软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(3):648−662 [doi: 10.13328/j.cnki.jos.005901]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于时空相关属性模型的公交到站时间预测算法*

赖永炫[1,2], 张 璐[1,2], 杨 帆[2,3], 卢 卫[4], 王 田[5]

[1](厦门大学 信息科学与技术学院 软件工程系,福建 厦门 361005)

[2](厦门大学 深圳研究院,广东 深圳 518057)

[3](厦门大学 航空航天学院 自动化系,福建 厦门 361005)

[4](中国人民大学 信息学院 计算机系,北京 100872)

[5](华侨大学 计算机科学与技术学院,福建 厦门 361021)

通讯作者: 赖永炫, E-mail: laiyx@xmu.edu.cn

摘 要: 公交车辆到站时间的预测是公交调度辅助决策系统的重要依据,可帮助调度员及时发现晚点车辆,并做出合理的调度决策.然而,公交到站时间受交通拥堵、天气、站点停留和站间行驶时长不固定等因素的影响,是一个时空依赖环境下的预测问题,颇具挑战性.提出一种基于深度神经网络的公交到站时间预测算法 STPM,算法采用时空组件、属性组件和融合组件预测公交车辆从起点站到终点站的总时长.其中,利用时空组件学习事物的时间依赖性与空间相关性.利用属性组件学习事物外部因素的影响.利用融合组件融合时空组件与属性组件的输出,预测最终结果.实验结果表明,STPM 能够很好地结合卷积神经网络与循环神经网络模型的优势,学习关键的时间特征与空间特征,在公交到站时间预测的误差百分比和准确率上的表现均优于已有的预测方法.

关键词: 到站预测;梯度提升树;卷积长短期记忆网络

中图法分类号: TP18

中文引用格式: 赖永炫,张璐,杨帆,卢卫,王田.基于时空相关属性模型的公交到站时间预测算法.软件学报,2020,31(3):648−662. http://www.jos.org.cn/1000-9825/5901.htm

英文引用格式: Lai YX, Zhang L, Yang F, Lu W, Wang T. Bus arrival time prediction algorithm based on spatio-temporal correlation attribute model. Ruan Jian Xue Bao/Journal of Software, 2020,31(3):648−662 (in Chinese). http://www.jos.org.cn/1000-9825/5901.htm

Bus Arrival Time Prediction Algorithm Based on Spatio-temporal Correlation Attribute Model

LAI Yong-Xuan[1,2], ZHANG Lu[1,2], YANG Fan[2,3], LU Wei[4], WANG Tian[5]

[1](Department of Software Engineering, School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

[2](Shenzhen Research Institute, Xiamen University, Shenzhen 518057, China)

[3](Department of Automation, College of Aerospace Engineering, Xiamen University, Xiamen 361005, China)

[4](Department of Computer Science, School of Information, Renmin University of China, Beijing 100872, China)

[5](School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

* 基金项目: 国家自然科学基金(61672441, 61872154); 深圳市基础研究计划(JCYJ20170818141325209); 福建省自然科学基金(2018J01097)

Foundation item: National Natural Science Foundation of China (61672441, 61872154); Basic Reasearch Plan of Shenzhen (JCYJ 20170818141325209); Natural Science Foundation of Fujian Province of China (2018J01097)

本文由人工智能赋能的数据管理、分析与系统专刊特约编辑李战怀教授、于戈教授和杨晓春教授推荐.

收稿时间: 2019-07-17; 修改时间: 2019-09-10; 采用时间: 2019-11-25; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-10 13:34:33, http://kns.cnki.net/kcms/detail/11.2560.TP.20200110.1333.004.html

Abstract: Bus arrival time prediction is an important basis for the decision-making assistant system of bus dispatching. It helps dispatchers to find late vehicles in time and make reasonable dispatching decisions. However, bus arrival time is influenced by traffic congestion, weather, and variable time when stopping at stations or travelling duration between stations. It is a spatio-temporal dependence problem, which is quite challenging. This study proposes a new algorithm called STPM for bus arrival time prediction based on deep neural network. The algorithm uses space-time components, attribute components and fusion components to predict the total bus arrival time from the starting point to the terminal. In this algorithm, time-dependence and space-time components are used to learn the internal spatio-temporal dependence. It uses attribute components to learn the influence of external factors, uses fusion components to fuse the output of temporal and spatial components, as well as attribute components, to predict the final results. Experimental results show that STPM can combine the advantages of convolutional neural network and recurrent neural network model to learn the key temporal and spatial features. The proposed algorithm outperforms existing methods in terms of the error percentage and accuracy of bus arrival time prediction.

Key words: arrival forecast; gradient boosting tree; ConvLSTM

随着我国城市的发展,私家车数量急剧增加,道路拥堵、车辆尾气排放造成环境污染等问题日益加剧[1].相比于私家车,公共交通工具具有承载量大、能源消耗较低、尾气排放相对较小等优势,对于缓解上述问题具有重要意义[2].相对于出租车等公共交通方式,公交具有投资成本更低、承载量更大,且覆盖范围更广等优势,成为城市出行的重要方式,提升其营运效率,是提升乘客满意度、吸引乘客使用该方式出行的必要手段[3,4].

目前,我国公交采取排班制发车,以达到公交公司和乘客之间的效益平衡.但由于道路交通、天气等因素复杂多变,导致车辆常常不能按照计划发车时间发班,进而会出现“串车”和“大间隔”现象[5].为应对各种原因导致车辆不能按原计划发班的情况,需要进行车辆的实时调度.现有的公交调度方式主要由手工完成,即公交调度员通过监视面板观察所负责线路的当前车辆分布状况.调度员根据自身经验估计车辆回场时间,进而进行下一班次发车时间的调整.现有调度方式仅依靠调度员的经验估计车辆到站时间,不仅工作量巨大,且常由于错误预估,导致调度策略无法被准确执行,仍无法缓解“串车”和“大间隔”现象的发生.图 1 给出了厦门市 11 路和 22 路公交线路 3 个月运行总时长的统计图(箱线图中线表示均值,其他线表示四分位点,圆圈表示异常点).从图中可以看出:同一线路方向,即使在同一时间段内的总行驶时长依然存在较大差异和异常点.因此,良好运行的调度系统迫切需要一个能相对准确地预测到站时间的算法,进而辅助调度员合理地预估车辆回场时间.这也是近年来智能交通(intelligent transportation system,简称 ITS)[6−8]应用的典型场景.

