

文章编号: 1009-6744(2019)02-0223-08

中图分类号: U491

文献标志码: A

DOI:10.16097/j.cnki.1009-6744.2019.02.032

基于深度学习的网约车供需缺口短时预测研究

谷远利¹, 李萌¹, 芮小平^{*2}, 陆文琦¹, 王硕¹

(1. 北京交通大学 综合交通运输大数据应用技术交通运输行业重点实验室, 北京 100044;

2. 河海大学 地球科学与工程学院, 南京 211100)

摘要: 城市不同区域网约车供需缺口预测可为车辆调度策略提供支持, 从而提高车辆运行效率和乘客服务水平。为实现网约车供需缺口短时预测, 提出一种基于时空数据挖掘的深度预测模型(Spatio-Temporal Deep Learning Model, S-TDL)。该模型由时空变量模型、空间属性变量模型和环境变量模型3个子模型融合而成, 可捕捉时空关联性、区域差异性和环境变化对供需缺口的影响。同时, 提出特征聚类—最大信息系数两阶段特征选择方法, 筛选与供需缺口相关性强的特征变量, 提高训练效率, 减少过拟合。滴滴出行实例分析证明, 特征选择后的S-TDL模型预测精度显著优于BP神经网络、长短期记忆网络和卷积神经网络。

关键词: 城市交通; 供需缺口预测; 深度学习; 网约车; 时空关联性

Short-term Forecasting of Supply-demand Gap under Online Car-hailing Services Based on Deep Learning

GU Yuan-li¹, LI Meng¹, RUI Xiao-ping², LU Wen-qi¹, WANG Shuo¹

(1. Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, Beijing 100044, China; 2. School of Earth Sciences and Engineering, Hohai University, Nanjing 211000, China)

Abstract: The results of supply-demand gap prediction for online car-hailing services in different areas can provide support for online car-hailing scheduling system, thereby improving efficiency and service levels. In order to realize the short-term forecast of supply-demand gap for online car-hailing services, this paper proposes a novel spatio-temporal deep learning model (S-TDL). The model is composed of three sub-models: spatiotemporal variable model, spatial attribute variable model and environment variable model. It can capture the impact of spatio-temporal correlation, regional difference and environmental change on supply-demand gap. Moreover, a feature selection method named feature clustering-maximum information coefficient two-stage feature selection is proposed to screen out the important features which are strongly correlated with the supply-demand gap, improve training efficiency. The experimental results show that the S-TDL model after feature selection achieves the better performance than the existing methods.

Keywords: urban traffic; supply-demand gap forecasting; deep learning; online car-hailing; spatio-temporal correlation

0 引言

随着互联网技术提升, 基于智能出行平台的网约车服务迅猛发展, 广泛应用于居民日常出行。与传统出租车行业相比, 网约车服务方便、快捷, 但同

样存在供需不平衡问题。实现城市不同区域网约车供需缺口短时预测可指导服务平台有效制定车辆调度策略, 鼓励司机从供给过剩区域行驶至供给不足区域, 提高车辆利用效率, 减少乘客等待时间。

收稿日期: 2018-10-08

修回日期: 2018-11-25

录用日期: 2018-12-05

基金项目: 国家自然科学基金(National Natural Science Foundation of China(41771478))。

作者简介: 谷远利(1973-), 男, 辽宁海城人, 副教授, 博士。

*通信作者: ruixpsz@163.com

出租车与网约车市场相似,其供需预测研究可为网约车提供借鉴^[1]。供需预测研究包括供给预测、需求预测和供需缺口预测。供给预测主要研究出租车规模发展趋势,预测方法包括供需平衡法、比例法、指数平滑法等。需求预测是乘客需求的时空分布预测,其预测方法与交通流预测类似,可划分为3类:参数方法,如差分自回归移动平均^[2]、加权时变泊松模型^[3];非参数方法,如随机森林^[4]、最小二乘支持向量机^[5];深度学习方法,如长短期记忆神经网络(Long Short-Term Memory, LSTM)^[6]、卷积长短期记忆神经网络(Convolutional LSTM Network, ConvLSTM)^[11]、基于时空图像的CLR预测模型^[7]。供给和需求预测无法直接反映某区域供需情况,相比之下,供需缺口预测更加直观。供需缺口预测研究多基于人工智能方法,如Zhang等^[8]提出二次集成梯度提升决策树方法;Li等^[9]应用LSTM实现网约车供需缺口预测;Wang等^[10]建立基于残差网络的深度供需预测模型。网约车供需缺口预测是时空分布预测,受时空关联性影响,而现有研究均未涉及空间关联性,未考虑周边区域供需情况、交通状态等对预测区域供需缺口的影响。

