基于大规模数据的公交到站时间预测方法比较

庞俊彪[1],胡安静[1],黄 晶[1],杜 勇[2],于海涛[2]
(1. 北京工业大学 信息学部 北京多媒体与智能软件技术重点实验室,北京 100124;
2. 北京市交通信息中心,北京 100161)

摘 要:公交到站时间预测作为提高公共交通运输服务水平的重要措施,能够鼓励用户使用公共交通出行,方便调度部门进行合理调度。通过研究现有文献,发现虽然已经提出了很多不同原理的公交到站时间预测方法,但由于各个文献中使用的数据集合、测试规模不同,所以在现有方法之间无法进行有效的比较,从而无法发现公交到站时间预测的基本问题。为了提供可靠准确的数据基础,实现在统一的数据集上公平地比较目前现有的方法,建立了北京市公交到站数据集。该公交到站数据集是目前为止最大的公交运营数据集,其中包含了各种复杂的路况和可能的情况。在北京市公交到站数据集上,通过选择典型到站预测方法,进行实验比较和结果分析,定位出公交到站时间预测的本质问题。

关键词:公交到站时间预测;性能比较;算法评估;GPS 数据
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2019)04-0024-05
doi:10.3969/j.issn.1673-629X.2019.04.005

Empirical Comparison among Classic Bus Arriving Time Prediction Methods Based on a Large-scale Dataset

PANG Jun-biao[1], HU An-jing[1], HUANG Jing[1], DU Yong[2], YU Hai-tao[2]
(1. Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Department of Information, Beijing University of Technology, Beijing 100124, China;
2. Beijing Transportation Information Center, Beijing 100161, China)

Abstract: As an important measure to improve the level of public transport service, the prediction of bus arriving time can encourage users to choose public transport and facilitate the scheduling department to make reasonable scheduling. Via the investigation of various existing methods for bus arriving time prediction, however resulting from using diverse datasets with different scale, the comparisons among them are not effective, so the intrinsic of bus arriving time prediction model cannot be found. Due to above shortcomings, the most comprehensive Beijing bus transportation dataset which includes sundry road conditions and providing reliable accurate platform to compare the algorithm is built. Based on the Beijing bus transportation dataset, through the analysis and comparison of existing prediction algorithms, as a result, the essential problem of bus arriving time prediction model is found.

Key words: bus arrival time prediction; performance comparison; algorithm evaluation; GPS data

0 引 言

交通在现代社会中扮演重要的角色,基于联合国人口署公布的报告:城市人口的增长率每年以大约 1.5% 的比例增长,预计在 2050 年,有大约 64 亿城市居住人口[1]。城市人口的增长将引发总出行需求的增长,就有限的交通资源来说,高效的公交系统既有利于降低城市交通拥堵,提高城市交通运转效率,同时也能减少汽车尾气排放,提升城市空气质量。在我国人口数量多、密度高,城市土地资源紧张,在城市化进程加快的过程中,优先发展城市公共交通是一项利于城市可持续发展的重要任务[2]。

考虑到发展公共交通带来的益处,城市公共交通管理部门希望引进相关的技术向乘客提供一些公交信息,从而提高交通服务水平[3]。在美国,有研究人员针对乘客所关心的公交信息种类进行问卷调查[4]:在众多的公交信息中,最让人关心的信息之一即为公交车

收稿日期:2018-04-16 **修回日期**:2018-08-17 **网络出版时间**:2018-12-20
基金项目:国家自然科学基金(61672069)
作者简介:庞俊彪(1980-),男,副教授,CCF 会员(25985M),研究方向为跨媒体建模与分析;胡安静(1994-),女,硕士,研究方向为跨媒体数据挖掘与建模分析。

到站时间。所以,准确预测公交车到站时间并及时将信息传递给用户,是提高服务水平的关键性一步[5]。准确的到站时间预测,可以有效减少用户的等待时间,提高服务质量[6],方便和鼓励用户使用公共交通出行。正因为公交到站时间预测的重要性,使其成为一项重要的研究课题[7]。