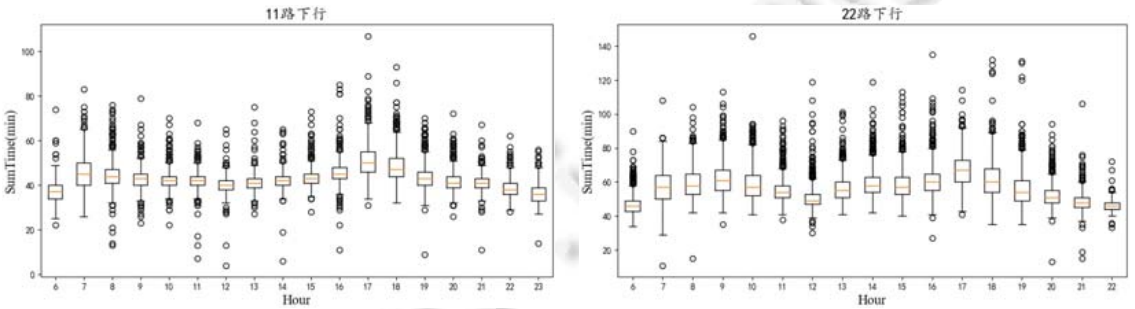


Fig.1 Distribution of bus travel time of route No.11 and 22 in Xiamen city

图 1 厦门市 11 路和 22 路公交行驶总时长分布图

公交车辆到站时间预测受站点停留和站间行驶时长等相关因素的影响,存在时空依赖性.从时间的角度看,无论是站点停留还是站间行驶,都具有一定的时间规律,这种规律可能是长期的历史规律、短期的周期规律或者近期的波动规律等.从空间的角度看,站点停留和站间行驶具有一定的空间规律.例如:对于站点停留来说,某一站用于乘客上车的停留时间必然影响另一站用于乘客下车的停留时间.对于站间行驶而言,相邻路段的行驶速度也相互影响.此外,天气状况、道路交通状况等难以准确预知,增加了预测的挑战性[9].已有的预测行程时长的方式包括两种,即分段预测[9,10]与全程预测[11,12].分段预测是指对路段进行划分预测,对预测时间加以累加.这

种方式可能产生的问题是误差累积,使得最终预测误差变大.全程预测是指直接预测从起点到达终点的时间的方式.这种方式可能产生的问题是当路径越长时,覆盖完整路径的轨迹点就越少,数据越稀疏导致结果往往不准确.由于公交通常具有“定线路”、“定站点”的特点,发车频率较快,同一时段有多辆同线路公交在运行,因此公交车辆本身即可收集各种交通数据,为即将发班的车辆预测提供依据.

近年来,深度学习技术在视觉、自然语言处理等领域得到了广泛应用.其中,卷积神经网络模型[13]已被证明可很好地捕捉空间规律,而循环神经网络[14]类似于一个存在记忆功能的神经网络,能够捕捉事物的时间规律.本文基于二者的融合,提出一种基于 ConvLSTM[15]模型,能够同时捕捉事物时间依赖性、空间相关性与外部影响因素的公交车到站时间预测算法——时空属性模型(spatio-temporal property model,简称 STPM).

本文的主要贡献包括:提出一种能够同时学习事物时空依赖性和外部因素影响的卷积时空属性模型,该算法通过时空组件捕获事物的时间依赖性和空间相关性,利用该组件分别学习与预测站点停留与站间行驶参数.通过属性组件,将外部因素如天气、时间、驾驶员、车辆、近期路段行驶状况等因素的复杂性融合到模型预测结果的考量中去.提出一种融合组件,将事物的时空特征与外部因素特征进行融合,预测公交车辆从起点到终点的总行驶时长.该算法可作为调度辅助决策系统的依据,帮助调度员及时发现晚点车辆,并作出合理的调度决策.在真实数据集上实现了该算法并进行了验证.实验结果表明:SPTM 算法能够更好地学习事物的时空特性,相对于单一使用卷积神经网络或循环神经网络而言,其预测准确率能提高至少 2.25 个百分点;且可以充分利用外部属性的因素,提高预测的准确率.

据我们所知,本文是首个利用深度学习算法进行公交车从起点到终点总时长预测的研究.第 1 节介绍国内外关于行程时长预测的研究.第 2 节介绍数据预处理过程.第 3 节对 STPM 算法进行详细展开和叙述.第 4 节利用真实公交到离站数据进行实验验证与分析.第 5 节总结全文并对未来工作进行展望.

1 相关工作

现有的关于行驶时长预测的方法主要可以分成基于传统方法和基于深度学习的方法.传统方法包括回归模型和卡尔曼滤波模型,利用历史数据和时间序列数据进行预测.Wu 等人[9]利用支持向量回归(support vector regression,简称 SVR)进行交通时长的预测.在该文献的研究中,通过使用过去 t 个时刻的真实交通时长数据,预测未来一段时间内的交通行驶时长.通过实验证明,该方法在预测旅行时间问题上具有一定可行性.但其在特征使用上,仅使用了过去时刻的数据,无法体现外在因素如驾驶员风格、车辆性能、道路交通状况的差异对预测结果的影响.Vanajakshi 等人[16]利用卡尔曼滤波技术预测不同交通条件下的出行时间.在该文献的研究中,对路段进行等距离划分,利用 2 辆前序车辆收集到的信息进行当前车辆的预测.在该方法中,假设任何时刻均有两辆前序车辆跑完全程为其收集信息,在实际应用中较难实现.Mathieu 等人[17]提出了一种用于预测到站时间的基于实时 GPS 数据的非参数算法,关键思想是,使用内核回归模型来表示位置更新与公交车站到达时间之间的依赖关系.实验结果表明:对于 50 分钟的时间范围,算法的预测误差平均小于 10%.在该文献的研究方法中,通过依据历史数据与当前状况的相似性为其赋予不同的权重,计算当前状况的预测值.这种方式对于模型训练的时间跨度要求较为严格,需要更长时间的数据样本.

近年来,深度神经网络逐渐应用到各个领域,包括行驶时长预测.Wang 等人[4]提出了一种宽深度递归(WDR)学习模型,预测在给定出发时间沿给定路线的行进时间.算法联合训练宽线性模型、深度神经网络和递归神经网络,以充分利用这 3 种模型的优势.但该方法并不考虑站点的停留时长,对路网的空间依赖考虑较少.Maiti 等人[18]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,结果表明所提出的 HD 模型分别比人工神经网络(artificial neural network,简称 ANN)模型和支持向量机(support vector machine,简称 SVM)模型执行速度更快,同时也具有比较高的预测精度.但这种方法仅利用车辆轨迹数据和时间序列数据,无法体现外在因素(如天气、驾驶员风格、车辆性能)等对结果的影响.王麟珠等人[19]提出一种基于 Elman 神经网络的公交车辆到站时间预测方法,并通过福州的公交数据进行验证.在该文献的研究方法中,以时间、天气、路段、当前路段的运行时间为特征进行预测,结果具有一定的精确度.但其对于天气的划分仅限于

是否下雨,未涉及到沙尘暴等影响能见度的天气因素,且其同样未考虑驾驶员风格等主观因素对于结果的影响,因此该方法还有待提升.张强等人[20]提出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.其将总时长分为站间行驶时长、站点滞留时长、交叉路口通行时长与等待时长,但其对这些组成因素的预测均采用基于历史时间序列进行预测,缺少对外在因素影响的研究.季彦婕等人[21]提出将粒子群算法与神经网络算法结合,从而减少预测误差.实验表明,该模型对于工作日与周末都有较高的预测精度.该文献在特征选择上同样是仅利用了历史数据,未考虑实际的外在因素影响.杨奕等人[22]将遗传算法与 BP 神经网络(back propagation neural network,简称 BPNN)结合,从而改进 BP 神经网络容易陷入局部最优的缺陷.通过对合肥某一公交线路数据的研究,实验表明,该算法确实有比较好的预测效果.但该算法在特征选择上仅利用最近一班班次的实际发生数据,缺少对于数据历史或周期规律的研究.此外,其同样忽略了外在因素的影响.张昕等人[23]提出利用遗传算法提升 SVM 的参数寻优效率,进而进行预测.该文献考虑道路因素、大型节假日、天气、路况、运行距离、运行时间、排班信息等 7 个因素的影响,能够更好地适应道路交通等的变化,具有较好的预测精度,但其忽略了司机或突发事件等因素的影响.谢芳等人[24]提出结合聚类与神经网络模型对车辆数据进行分段预测,并利用 Map Reduce 的并行化框架减少算法的计算时间.在该算法中,使用站点停留数据、站间行驶数据、星期几和是否为节假日作为特征,同样忽视了驾驶员、车辆、天气等因素对其的影响.