为弥补以上缺陷,本文在FCL-Net模型^[1]基础上加以改进,综合考虑时空关联性、区域差异性和环境变化对网约车供需缺口的影响,创新提出基于时空数据挖掘的深度预测模型(Spatio-Temporal Deep Learning model, S-TDL)。并提出基于特征聚类—最大信息系数(Feature Clustering-Maximal Information Coefficient, FC-MIC)的两阶段特征选择方法,提取重要特征,减少过拟合。

1 网约车供需缺口影响变量

城市不同区域网约车供需缺口变化规律复杂,受区域特征、交通状态和天气条件等多种因素影响,因此,仅利用历史供需缺口作为模型输入变量难以准确预测供需缺口变化。本文充分考虑各种因素,将影响网约车供需缺口变化的特征变量划分为3类:时空变量、空间属性变量和环境变量。

1.1 时空变量

(1) 供需缺口。

将城市划分为 $A \times B$ 个互不重叠的正方形网

格,表示 $A \times B$ 个不同区域,将1天24 h划分为 I 个相等时段。定义 t 时段不同区域网约车供需缺口 G_t 为

$$G_t = \begin{pmatrix} g_{(1,1,t)}^{ap} & g_{(1,2,t)}^{ap} & \cdots & g_{(1,B,t)}^{ap} \\ g_{(2,1,t)}^{ap} & g_{(2,2,t)}^{ap} & \cdots & g_{(2,B,t)}^{ap} \\ \vdots & \vdots & & \vdots \\ g_{(A,1,t)}^{ap} & g_{(A,2,t)}^{ap} & \cdots & g_{(A,B,t)}^{ap} \end{pmatrix} \quad (1)$$

$$g_{(a,b,t)}^{ap} = d_{(a,b,t)} - s_{(a,b,t)} \quad (2)$$

式中: $g_{(a,b,t)}^{ap}$ 为区域 $D_{(a,b)}$ 在 t 时段的供需缺口, $g_{(a,b,t)}^{ap} \geq 0$; $d_{(a,b,t)}$ 为该区域网约车需求量; $s_{(a,b,t)}$ 为供应量(需求满足量)。

(2) 交通状态。

将道路交通状态划分 C 个级别($C > 1$),级别越大表示道路越拥堵。定义 t 时段不同区域第 i 级别交通状态下的路段数量为 R_t^i ($A \times B$ 矩阵), $i = 1, 2, \dots, C$,用于表征各区域交通状态,其中第 (a,b) 个元素为 $r_{(a,b,t)}^i$,表示区域 $D_{(a,b)}$ 在 t 时段第 i 级别交通状态下的路段数量。

1.2 空间属性变量

区域供需缺口大小与区域及其周边土地利用模式紧密相关。将区域内各类兴趣点(Point of Interest, POI)数量作为空间属性变量,表示区域特征。POI是指具有地理标识的空间特征物,涵盖城市各类设施位置与属性信息。将城市POI划分 S 类,定义不同区域第 i 类POI数量为 P_t^i ($A \times B$ 矩阵), $i = 1, 2, \dots, S$,其中第 (a,b) 个元素为 $p_{(a,b)}^i$,表示区域 $D_{(a,b)}$ 第 i 类POI数量。

1.3 环境变量

(1) 时间。

依据供需缺口变化规律,将1天24 h划分为3个时段^[1]: 7:30-9:30和17:30-19:30为高峰时段,23:00至次日6:00为入眠时段,其余为非高峰时段。定义表征预测时段的变量为

$$t^h = \begin{cases} 2, & \text{预测时段为高峰时段} \\ 1, & \text{预测时段为非高峰时段} \\ 0, & \text{预测时段为入眠时段} \end{cases} \quad (3)$$

(2) 节假日。

定义表示预测时段是否属于节假日的变量。

$$l = \begin{cases} 0, & \text{预测时段属于节假日} \\ 1, & \text{预测时段属于工作日} \end{cases} \quad (4)$$

(3) 天气条件.