在现实生活中,由于交通运输的随机性,如路口的延时、用户在时间和空间上需求的波动、天气情况等因素,公交到站时间很难进行预测。早年间,大多数公交车到站时间预测的研究,主要是通过大量调查,着重于预测模型的建立[8-9]。直到 1999 年,基于自动车辆定位技术(automatic vehicle location,AVL)的全球定位系统(global position system,GPS)的出现,使研究发生了质的飞跃[10-11]。随后,简单的线性预测模型被非线性预测模型取代,例如核函数机[12]、神经网络[13]等。然而正如 O'Sullivan 在文献[14]中关于公交到站时间预测不确定性的讨论所提到的:"公交到站时间预测问题的复杂性在于,公交运行时间是非线性和复杂的多种因素交互作用的结果。"

1　现有方法

回顾现有的大多数预测算法,公交到站时间预测算法根据原理的不同,大体可以分为三大类:基于回归算法、基于滤波算法、基于搜索算法。

1.1　基于回归算法

在基于回归算法中,将公交到站时间作为受其他时空因素控制函数的响应量。因此这类算法之间最大的差异在于,如何定义到站时间和其他时空因素之间的非线性关系。该类算法包括以下四种:

(1) SPB[8](SVM on probe buses)算法中首先将路径划分成路段,随后从运行在下游路段的前一辆公交作为探针,提出了四种特征作为模型的输入。支持向量机和支持向量回归的非线性,从数学本质上来说都来源于核函数。但确定核函数及其相应参数较为困难,计算复杂度大,不适用于解决大规模计算量问题[15]。实验中通过公开的源代码实现 SVM 算法。

(2) LRT[16](linear regression on trajectories)算法假设公交行进时间服从多元高斯分布,故而预计到站时间为公交到达目标站点的后验期望。

(3) AMM[17](additive mixed model)算法使用半参数模型来建立多因素模型,使用周末与否、当前时间、行进距离等因素,用函数来描述这些因素与公交到站时间之间的关系。文中利用 R 语言的 MGCV 包实现了 AMM 算法。

(4) MLP[18](multilayer perceptron)算法中使用了三种特征:当前时间、当前位置、与目标站点之间的距离。多层感知机是一种特殊的前向传播神经网络算法[19],在本次实现中遵循原文的设计,选择和调整隐含层节点个数,报告节点数为 15 时得到的最好结果。文中使用 MATLAB 中的 newff 函数实现该算法。

1.2　基于滤波算法

遵循经典贝叶斯预测原理,基于滤波算法根据预先提供的路段信息,预测给定路段的行程时间,从而实现公交到站时间的预测。通过定义或者利用相应的模型学习公交车运行的动力系统,以递归的方式进行公交车到站时间预测。

卡尔曼滤波器是这类模型的典型代表。它是现代控制领域的一种重要方法,通过一系列的递归数学公式,即状态方程和观测方程组成的线性随机系统来描述,它是一种最优化自回归数据处理算法,采用线性无偏最小均方误差估计准则,对过程的状态进行预测。

KFP[20](Kalman filter with probe buses)算法首先将路径划分为路段,而后学习动态滤波器。实现过程中使用期望最大化实现线性动态滤波。

1.3　基于搜索算法

影响交通的因素众多,而这一类模型采取了"用替换代替预测"的策略来避免考虑这些因素。简单地假定在相似的交通环境下,公交车到达各个站点的时间是接近的。这种方法一般会以当前车辆的通行时间作为依据,去搜索一些相似的历史数据。

基于搜索算法在所预测公交运行的交通状况与历史数据相似的情况下,预测结果较好。为了提高准确率,将一天的时间按照定义的时间块进行分层[21]。但当处理大数量轨迹数据时,k-NN 的计算复杂度将会很大。

k-NN[20]算法和 KR[16](kernel regression)算法均是基于搜索的算法,通过权值化历史相似运营轨迹进行预测。如 k-NN 算法选择 k 条和当前轨迹最接近的历史轨迹,获得这些历史轨迹到达下游待预测站点的时间,并分别乘以权值(如 $1/K$)得到当前公交车到达下游站点的时间。KR 为了达到非线性[16],采用基于核函数(如高斯核,$\exp(-\|x-y\|^2/b)$)计算权值的方法,其中 x 和 y 代表当前的运营轨迹和历史运营轨迹,b 是核的带宽。实验中根据文中的最优参数设置,设 $b=1$。