现有的关于行程预测或公交预测的研究无法被直接用于公交辅助调度决策,主要原因包括:由于公交车与出租车或私家车的行驶特点不同,其需要在固定站点进行停留,因此站点停留时长是其总时长的重要可变组成成分,直接利用出租车或私家车关于行程时长预测的方法往往不够准确.现有关于公交到离站的预测多基于对邻近站点的站数、距离或站间行驶时长的预测,缺少关于起点站到终点站总时长预测的研究.利用站间预测的方式预测总时长会出现误差累加以及无法预测各站点停留时长的情况.目前,对于行程时长的预测仅考虑某几个影响因素,缺乏对其时空特性和外在因素影响的融合研究.但公交车辆总时长的问题是一个包含时间依赖性、空间相关性和外部因素综合影响的复杂问题,需要一种能够直接对公交车辆从起点站至终点站总时长进行预测的算法.

2 数据预处理

数据预处理的过程如图 2 所示,包括线路静态特征处理、动态特征处理、天气特征处理以及缺失数据填充等步骤,最后得到特征数据集.

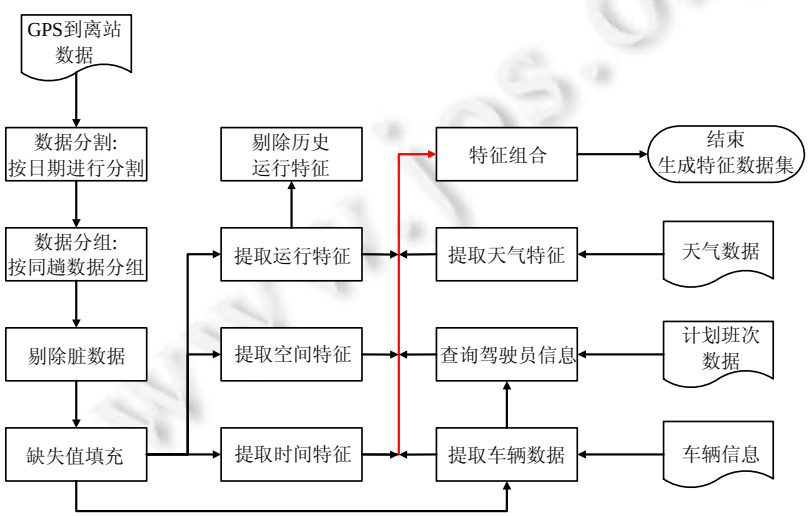


Fig.2 Data preprocessing process

图 2 数据预处理流程

2.1 线路静态特征处理

(1) 数据清洗

原始到离站数据文件包括一段时间内公交线路收集到的所有车辆进站及出站信息.进行特征提取前,需先将数据文件按日期进行分割,并将分割后的每日数据根据上下行、车辆 id、获取到离站数据的时间进行排列.然后将同趟数据划分到同组中,剔除组内脏数据.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、到达后一站的时间在到达前一站之前的记录以及同一趟的到离站数据大量丢失的记录等.

(2) 线路特征处理

线路特征包括线路、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 id 查询驾驶员 id,以车辆 id 和驾驶员 id 分别作为车辆特征和驾驶员特征的区分.

(3) 时间特征处理

公交车辆的运行时长具有一定的时间规律,比如工作日与周末、节假日与平时、高峰期与平峰期的区别等.为了研究以上时间因素对于研究结果的影响,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、星期几、是否为节假日、是否为工作日等特征.

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数据特征,包括站间行驶时长和站点停留时长.根据线路运行具有长期趋势、周期规律、短期影响等特征,本文将线路动态特征定义为近 1 周、近 3 天、最近一段时间内的站间行驶时长和站点停留时长.其中,近 1 周、近 3 天的站间行驶或站点停留时长是根据相同时段下各车辆实际运行时对应数据的均值计算所得.最新各个站点的停留时长与行驶时长通常是从起点站依次查询经过该站最近一班的车辆,以该车辆在该站的站间运行时长和站点停留时长作为该站最新运行时长和停留时长的代替,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.

2.3 天气特征处理

在本文获得的原始天气数据中,利用到的主要字段包括时间、能见度与天气描述.其中,时间需要处理成为日期与时间段,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.

2.4 缺失数据填充

由于数据期间的局限以及部分 GPS 到离站原始数据的丢失,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存在缺失,需要进行填充.在本文的缺失数据填充中,对于站点停留和站间行驶时长的缺失主要是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,则用临近班次进行填充.对于天气数据的缺失,主要是利用临近小时内的天气状况进行代替.

2.5 特征数据集

按照图 2 的流程完成数据预处理,将预处理过程中提取到的空间特征、时间特征和外部特征(包括天气特征、驾驶员信息和车辆信息等)合并为最终的特征数据集.特征数据集的维度为 11202×51,包含厦门 22 路公交车在 2018 年 9 月、12 月和 2019 年 1 月、2 月这 4 个月里的所有班次.其中,每一行表示一个班次内的所有特征值.表 1 展示了特征数据集中的部分特征及其说明.

3 STPM 算法

3.1 算法概述

公交车辆的站点停留时长和站间行驶时长可能受不同因素的影响,呈现不同的数据表征.本文首先利用融合卷积与 LSTM 特点的 ConvLSTM[15]分别预测车辆的站间行驶时长和站点停留时长,再利用 Stack-LSTM[25]预测总行驶时长.算法分为 3 个部分:时空组件、外部属性组件和融合组件,整体的网络结构如图 3 所示.

(1) 时空组件用于处理站点停留或是站间行驶等相关特征,捕捉其时间依赖性与空间相关性;

(2) 外部属性组件用于处理与模型有关的外部因素,诸如星期几、发车时间、天气、是否是工作日、是否是节假日、驾驶员、车辆、近期运行状况等,其输出将作为融合组件输入的一部分;

(3) 融合组件是根据时空组件与外部属性组件的输出估计车辆的总行驶时长.