将天气类型 w 、温度 e 和PM2.5指数 p^q 作为表征天气的变量.其中,天气类型可划分为5类:晴天、多云、小雨、中雨和大雨.

2 基于时空数据挖掘的深度学习预测模型

S-TDL可同时预测城市 $A \times B$ 区域 i 时段内的供需缺口($A \times B$ 矩阵).如图1所示,S-TDL由3个子模

型构成:基于ConvLSTM的时空变量模型、基于卷积神经网络(Convolutional Neural Network, CNN)的空间属性变量模型和基于全连接层(Fully Connected Layer, FC)的环境变量模型.子模型输出变量以不同权重融合,得到各区域供需缺口预测值.3个子模型分别对时空、空间属性和环境变量进行处理,可有效捕捉时空关联性、区域差异性和环境变化对供需缺口的影响.

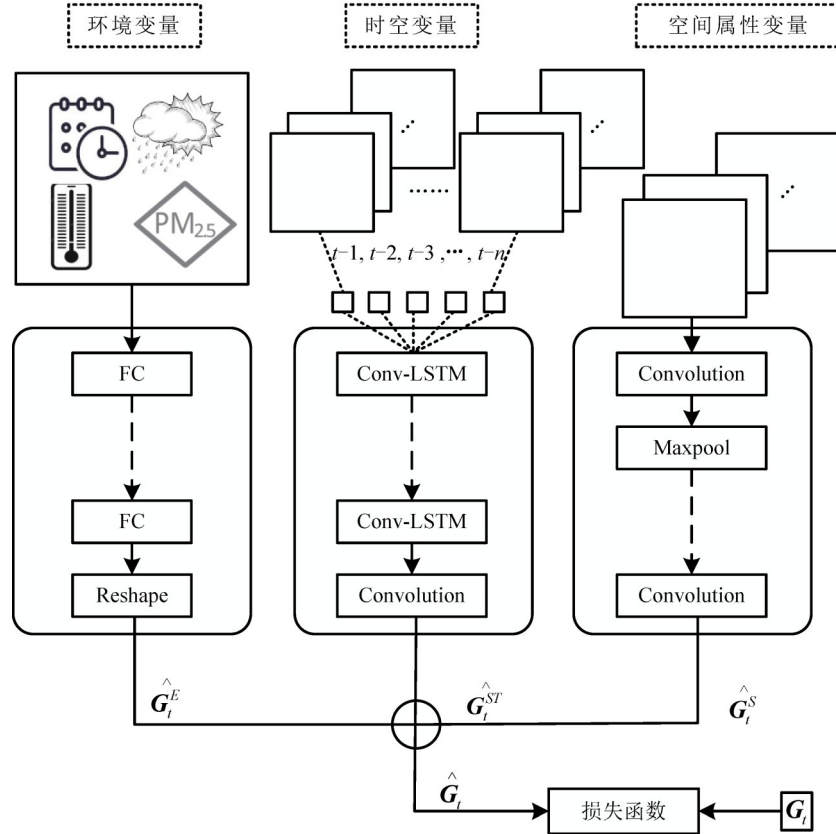


图1 S-TDL结构示意图

Fig. 1 Structure of the S-TDL

2.1 子模型1 时空变量模型

输入变量 X^{ST} 为各区域供需缺口和交通状态的历史数据,即

$$X^{ST} = (G_{t-1}, G_{t-2}, \dots, G_{t-n}, R_{t-1}^1, R_{t-2}^1, \dots, R_{t-n}^1, R_{t-1}^2, \dots, R_{t-n}^2, \dots, R_{t-1}^C, \dots, R_{t-n}^C) \quad (5)$$

式中: G_{t-n} 为 $t-n$ 时段不同区域供需缺口; R_{t-n}^i 为 $t-n$ 时段不同区域第 i 类交通状态下路段数量.模型可定义为 $f^{ST}(X^{ST}): R^{T \times A \times B \times 1} \rightarrow R^{A \times B \times 1}$ (R 为实数集), T 表示历史数据的时间序列长度, $T=n$; L 表示输入特征个数, $L=(C+1) \cdot n$.