1.4　讨　论

虽然现有很多算法,但因为数据获取较为困难,导致大多数算法并没有使用大规模的数据集对其进行验证,只使用极少的线路对算法进行性能评价,所以无法比较算法的准确性。并且在不同的数据集上进行的实验无法公平地比较各算法的性能,每种算法在不同的条件下会有不同的结果,比如某算法可能在短距离上

的预测比较差,但在长距离的预测上却表现较好,所以各种算法缺少在同一数据集上进行对比。

2 大规模公交运营数据集的构建

目前,针对公交车到站预测问题,尚没有公认的大规模数据集,故文中选择了北京 47 条公交线路进行数据采集和预处理,并建立了北京市公交运营数据集。这些运营线路在北京市的分布如图 1 所示,其中黑色线条表示公交车的路径轨迹,它们几乎覆盖了北京市城区的主要干道。

将该数据集和其他研究所用到的数据集在线路数量、线路长度、站点数量等方面进行了比较,如表 1 所示,说明了数据集的无偏性和全面性。

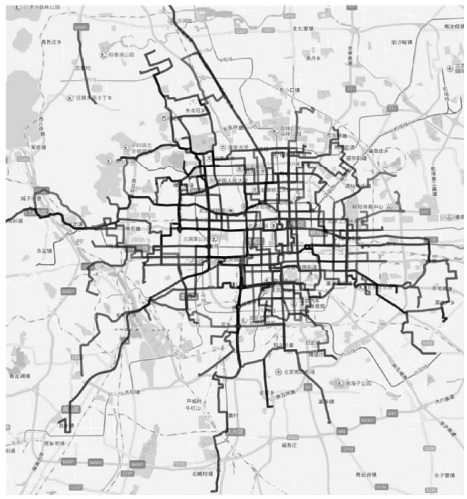

图 1 47 条线路在北京市的空间分布

表 1 数据集的比较

数据集	线路数量	线路长度/km		站点数量		车辆数量		运营次数	
		最小值/最大值	均值±标准差	最小值/最大值	均值±标准差	最小值/最大值	均值±标准差	最小值/最大值	均值±标准差
文献[14]	1	18.87	18.87	59	59	–	–	1 570	1 570
文献[15]	4	4/15	11±5.2	15/54	27.8±17.9	–	–	1 276/7 882	3 324.5±3 082.8
文献[8]	2	0.6/0.7*	0.7±0.1	3	3	–	–	224/237	229.7±6.7
文中	47	5.4/24.6	12.0±4.3	6/39	18.9±6.3	11/58	22.7±8.8	718/3 009	1 606.5±527.6

注:文献[8]只使用了两条线路的部分只含有 3 个站台的区间,作为个案研究,所以线路长度相对较短。

2.1 公交运营数据描述

数据集共两部分:公交的动态 GPS 数据;道路网的静态信息,如公交站台的位置,路口的位置,每个公交站台的相关统计量。

公交动态 GPS 数据从北京的 47 条线路,共 1 089 辆公交车中收集,数据时间跨度为 2015 年 2 月 1 号至 2015 年 2 月 28 号。每条 GPS 数据包括以下信息:公交车的经纬度坐标,时间戳,公交号,线路号。每 30 s 产生一条 GPS 数据,且最大位置误差为 50 m。但由于高层建筑或天桥造成的信号传输的阻塞,GPS 数据中存在误差数据。因此通过建立城市的边界,过滤这种错误的 GPS 数据,实验中设置经度范围为 110 ~ 200,纬度范围为 30 ~ 50,作为北京市的边界。

道路网的静态信息包括:公交站台的位置,线路号,行车方向,路口位置,公交站台的规模(每个站台有多少种线路的公交停靠)。其中路口预示有时间延后的可能性,公交站台的规模间接提供了公交在该站台可能停驻的时长。

2.2 大规模公交运营数据集构建

将 GPS 数据和静态信息数据协同映射为“到达时间-行驶距离数据对”。通过计算两个连续 GPS 数据点之间的距离,首先累积两点之间的距离,得到了行进距离。接着需要解决以下两个问题:将每个公交站台映射到行进距离中;获得每个站台的到达时间。