Table 1 Feature in dataset and their meaning

表 1 数据集中的特征及其说明

特征	解释	特征	解释
carID	车辆 ID	isworkday	是否工作日
datetime	车辆发车时间	minuted	发车时间的分钟数
direction	方向	pdistance	预定的行驶距离
driverid	驾驶员 ID	routeid	线路 ID
everyruntime	车辆在每一站的站间行驶时长	sunruntime	实际从起点站到终点站的时长
everystaytiime	车辆在每一站的站点停留时长	time_group	时间分组
fdate	发车日期	weatherid	天气 id
fsitenum	计划站点数量	weekdayid	星期几
fstime	发车时间	lats	站点的经度集合
groupid	发车时间所在的段内分组	lngs	站点的纬度集合
hourid	发车时间的小时数	sum_runtime	车辆的总运行时长
isholiday	是否节假日	sum_staytime	车辆的总停留时长

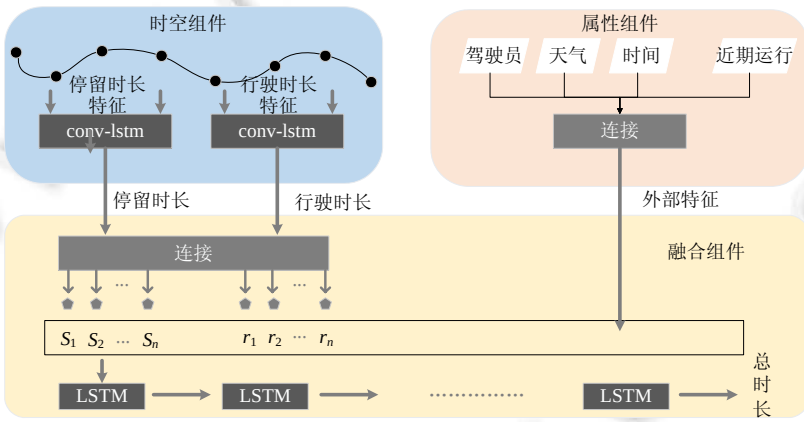


Fig.3 STPM model architecture

图 3 STPM 模型架构

模型中涉及的符号说明见表 2.

3.2 属性组件

公交车辆的行驶时长受多种复杂因素的影响,包括:

(1) 时间信息:如公交车辆的发车时刻点、星期几、是否是工作日、是否是节假日等.对于公交车辆来说,通常具有时段特征,即一天中存在高峰期、低峰期与平峰期.高峰期的行驶时长通常高于低峰期的行驶时长,且由于周末或其他节假日出行人数较多,道路私家车数量相对于工作日有所增长,更易出现交通拥堵增加车辆行驶时长的现象;

(2) 天气信息:当天气出现大雾、沙尘暴、暴雨等情况时,由于其会影响道路的能见度,进而减慢道路通行速度,使得通行时长要比其他情况要长;

(3) 驾驶员信息:由于驾驶员行驶习惯的差异,即使是同一路段,通行时长亦有所不同;

(4) 车辆信息:不同车的性能、能容纳的乘客数、车辆结构的差异亦会影响车辆的行驶与停留消耗时间;

(5) 近期交通状况信息:由于现在城市发展较为迅速,道路交通状况处于不断变化的过程中,因此有必要

将近期线路的交通运行状况作为外部属性输入到模型当中.

Table 2 Symbols in STPM and their meanings

表 2 STPM 中的符号及其含义

符号	含义	符号	含义
x_{mean}	x 的均值	x_{std}	x 的标准差
$attr_i$	第 i 个样本的外部属性特征	C_i^c	第 i 个样本的车辆特征 C 经嵌入方法转换后的一个 c 维的向量
D_i^d	第 i 个样本的驾驶员特征 D 经嵌入方法转换后的一个 d 维的向量	W_i^w	第 i 个样本的天气特征 W 经嵌入方法转换后的一个 w 维的向量
DT_i^{dt}	第 i 个样本的星期几特征 DT 经嵌入方法转换后的一个 dt 维的向量	H_i^h	第 i 个样本的时段特征 H 经嵌入方法转换后的一个 h 维的向量
M_i^m	第 i 个样本的分钟特征 M 经嵌入方法转换后的一个 m 维的向量	SW_i	第 i 个样本近一周的运行特征经标准化方法转换后的一个值
ST_i	第 i 个样本近三天的运行特征经标准化方法转换后的一个值	SC_i	第 i 个样本最新的运行特征经标准化方法转换后的一个值
r_i	预测的第 i 站的行驶时长	s_i	预测的第 i 站的停留时长
w_s	第 i 站停留时长的重要性	$\max_{s_i}$	第 i 站停留时长的最大值
$\min_{s_i}$	第 i 站停留时长的最小值	var_{s_i}	第 i 站停留时长的方差
$mean_{s_i}$	第 i 站停留时长的平均值	w_{r_i}	第 i 站行驶时长的重要性
var_{r_i}	第 i 站行驶时长的方差	$mean_{r_i}$	第 i 站行驶时长的平均值
$\hat{y}$	Y 的预测值	y	Y 的真实值

在属性组件中,对于车辆、驾驶员、天气、时间等外部特征,采用嵌入的方式将这些外部属性特征转换为低维实向量.这种处理方式的优势在于:其一,它可以将分类值较多的特征降维到较小的输入维度,进而提高运算效率;其二,已有的相关研究发现,具有相似语义的范畴值通常会被嵌入到相近的位置,使得这种方法在本文的研究中有助于发现不同的运营数据之间相似的部分,如驾驶员的驾驶风格、车辆属性等.

对于近期运行状况特征,包括近 1 周、近 3 天、最新的各站点的站间行驶和站点停留特征,利用其均值与标准差进行标准化.假设近期运行状况中一个特征为 x,利用公式(1)对其进行转换得到 $\tilde{x}$:

$$\tilde{x}=\frac{x-x_{mean}}{x_{std}} \tag{1}$$

最后将经过嵌入和标准化处理的外部特征进行连接,作为其他组件的外部属性输入.若经过嵌入方法转换后,车辆属性特征为 C,驾驶员属性特征为 D,天气属性为 W,星期几属性为 DT,时间段属性为 H,段内分组属性为 M;经过标准化的方式转换后,近一周的总时长特征为 SW,近 3 天的总时长特征为 ST,最新的总时长特征为 SC,则经过连接后的输出向量 $attr$ 为

$$attr_i=C_i^c\circ D_i^d\circ W_i^w\circ DT_i^{dt}\circ H_i^h\circ M_i^m\circ SW_i\circ ST_i\circ SC_i \tag{2}$$

其中,$attr_i$ 表示第 i 个样本经过属性组件后输出的属性特征集;C_i^c 表示第 i 个样本的车辆特征经嵌入方法转换后的一维向量,其他属性特征(驾驶员、星期几、时间段、段内分组)的表示与之相同;SW_i 表示第 i 个样本近一周的运行特征经标准化方法转换后的一个值,3 天与最新运行特征的转换与之相同.故经过属性组件后的特征都被处理为一个 $n=c+d+w+dt+h+m+3$ 的一维向量.

3.3 时空组件

在时空组件中,相当于是由两个子件组成.其中一个子件用于预测各个站点的停留时长,另一个子件用于预测各个站间的行驶时长.利用 ConvLSTM 模型捕获站间行驶与站点停留的时间依赖性与空间相关性.虽然行驶时长预测子件与停留时长预测子件在整体结构上存在差异,但其核心均是依赖于 ConvLSTMCell.

(1) ConvLSTMCell

ConvLSTMCell 是基于 LSTM 结构的.LSTM 作为一种特殊的 RNN,它的时间记忆性能够在一定程度上解决时间依赖的问题.一个抽象的 LSTM 如图 4 所示(各符号的定义见表 1).