该模型主要结构为ConvLSTM,如图2所示,

在LSTM基本结构中加入卷积操作,使其既可以像LSTM一样建立时序关系,又可以像CNN一样提取空间特性.ConvLSTM最早由Shi提出并证明它在获取时空关系上比LSTM有更好的效果^[10].故本文利用ConvLSTM处理时空变量,可有效捕捉时空关联性.通过 T 次循环,模型将时空变量 X 映射为 H ,每次循环计算公式为

$$I_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (6)$$

$$F_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (7)$$

$$C_t = F_t \circ C_{t-1} + I_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (8)$$

$$O_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \quad (9)$$

$$H_t = O_t \circ \tanh(C_t) \quad (10)$$

式中: X_t, F_t, I_t, O_t 和 H_t 为三维张量, 其中两个维度表示行和列的空间信息; X_t 表示第 t 时间步输入变量, $t = 1, 2, \dots, T$; F_t, I_t, O_t 分别表示遗忘门、更新门和输出门; C_t 是记忆向量, 可传递到第 $t+1$ 次循环中; F_t, I_t, O_t 和 C_t 共同决定 H_t . “ $\circ$ ” 表示 Hadamard 乘法; “ $*$ ” 表示卷积操作, W 为卷积核, 可提供权重共享; σ 和 $\tanh$ 表示非线性激活函数, 即

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (12)$$

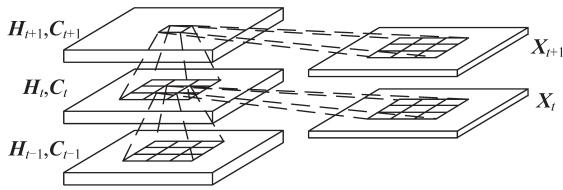


图2 ConvLSTM 内部结构

Fig. 2 Inner structure of ConLSTM

2.2 子模型2 空间属性变量模型

该模型可有效捕捉区域差异性对供需缺口的影响, 主要结构为 CNN, 输入变量 X^s 为空间信息数据, 即

$$X^s = (P_1, P_2, \dots, P_s) \quad (13)$$

式中: P_i 表示不同区域第 i 类 POI 数量. 模型可定义为 $f^s(X^s): R^{A \times B \times S} \rightarrow R^{A \times B \times 1}$.

2.3 子模型3 环境变量模型

该模型可有效捕捉环境变化对供需缺口的影响, 主要结构为全连接网络, 输入变量 x^E 包括时间和天气条件, 即

$$x^E = (t^h, l, w_{t-1}, e_{t-1}, p_{t-1}^q) \quad (14)$$

式中: t^h 表示预测时段; l 表示是否为节假日; w_{t-1}, e_{t-1} 和 p_{t-1}^q 分别表示 $t-1$ 时段的天气类型、温度和 PM2.5 指数. 该模型可定义为 $f^E(x^E): R \rightarrow R^{A \times B \times 1}$.

2.4 模型融合

将 3 个子模型输出结果融合, 得到 t 时段网约车供需缺口预测值, 计算公式为

$$\hat{G}_t = w^E \hat{G}_t^E + w^{ST} \hat{G}_t^{ST} + w^S \hat{G}_t^S \quad (15)$$

式中: $\hat{G}_t$ 为城市不同区域 t 时段的网约车供需缺口预测值; $\hat{G}_t^E, \hat{G}_t^{ST}$ 和 $\hat{G}_t^S$ 分别为 3 个子模型输出结果;

w^E, w^{ST} 和 w^S 为融合权重.

S-TDL 训练目标是最小化供需缺口预测值 $\hat{G}_t$ 与实际值 G_t 的均方误差, 即

$$\min_{w,b} \left\| \hat{G}_t - G_t \right\|_2^2 + \lambda \|W\|_2^2 \quad (16)$$

式中: $\lambda \|W\|_2^2$ 表示 L2 正则化项, 可减少过拟合; λ 为正则化系数; W 表示模型中所有权重系数.

3 特征选择

提出 FC-MIC 两阶段特征选择方法, 先利用特征聚类别除冗余空间信息, 再利用最大信息系数度量特征变量与供需缺口相关性, 提取重要特征.

(1) 第1阶段.

针对空间属性变量特征多、冗余性大、无时序性特点, 采用聚类分析对其进行特征提取. 假定空间属性变量 X^s 有 S 个特征, 特征集 $P = \{P_1, P_2, \dots, P_s\}$, $P \in R^{A \times B \times S}$, 以特征集 P 为研究对象, 依据平方欧式距离进行聚类. 聚类后, P 被划分为 K 个不同特征簇, 可表示为 $P = \{C_1, C_2, \dots, C_K\}$, $1 \leq K \leq S$.