为了解决这两个问题,将每个站点的 GPS 坐标映射到行进距离中,通过插值得到相应的到达时间。

这一步骤中通过克里金插值法(Kriging interpolation)[21]的泛化算法,高斯过程(Gaussian process)[16],得到插值点。由此,GPS 数据和站点位置最终转化为一个时间空间联合的轨迹。

3 实 验

3.1 评估量设置

为了公平地评估每种算法的总体性能,使用平均绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)以及平均绝对百分比(mean absolute percentage error, MAPE)来评价预测精度,具体公式如下:

$$\mathrm{MAE}=\frac{2\sum_{i=1}^{N}\sum_{k=1}^{K-1}\sum_{\Delta=1}^{K-k}\left|t_{k+\Delta}^{i}-\hat{t}_{k+\Delta}^{i}\right|}{NK(K-1)} \tag{1}$$

$$\mathrm{RMSE}=\frac{\sum_{i=1}^{N}\sum_{k=1}^{K-1}\sqrt{\frac{1}{K-k}\sum_{\Delta=1}^{K-k}\left(t_{k+\Delta}^{i}-\hat{t}_{k+\Delta}^{i}\right)^{2}}}{N(K-1)} \tag{2}$$

$$\mathrm{MAPE}=\frac{2\sum_{i=1}^{N}\sum_{k=1}^{K-1}\sum_{\Delta=1}^{K-k}\left|\frac{t_{k+\Delta}^{i}-\hat{t}_{k+\Delta}^{i}}{t_{k+\Delta}^{i}}\right|}{NK(K-1)}\times 100\% \tag{3}$$

其中,i 表示该线路的第 i 次运营;N 表示运营的总量;K 表示该线路站点的数量。

3.2　结果与讨论

分别用上述七种算法在测试集上进行实验对比。将 47 条线路按运营距离分为三类,分别为短距离(0 ~ 10 km)、中距离(10 ~ 15 km)和长距离(>15 km),计算各线路的 MAE、RMSE、MAPE。表 2 记录了三种评价指标在各组的均值和标准差。

表 2　现有算法在本数据集上实现结果的比较

算法	<10 km			10 ~ 15 km			>15 km		
	MAE/min	RMSE/min	MAPE/%	MAE/min	RMSE/min	MAPE/%	MAE/min	RMSE/min	MAPE/%
k-NN[21]	1.31±0.22	1.52±0.27	18.16±1.98	1.49±0.24	1.74±0.28	17.26±1.73	1.43±0.11	1.66±0.14	16.00±1.35
KR[16]	1.23±0.20	1.42±0.25	17.29±1.82	1.41±0.22	1.64±0.26	16.59±1.75	1.38±0.15	1.61±0.19	15.71±1.36
KFP[20]	2.19±0.94	2.31±0.98	84.69±59.19	3.07±1.16	3.23±1.18	78.63±38.53	2.68±0.40	2.84±0.39	71.27±39.16
SPB[8]	1.50±1.51	1.80±0.64	48.10±41.58	2.51±0.78	3.06±0.93	44.79±15.21	2.78±0.84	3.51±1.05	60.20±49.84
LRT[16]	1.21±0.19	1.40±0.23	17.28±2.35	1.37±0.22	1.60±0.26	16.21±1.97	1.35±0.12	1.57±0.14	15.64±1.65
AMM[17]	1.29±0.94	1.47±0.34	31.17±15.81	1.55±0.34	1.78±0.38	26.40±21.65	1.44±0.25	1.68±0.29	36.31±55.43
MLP[18]	1.22±0.34	1.37±0.39	28.59±12.35	1.70±0.33	1.96±0.38	23.01±3.71	1.95±0.64	2.30±0.95	23.07±5.78

根据表 2 可知,KFP 算法的总体性能最低,主要原因是 KFP 作为一种基于滤波算法,在预测时间时,只依靠下一步准确的测量对预测模型进行“纠正”。如果从历史数据学习到的动力系统和现实的交通情况出入较大,结果会迅速变差,并且不断累积和传播。

k-NN 和 KR 算法在各种情况和各种指标下的表现很相近,正如所预期的那样,基于搜索的算法只有在当前的交通环境和搜索到的历史轨迹所对应的交通环境相似时才能预测较为准确的结果。