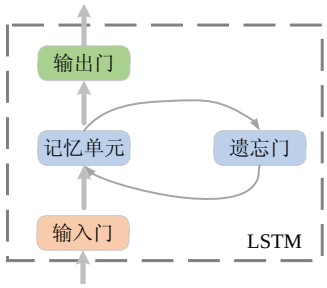


Fig.4 LSTM structure

图 4 LSTM 结构

当有新的输入时,如果输入门 i_t 被激活,则新输入的信息将会被累加到细胞单元中.输入门是否被激活,是由 t–1 时刻的网络输出 h_{t-1} 和这一步的网络输入 x_t 决定(见公式(3)).在计算这一步的网络输出前,还需要考虑 t–1 时刻网络中的记忆单元 c_{t-1}.当 c_{t-1} 传入到 t 时刻的网络中时,首先,网络需要先决定它被遗忘的程度,即,将 t 时刻之前的记忆状态乘以一个记忆衰减系数 f_t.这个记忆衰减系数 f_t 是根据 t 时刻的网络输入 x_t 与 t–1 时刻的网络输出 h_{t-1} 所决定(见公式(5)).也就是说,网络所要保留的记忆是由前一时刻的输出和这一时刻的输入共同决定.

新时刻学到的记忆 c_t 是经过线性变化和激活函数所得到(见公式(4)).在得到 t–1 时刻的记忆需要保留多少 f_t*c_{t-1} 以及新时刻学到什么样的记忆后,将 t–1 时刻保留的记忆加上 t 时刻学到的记忆及其对应的衰减系数 i_t,则得到了 t 时刻的记忆状态(见公式(6)).

t 时刻网络的输出 h_t 是由 t 时刻的输入 x_t、t–1 时刻网络的输出和 t 时刻记忆状态所决定.使用类似计算记忆衰减系数的方式计算输出门的系数 o_t(见公式(7)),由决定网络的输出,即最终网络的输出是由公式(8)所计算:

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{3}$$

$$\tilde{c}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{4}$$

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{5}$$

$$c_t=f_t\times c_{t-1}+i_t\times\tilde{c}_t \tag{6}$$

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{7}$$

$$h_t=o_t\times\tanh(c_t) \tag{8}$$

对于公交到站时间预测,有必要使用一个能够同时捕捉时空特征的网络结构.在 ConvLSTM 中,通过将卷积层融入到传统的 LSTM 中,使得某一单元的输入不再是仅由过去时刻的状态所决定,还与其邻近的邻居状态有关(如图 5 所示).可以理解为:某一站点的停留时长或站间的行驶时长,不仅与这一站点过去的停留时长或站间行驶时长有关,还与其邻近站点的停留时长或站间行驶时长有关.

ConvLSTM 上述特性的实现依赖于其将卷积的操作融入到 LSTM 各个门的计算当中,即在以下公式中的 $W_{xi},W_{hi},W_{xf},W_{hf},W_{xc},W_{hc},W_{xo},W_{ho}$ 等参数中利用卷积操作进行运算(各符号的定义见表 1).

$$i_t=\sigma(W_{xi}\times x_t+W_{hi}\times h_{t-1}+W_{ci}\circ c_{t-1}+b_i) \tag{9}$$

$$f_t=\sigma(W_{xf}\times x_t+W_{hf}\times h_{t-1}+W_{cf}\circ c_{t-1}+b_f) \tag{10}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ\tanh(W_{xc}\times x_t+W_{hc}\times h_{t-1}b_c) \tag{11}$$

$$o_t=\sigma(W_{xo}\times x_t+W_{ho}\times h_{t-1}+W_{co}\circ c_{t-1}+b_o) \tag{12}$$

$$h_t=o_t\times\tanh(c_t) \tag{13}$$

(2) 行驶时长子件

对于行驶时长子件,其可以看做由多个 ConvLSTMCell 组成的一个 ConvLSTM.对于 ConvLSTM 来说,其卷积核的大小从某种程度上反映了一个单元的状态由多大范围内的邻近单元状态决定.对于行驶时长子件,其网络结构如图 6 所示.

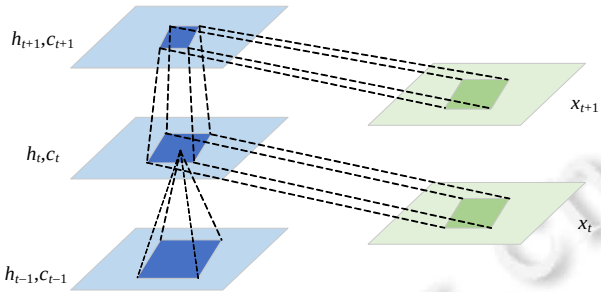


Fig.5 Convolution structure of ConvLSTM

图 5 ConvLSTM 的卷积结构

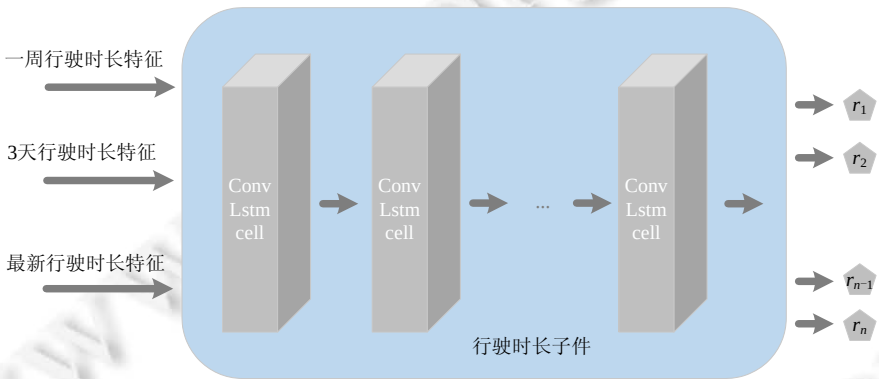


Fig.6 Structure of the driving time component

图 6 行驶时长子件结构

(3) 停留时长子件

对于停留时长子件,同样可以看做由多个 ConvLSTMCell 组成的一个 ConvLSTM.其网络结构可以如图 7 所示.

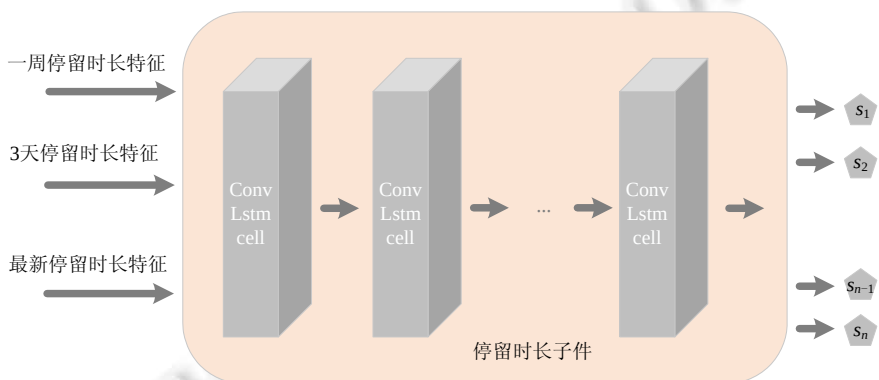


Fig.7 Structure of the staying time component

图 7 停留时长子件结构

3.4 融合组件

融合组件将根据行驶时长组件和停留时长组件输出的预测各站点的停留时长参数和行驶时长参数,以及属性组件经嵌入、标准化和连接方式输出的驾驶员特征、天气特征、时间特征和近期运行特征,预测公交车辆

由起点站到终点站所需的总时长.融合组件采取 Stack-LSTM 的方式进行预测,其网络结构如图 8 所示.