对任意特征簇 $C_k = \{P_{k1}, P_{k2}, \dots, P_{kN}\}$, $k = 1, 2, \dots, K$, 计算特征簇内 N 个特征的平均值作为该特征簇提取特征, $P_k^c = \sum_{i=1}^N P_{ki} / N$. 则提取后新特征集为 $P^c = \{P_1^c, P_2^c, \dots, P_K^c\}$.

(2) 第2阶段.

最大信息系数 (Maximal Information Coefficient, MIC) 可对变量间线性、非线性和非函数依赖关系进行有效度量. 特征选择第 2 阶段利用 MIC 度量特征与供需缺口相关性. 第 1 阶段提取的空间属性特征 P^c 和其余特征组成集合 $F = \{f_1, f_2, \dots, f_g\}$, g 为特征数量. 对任意特征 $f_i \in F$, 计算该特征与供需缺口的 MIC, 计算公式如式 (17) 和式 (18) 所示. 最后, 依据 MIC 排序筛选出数值较大的 Q 个特征.

$$MI(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (17)$$

$$MIC[X; Y] = \max_{|X|Y| < B} \frac{MI[X; Y]}{\log(\min(|X|, |Y|))} \quad (18)$$

式中: $MI(X; Y)$ 为随机变量 X 和 Y 的互信息, 表示两

个变量共同拥有的信息含量;变量 X, Y 取值的有序对集合 $V = \{(x_j, y_j), j = 1, 2, 3, \dots, m\}$, m 为样本数量, 利用直方图估计 X 和 Y 边缘概率密度 $p(x), p(y)$, 和它们的联合概率密度 $p(x, y)$; $MIC[X; Y]$ 为变量 X, Y 最大信息系数. 将两随机变量画成散点图, 利用网格划分原理将变量 X 值域划分 $|X|$ 段, Y 值域划分 $|Y|$ 段, M 为网格划分 $|X| \times |Y|$ 限制条件, $M(m) = m^{0.6}$.

4 实例分析

4.1 数据来源

数据来源为滴滴出行Z市2016年1月1~21日网约车订单信息, 原始数据包含订单、区域、POI、拥堵信息和天气信息【<http://cn-static.udacity.com/mlnd/didi/didi.html>】. 数据中拥堵和天气信息获取间隔分别为10 min和5 min, 故将天气和订单信息以10 min为间隔汇聚, 则1天可划分为144个相等时段. 原始数据将Z市划分为66个互不重叠的正方

形区域, 为提高运算效率, 选取前 8×8 区域, 预测各区域 t 时段供需缺口. 处理后样本数量为3 024, 其中70%为训练集, 30%为测试集.

4.2 特征选择

历史输入数据时间窗为5; 交通状态划分4类: 畅通、缓行、拥挤和阻塞; POI划分25类, 特征集合为

$$F = \{G_{t-1}, G_{t-2}, \dots, G_{t-5}, R_{t-1}^1, R_{t-2}^1, \dots, R_{t-5}^1, R_{t-1}^2, R_{t-2}^2, \dots, R_{t-5}^2, R_{t-1}^3, R_{t-2}^3, \dots, R_{t-5}^3, R_{t-1}^4, R_{t-2}^4, \dots, R_{t-5}^4, P_1, P_2, \dots, P_{25}, t^h, l, w_{t-1}, e_{t-1}, p_{t-1}^q\}$$

第1阶段聚类并重构特征后, 空间属性特征由25减少至6个, 表示为 $P^c = \{P_1^c, P_2^c, \dots, P_6^c\}$; 第2阶段选择MIC较大的前20个特征, 如图3所示. FC-MIC将特征变量由55减少至20个, 新特征集合为

$$F_s = \{G_{t-1}, G_{t-2}, \dots, G_{t-5}, R_{t-1}^3, R_{t-2}^3, \dots, R_{t-5}^3, R_{t-1}^4, R_{t-2}^4, \dots, R_{t-5}^4, P_1^c, P_2^c, P_5^c, t^h, w_{t-1}\}$$

