曹佳, 齐岩. 城际公交一体化发展模式研究. 综合运输, 2013, (10): 56–59.
3 曲思源, 徐行方, 洪玲. 城际公交车高峰时段发车间隔时间优化. 武汉理工大学学报 (交通科学与工程版), 2012, 36(1): 95–97, 102. [doi: 10.3963/j.issn.1006-2823.2012.01.022]
4 廖勇. 公交化城际列车开行间隔优化. 铁道学报, 2010, 32(1): 8–12. [doi: 10.3969/j.issn.1001-8360.2010.01.002]
5 曹兴举, 杨意品. 城际公交的发展现状及实施效果评估. 公路交通科技: 应用技术版, 2016, 12(4): 285–288.
6 Lin YF, Wang Y. Public traffic planning based on green traffic concept. Proceedings of 2011 Fourth International Conference on Intelligent Computation Technology and Automation. Shenzhen, China. 2011. 1146–1148.
7 王泓正. 机器学习在数据挖掘中的应用. 中国新技术新产品, 2018, (22): 98–99. [doi: 10.3969/j.issn.1673-9957.2018.22.055]
8 王旸. 大数据背景下机器学习在数据挖掘中的应用. 信息与电脑, 2018, (21): 138–139, 142.
9 Liu LQ, Zhang Y. Research of urban bus stop planning based

on optimization theory. Proceedings of 2009 International Conference on Measuring Technology and Mechatronics Automation. Zhangjiajie, China. 2009. 551–554.

10 樊超, 郭进利, 韩筱璞, 等. 人类行为动力学研究综述. 复杂系统与复杂性科学, 2011, 8(2): 1–17. [doi: 10.3969/j.issn.1672-3813.2011.02.001]

11 韩筱璞, 汪秉宏, 周涛. 人类行为动力学研究. 复杂系统与复杂性科学, 2010, 7(2–3): 132–144.

12 方小妹, 葛璞, 谢超, 等. 校园网论坛评论的人类动力学分析研究. 智能计算机与应用, 2017, 7(2): 90–93, 98. [doi: 10.3969/j.issn.2095-2163.2017.02.025]

13 Han ZF. Research on dynamics modeling based on human-structure interaction. Proceedings of 2018 2nd International Conference on Computer Science and Intelligent Communication. Hohhot, China. 2018. 278–283.

14 李苹, 刘昆, 徐坚, 等. 一元线性回归在成绩预测中的应用. 电脑知识与技术, 2016, 12(24): 125–126.

15 李玉毛, 何涛, 刘冬. 一元线性回归方法的理论及其应用. 赤峰学院学报 (自然科学版), 2017, 33(15): 1–2. [doi: 10.3969/j.issn.1673-260X.2017.15.001]

16 Yang Q, Yuan PL, Zhang Q, *et al*. Mobile phone user behavior prediction base on multivariable linear regression model. DEStech Transactions on Computer Science and Engineering, 2019.