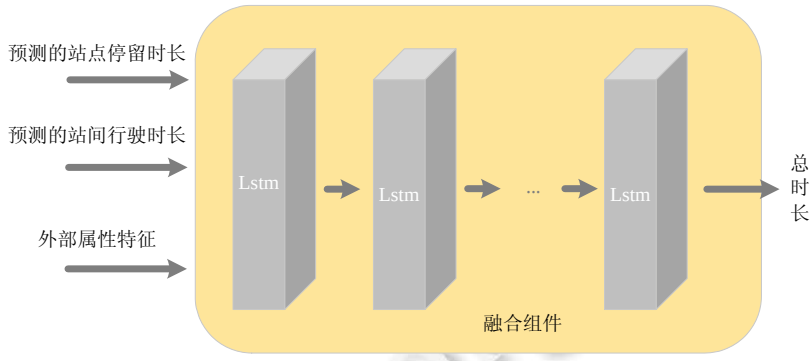


Fig.8 Structure of the fusing component

图 8 融合组件结构

4 实验与分析

4.1 环境设置

本次实验基于 2018 年 9 月 1 日~2019 年 2 月 28 日厦门市 22 路公交到离站数据、计划班次信息、车辆信息以及厦门天气状况等数据进行.使用 python3.7 和 Pytorch1.1 进行数据处理与算法编写,利用 sklearn0.18.1 库中的模型接口进行部分算法实现.如图 9 所示:厦门 22 路公交上行从胡里山站出发,途经厦大白城站等站点,终点为麦德龙站.预处理后的特征数据集包含到离站数据 280 050 条,班次信息 11 202 条,天气数据为每小时天气状况.

Fig.9 No.22 bus route of Xiamen

图 9 厦门市 22 路公交线路图

4.2 模型设置

4.2.1 时空组件——停留时长预测子件

本节在实验中修改网络的结构,包括网络层数、每层神经元的数量、卷积层核的大小、每次参与训练的

batch的数量以及网络的学习率.通过实验发现,当停留时长预测子件的参数如下时,模型效果最优:网络结构由4层 ConvLSTMCell 组成;每层神经元的设置分别是 128,64,32,1;各层卷积核的大小分别是 7,5,5,3;每次进行训练的 batch 大小为 600;学习率为 0.01.此外,为了体现不同站点对最终结果的影响程度不同,在该网络中,为不同站点的损失赋予不同的权重(公式(14)):

$$w_{s_i}=(\max_{s_i}-\min_{s_i})\times\frac{\mathrm{var}_{s_i}}{mean_{s_i}} \tag{14}$$

4.2.2 时空组件——行驶时长预测子件

经实验发现,当行驶时长预测组件由一个 3 层 ConvLSTMCell 组成,各层参数设置值如下时效果最优:该网络每层神经元的设置分别是 128,64,1;各层卷积核的大小分别是 5,5,5;每次进行训练的 batch 大小为 600;学习率为 0.01.在该网络中,不同站点损失的权重将根据公式(15)进行计算:

$$w_{ti}=\frac{\mathrm{var}_{t_i}}{mean_{t_i}} \tag{15}$$

4.2.3 属性组件

属性组件对于其中的驾驶员、车辆、天气、星期几、发车时间、时段分组特征,模型采用嵌入的方式,对其纬度变化见表 3.

Table 3 Attribute feature latitude change

表 3 属性特征纬度变化

特征	输入纬度	输出纬度
carid (*C*)	20	3
driverid (*D*)	39	3
weekdayid (*DT*)	7	3
hourid (*H*)	24	5
minuteid (*M*)	60	6
weatherid (*W*)	9	4

4.2.4 融合组件

在本节的实验中,融合组件在其参数设置如下时,达到最优状态:层数为 4;各层神经元的数量分别为 96,48,24,1;学习率为 0.1.在该网络结构中,其损失函数为公式(16):

$$loss=|\hat{y}-y| \tag{16}$$

4.3 实验结果分析

由于现有关于公交车辆从起点站到终点站的研究较少,故本文采用几种常用于回归预测的机器学习方法进行误差百分比和准确率两个方面的对比.其中,准确率是绝对误差在 5 分钟以内被判定为正确的数量占测试集的比例.其实验结果见表 4.

(1) STPM:该算法为本文所提出的算法,融合卷积与 LSTM 组成时空组件 CT,用于学习事物的时间依赖性和空间相关性;利用嵌入、标准化与连接的方式组成属性组件,用于学习外部因素的影响;利用融合组件对时空组件与属性组件的输出进行融合进而预测总行驶时长;

(2) CNN[13,26]:由卷积层和池化层组成的一种前馈神经网络,在对比实验中,使用与 STPM 相同的特征进行预测;

(3) LSTM[14,27]:由输入门、输出门、遗忘门组成的基本长的短时记忆网络,使用的特征与 STPM 相同;

(4) Adaboost[28]:利用训练数据训练多个弱分类器,融合弱分类器的的训练结果进行预测的一种集成方法.在对比实验中,使用与 STPM 相同的特征进行预测;

(5) DecisionTree[29]:决策树,一种用于预测的树结构.在该算法实验中,使用与 STPM 相同的特征进行预测;

(6) SVM[30]:支持向量机.一种常用于模式识别、分类及回归分析的监督学习方法;

(7) HP:基于历史相同条件(是否为工作日、时段)下平均值预测.

Table 4 Comparative experiment results

表 4 对比实验结果

模型名	误差百分比(%)	准确率(%)
STPM	**5.68**	**80.23**
CNN	9.94	58.69
LSTM	6.05	77.98
Adaboost	7.32	66.97
DecisionTree	8.49	64.97
SVM	9.69	57.05
HP	29.15	1.18

从实验结果上看,本文所提出的基于深度神经网络的预测算法 STPM 的误差百分比相对于其他算法更低.图 10 与图 11 展示了利用传统方法与深度学习方法进行预测的结果,其中,纵轴为总时长值,横轴为样本 ID,拟合线为预测值,散点为真实值.根据真实值与预测值的绝对误差,对真实值的散点颜色进行区分.十字代表 300s 的误差内,圆圈代表 300s~400s 的误差,叉号代表 400s~500s 的误差,三角代表 500s~600s 的误差,方形代表误差超过 600s.