值得注意的回归类算法,如 LRT 和 AMM 都取得了较好的结果,因为这些方法避免了多步超前预测的难点,不需要迭代进行预测。其中 LRT 并不像 k-NN 和 KR 算法,只统计了历史相似轨迹,它捕获了公交不同路段之间运行状态的线性叠加关系,故而在大量数据的情况下,预测效果最好。

AMM 是半参数回归模型的一种。该算法考虑多种因素,通过函数来描述因素和到站时间之间的关系,是一个统计型原理模型建立框架,灵活地整合了多种因素,算法总体性能较好。

SPB 由于其预测结果与支持向量机中核函数及其参数的选择关系较大,所以参数较难控制,且 SPB 算法使用下游前一辆公交作为探针,所以对于前一辆公交的信息较为依赖,在某些情况下结果较差。

MLP 算法的核心是典型的多层感知机网络,这是一种具有三层(输入层、隐藏层和输出层)神经元的前向型神经网络,文中使用了三种特征,分别为当前时刻,当前位置和目标站点之间的距离,输入特征较少,没有考虑时序间的关联性。

此外,由于交通的复杂性,预测其在较远时空下的状态相对较难,还将考虑在不同预测距离的情况下,算法的预测能力是否会有较大的变化。将预测距离划分为 2 km 一段,即[0,2),[2,4),[4,6)……通过图 2 来描述绝对误差在不同预测距离区间内的分布。

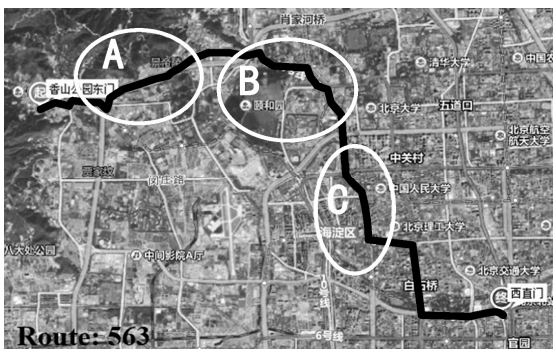

图 2　线路 563 的分布

为了清晰地比较几种算法的误差分布,利用最长的 563 路公交线路对上述 7 种算法进行测试,如图 2 所示。该线路具有 29 个站点,且线路穿过了旅游胜地(香山),著名的历史景点(颐和园),集中商业区(中关村),分别对应图 2 中 A 区域、B 区域、C 区域,通常拥有较大客流量,线路总体路况复杂,对于预测公交到站时间来说,具有挑战性,适合进行算法性能的比较。实验结果如图 3 所示,其中图中折线为第 95 个百分位的绝对误差。

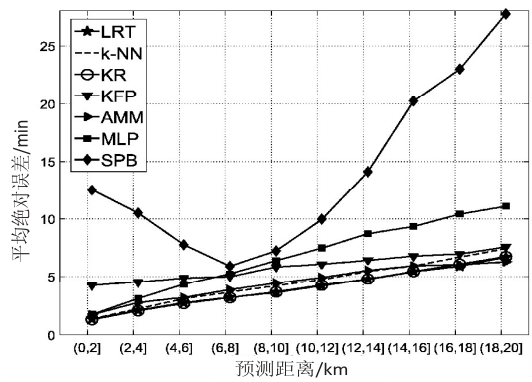

图 3　各算法在不同预测距离的平均绝对误差对比

从图 3 可以看出,各算法性能大致与表 2 中相同,但随着到达站台距离起点站台的距离逐渐增加,预测到站时间的绝对误差随之增加。值得注意的是,由于 SPB 算法是通过多条线路来预测公交到站时间,所以在只使用一条线路时,SPB 算法性能最差也是预料之中的。