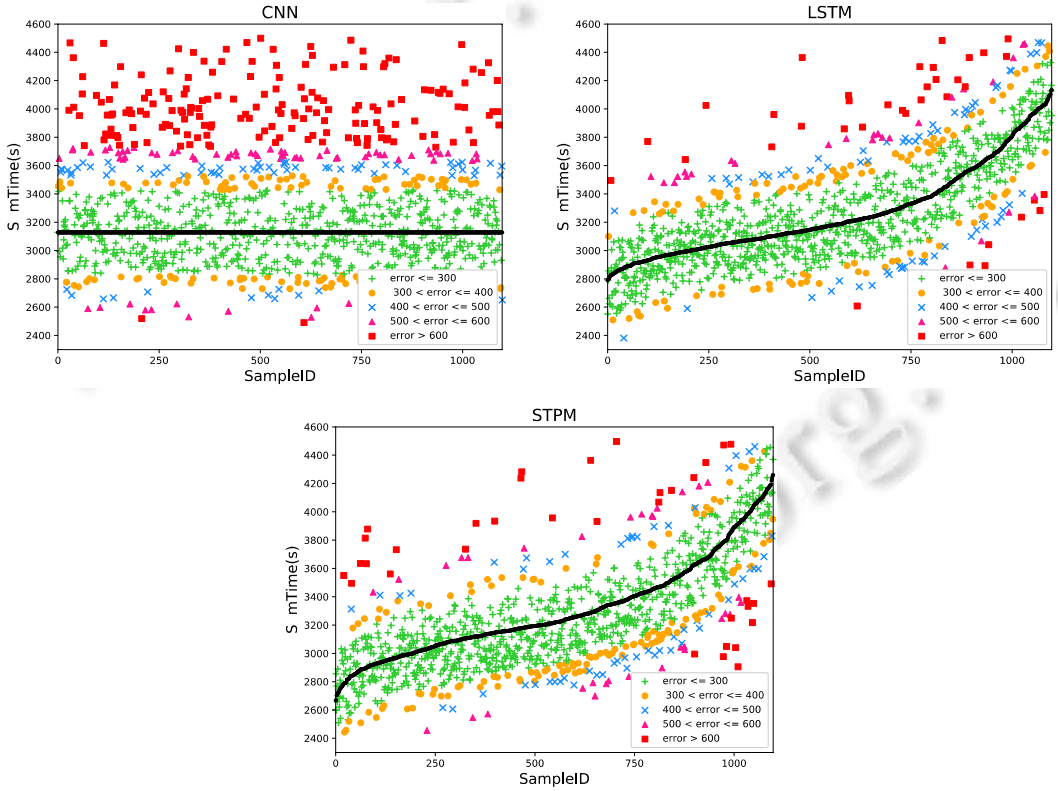


Fig.10 Comparison of prediction results of deep learning methods

图 10 深度学习方法预测结果对比

实验结果分析如下.

(1) 对比 STPM 与 HP 发现,STPM 的效果要优于单纯根据历史条件进行预测的模型 HP.这是因为公交总行驶时长是一个受多种复杂因素综合影响的问题,天气与道路交通状况等因素不可预知,仅根据历史条件进行预测缺乏灵活性;

(2) 对比 STPM,CNN 与 LSTM 发现:STPM 在误差百分比与准确率上要优于 CNN 与 LSTM,这可能是由于 STPM 结合了 CNN 卷积与 LSTM 的记忆优势.公交总行驶时长,无论停留还是行驶,都具有时间依

赖性和空间相关性.因此,使用一个可以同时捕获时空特征的 ConvLSTM,相对于单用一个模型较好;

(3) 对比 STPM 与其他机器学习模型发现:其误差百分比与准确率均优于其他模型;且 Adaboost 的效果要好于 SVM、决策树.这是由于 Adaboost 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.

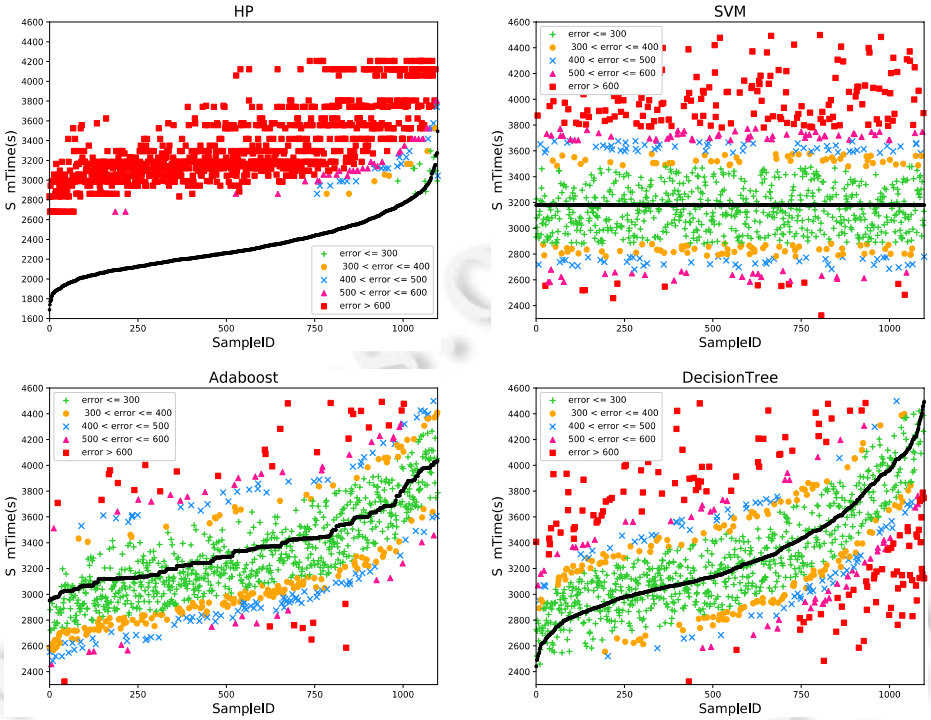


Fig.11 Comparison of prediction results of conventional methods

图 11 传统方法预测结果对比

4.4 时间性能分析

表 5 展示了上述 7 种方法的训练时间(即 STPM,CNN,LSTM,HP,SVM,Adboost,DecisionTree),训练过程使用的 GPU 为 TITAN X.

Table 5 Comparative training time and testing time

表 5 训练时间、模型参数和预测时间的对比

模型名	训练时间(s)	预测时间(s)	参数数量(M)
STPM	**1 944.5**	**0.079 6**	**7.839**
CNN	306.8	0.005 6	0.008
LSTM	560.8	0.055 9	0.194
Adaboost	23.9	0.019 2	–
DecisionTree	29.9	0.017 7	–
SVM	24.2	0.987 6	–
HP	19.1	0.008 2	–

由于 STPM 模型的参数数量大于其他的深度模型,因此其训练也是最耗时的.然而通过实验结果可以看出,STPM 的精度好于其他方法.虽然具有较长的训练时长,但是在离线训练中,这样的时长是可以接受的.通常在实际应用当中,我们能够获得大量的离线资源进行预训练.此外,当模型训练完毕之后,STPM 的预测时间与其他方法相当,可在 80ms 内得到预测结果.

5 结束语

深度学习现已在人脸识别、计算机视觉、自然语言处理等领域发挥着重要的作用,是各界学术研究者研究的热点之一.但是在智能交通领域,特别是公交到站时间预测的研究与应用还较少.本文提出利用 ConvLSTM 捕获事物的时间依赖性与空间相关性,分别对站点的停留时长和站间的行驶时长进行预测,利用属性组件对诸如驾驶员特征、车辆特征、时间特征、天气特征、近期运行等特征进行嵌入操作,将时空组件与属性组件的输出作为融合组件的输入.然后,由多个 LSTM 组成的融合组件对来自时空组件和属性组件的输入进行融合,预测最终车辆从起点站到终点站的总时长.实验结果表明,算法在误差百分比与准确率上的表现优于已有的算法.

对于未来的工作,可以增加非起点站到终点站的预测,辅之以修正机制.通过车辆到达非终点站的真实时间与预测时间之间的误差,不断改进对到达终点站的预测时间.同时,由于道路交通状况常常是一个复杂且不可准确预知的问题,特别是何时会发生交通事故等难以预料,因此,加强交通系统中车辆信息收集与相互之间的信息交互,对于及时了解路段状况、改进模型预测效果也是有一定帮助的.

References:

[1] Feng K. Development status and trends of urban intelligent transportation system. Global Market Information Guide, 2017,(1): 111−112 (in Chinese with English abstract).

[2] Geng YB. Quantitative research on low carbon index of urban public transport. Low Carbon World, 2018,185(11):28−29 (in Chinese with English abstract).

[3] Yang ZS. Urban Intelligent Public Transportation System Theory and Method. Beijing: China Railway Publishing House, 2004 (in Chinese).

[4] Wang Z, Fu K, Ye J, *et al*. Learning to estimate the travel time. In: Proc. of the 24th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2018. 858−866.

[5] Ma Q. The applied research of intelligent transportation system in Luoyang's public transportation filed [Ph.D. Thesis]. Zhengzhou: Henan University Of Science And Technology, 2013 (in Chinese with English abstract).

[6] Zhang J, Wang F, Wang K, *et al*. Data-driven intelligent transportation systems: A survey. IEEE Trans. on Intelligent Transportation Systems, 2011,12(4):1624−1639.

[7] Ghosh R, Pragathi R, Ullas S, *et al*. Intelligent transportation systems: A survey. In: Proc. of the Int'l Conf. on Circuits. 2017.

[8] Li YG, Fu K, Wang Z, Shahabi C, Ye JP, Liu Y. Multi-task representation learning for travel time estimation. In: Proc. of the 24th ACM SIGKDD Conf. on Knoledge Discovery and Data Mining. 2018. 1695−1704.

[9] Wu CH, Ho JM, Lee DT. Travel-Time prediction with support vector regression. IEEE Trans. on Intelligent Transportation Systems, 2004,5(4):276−281.

[10] Lai YX, Yang X, Cao Q, *et al*. A bus running length prediction method based on gradient boosting. Big Data Research. 2019,5(5): 58−78 (in Chinese with English abstract)

[11] Ding HF, Li YH, Liu B, *et al*. Expressway's travel time prediction based on combined BP neural network and support vector machine approach. Application Research of Computers, 2016,33(10):2929−2932 (in Chinese with English abstract).

[12] Xu TD, SunLJ, Hao Y. Real-time traffic state estimation and travel time prediction on urban expressways. Journal of Tongji University (Natural Science), 2008,36(10):1355−1361 (in Chinese with English abstract).

[13] Le CY, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989,1(4):541−551.

[14] Sundermeyer M, Schluter R, Ney H, *et al*. LSTM neural networks for language modeling. In: Proc. of the Int'l Conf. on Speech Communication Association. 2012. 194−197.

[15] Shi X, Chen Z, Wang H, *et al*. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proc. of the Neural Information Processing Systems 2015. 2015.

[16] Vanajakshi L, Subramanian SC, Sivanandan R. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. IET Intelligent Transport Systems, 2009,3(1):1−10.

[17] Sinn M, Yoon J W, Calabrese F, *et al*. Predicting arrival times of buses using real-time GPS measurements. In: Proc. of the Int'l IEEE Conf. on Intelligent Transportation Systems. IEEE, 2012.

[18] Maiti S, Pal A, Pal A, *et al*. Historical data based real time prediction of vehicle arrival time. In: Proc. of the IEEE Int'l Conf. on Intelligent Transportation Systems. IEEE, 2017.

[19] Wang LZ, Su QL, Zheng RB. Bus arrival time prediction based on elman dynamic neural network. Mechanical & Electrical Technology, 2012,35(1):135−137 (in Chinese with English abstract).

[20] Zhang Q, Zhang YY. Research on prediction model of bus arrival time based on time segmentation. Digital Technology and Application, 2014,(11):60,62 (in Chinese with English abstract).
[21] Ji YJ, Lu JW, Chen XS, *et al*. Prediction model of bus arrival time based on particle swarm optimization and wavelet neural network. Journal of Transportation Systems Engineering and Information Technology, 2016,16(3):60−66 (in Chinese with English abstract).
[22] Yang Y, Zhang WR, Zhang C. Application of BP neural network based on genetic algorithm in bus arrival time prediction. Modern Business, 2017,(16):38−40 (in Chinese with English abstract).
[23] Zhang X, Jiang JJ, Liu J. Coach bus arrival time prediction based on SVM and GA. Computer & Digital Engineering, 2017,45(6): 1062−1066, 1085 (in Chinese with English abstract).
[24] Xie F, Gu JH, Zhang SQ, *et al*. Predicting model of bus arrival time based on MapReduce clustering and neural network. Journal of Computer Application, 2017,37(S1):118−122 (in Chinese with English abstract).
[25] Dyer C, Ballesteros M, Ling W, *et al*. Transition-based dependency parsing with stack long short-term memory. Computer Science, 2015,37(2):321−332.
[26] Kim P. Convolutional neural network. MATLAB Deep Learning, 2017. 121−147.
[27] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735−1780.
[28] Dietterich TG. Ensemble Methods in Machine Learning. In: Proc. of the Int'l Workshgp on Multiple Classifier Systems. 2000.
[29] Quinlan JR. Induction on decision tree. Machine Learning, 1986,1(1):81−106.
[30] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995,20(3):273−297.

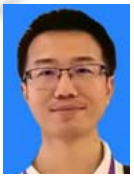
附中文参考文献:

[1] 冯凯.城市智能交通系统的发展现状与趋势.环球市场信息导报,2017,(1):111−112.
[2] 耿宇博.城市公共交通工具低碳指标量化研究.低碳世界,2018,185(11):28−29.
[3] 杨兆升.城市智能公共交通系统理论与方法.北京:中国铁道出版社,2004.
[5] 马卿.智能交通系统在洛阳市公共交通领域的应用研究[博士学位论文].郑州:河南科技大学,2013.
[10] 赖永炫,杨旭,曹琦,等.一种基于 Gradient Boosting 的公交车运行时长预测方法.大数据,2019,5(5):58−78.
[11] 丁宏飞,李演洪,刘博,等.基于 BP 神经网络与 SVM 的快速路行程时间组合预测研究.计算机应用研究,2016,33(10):2929−2932.
[12] 徐天东,孙立军,郝媛.城市快速路实时交通状态估计和行程时间预测.同济大学学报(自然科学版),2008,36(10):1355−1361.
[19] 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.机电技术,2012,35(1):135−137.
[20] 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.数字技术与应用,2014,(11):60,62.
[21] 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.交通运输系统工程与信息,2016,16(3):60−66.
[22] 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.现代商业,2017,(16):38−40.
[23] 张昕,姜佳佳,刘进.基于 SVM+GA 的客运车辆到站时间预测.计算机与数字工程,2017,45(6):1062−1066,1085.
[24] 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.计算机应用,2017,37(S1):118−122.

赖永炫(1981－),男,福建龙岩人,博士,副教授,CCF 专业会员,主要研究领域为大数据分析和管理,智能交通,深度学习,车载网络数据管理.

卢卫(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为大数据分析与管理.

张璐(1994－),女,硕士,主要研究领域为数据分析与挖掘.



王田(1982－),男,博士,教授,主要研究领域为边缘计算,人工智能.

杨帆(1982－),男,博士,副教授,主要研究领域为人工智能,机器学习,数据挖掘.