4 结束语

文中提供了一个大规模公交运营数据集,包括不同运营距离的线路,几乎覆盖了北京市城区的主要干道。与其他研究所用到的数据集进行了比较,表明该数据集有无偏性、全面性以及大规模等特点,为公平地比较现有算法提供了可靠准确的数据基础。在将各算法按照原设计实现后,在该数据集上进行实验对比,发现基于回归算法的各方法具有更稳定的性能。且各方法在不同长度线路上的结果中 LRT 算法的结果最好且稳定。

随着线路的行进,公交到站时间预测的准确性越来越差,从绝对误差看出预测到站时间越来越不准确。所以从预测准确度的角度看,到站预测算法还有很大的研究空间。尤其是如何融合多种异构信息进行预测,目前只能关联时间、空间两种因素,而对于路况、道路约束、天气等因素无法进行建模和利用。这也将是到站时间预测未来努力的方向之一。

参考文献:

[1] 张 婷. 基于健康力分析框架的生态城市建设研究[J]. 中国工程咨询,2015(3):19-22.

[2] 王海洋,陈 昕,苗晓坤,等. 基于国情的我国城市交通问题分析与对策研究[J]. 辽宁工业学院学报:自然科学版,2006,26(2):118-119.

[3] 张 凡,上官晨博,邹吉聪,等. 公共汽车交通服务水平评价系统[J]. 山西大学学报:自然科学版,2010,33(2):198-206.

[4] 罗 虹. 基于 GPS 的公交车辆到达时间预测技术研究[D]. 重庆:重庆大学,2007.

[5] 周雪梅,彭昌溆,宋兴昊,等. 基于前车数据的动态公交车辆到站时间预测模型研究[J]. 交通与运输,2011(2):52-56.

[6] 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.

[7] 任 远,吕永波,马继辉,等. 基于粒子滤波的公交车辆到站时间预测研究[J]. 交通运输系统工程与信息,2016,16(6):142-146.

[8] YU Bin,LAM W H K,TAM M L. Bus arrival time prediction at bus stop with multiple routes[J]. Transportation Research Part C:Emerging Technologies,2011,19(6):1157-1170.

[9] ALTINKAYA M,ZONTUL M. Urban bus arrival time predication:a review of computational models[J]. International Journal of Recent Technology Engineering,2013,2(4):164-169.

[10] 马 飞. 基于 GPS/GIS 实时定位系统的设计与实现[D]. 成都:西南交通大学,2005.

[11] 刘 旭,杨东凯,张其善. 自动车辆定位系统的 GPRS 链路性能分析[J]. 交通运输系统工程与信息,2007,7(2):46-51.

[12] YU Bin,YANG Zhongzhen,YAO Baozhen. Bus arrival time prediction using support vector machines[J]. Journal of Intelligent Transportation Systems,2006,10(4):151-158.

[13] CHIEN I J,DING Y,WEI C. Dynamic bus arrival time prediction with artificial neural networks[J]. Journal of Transportation Engineering,2002,128(5):429-438.

[14] O' SULLIVAN A,PEREIRA F,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):3286-3296.

[15] CRISTIANINI N,SHAWE-TAYLOR J. An introduction to support vector machines[M]. New York,NY,USA:Cambridge University Press,2000.

[16] SINN M,YOON J W,CALABRESE F,et al. Predicting arrival times of buses using real-time GPS measurements[C]// International IEEE conference on intelligent transportation systems. Anchorage,AK,USA:IEEE,2012:1227-1232.

[17] KORMÁKSSON M,BARBOSA L,VIEIRA M R,et al. Bus travel time predictions using additive models[C]//IEEE international conference on data mining. Shenzhen,China:IEEE,2014:875-880.

[18] GURMU Z,WEI F. Artificial neural network travel time prediction model for buses using only gps data[J]. Journal of Public Transportation,2014,17(2):45-65.

[19] HAGAN M T,DEMUTH H B,BEALE M. Neural Network design[M]. Beijing:China Machine Press,2002.

[20] VANAJAKSHI L,SUBRAMANIAN S C,SIVANANDAN R. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses[J]. IET Intelligent Transport Systems,2009,3(1):1-9.

[21] COFFEY C,POZDNOUKHOV A,CALABRESE F. Time of arrival predictability horizons for public bus routes[C]// ACM SIGSPATIAL international workshop on computational transportation science. [s.l.]:ACM,2011:1-5.