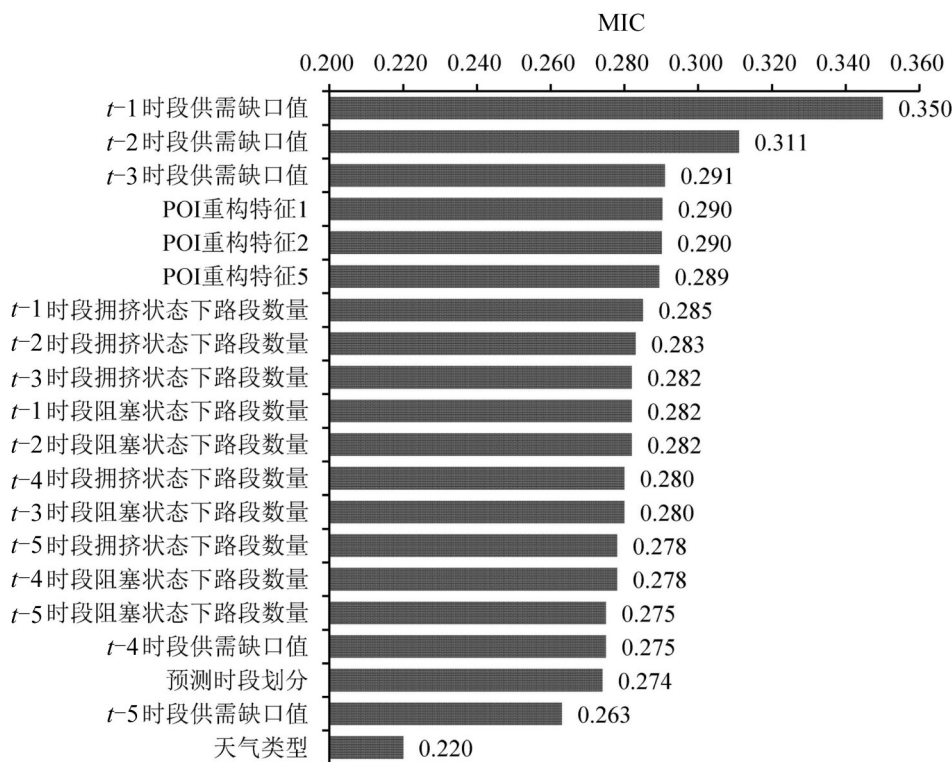


图3 20个选择特征的MIC

Fig. 3 MIC of 20 selected features

4.3 预测结果分析

利用Keras深度学习框架实验, 模型优化算法

为Adam, 训练代数数为1 000, 批大小设置为64. 模型主要结构如表1所示.

表1 S-TDL 网络结构
Table 1 Structure of the model

环境变量模型			
隐藏层		隐藏单元数	
FC1		512	
FC2		256	
FC3		128	
时空变量模型			
隐藏层	隐藏单元数	卷积核数量	卷积核大小
ConvLSTM1	128	40	(3,3)
ConvLSTM2	64	40	(3,3)
ConvLSTM3	32	40	(3,3)
Conv3D	—	1	(3,3,3)
空间属性变量模型			
隐藏层	卷积核数量		卷积核大小
Conv1	8		(5,5)
Conv2	8		(5,5)
Maxpool1	1		(2,2)
Conv3	16		(3,3)
Conv4	16		(3,3)
Maxpool2	1		(2,2)
Conv5	32		(3,3)
Conv6	32		(3,3)
Maxpool3	1		(2,2)
FC-FC	—		—

表1中,Conv3D为3D卷积层,Conv为2D卷积层,Maxpool为最大池化层,激活函数为

$$\text{ReLU}(x) = \max\{0, x\} \quad (19)$$

(1) 与其他模型预测结果比较.

选择ConvLSTM、LSTM、CNN和BP神经网络(BPNN)与S-TDL预测结果进行对比.模型评价指标选取平均绝对误差(Mean Absolute Error, MAE)和均方根误差(Root Mean Squared Error, RMSE).如表2所示,S-TDL平均预测误差最小,MAE和RMSE分别为7.09和8.92.与ConvLSTM相比,S-TDL运算时间仅增加5.88%,而预测误差(RMSE)降低14.64%;与S-TDL和ConvLSTM相比,其他模型预测误差较大,证明了时空关联性对供需缺口预测的重要影响.观察发现,CNN比LSTM预测效果好,MAE和RMSE分别比LSTM减小2.32和4.25,表明了空间关联性对供需缺口预测的重要影响.为深入分析S-TDL预测效果,选取3个区域,比较ConvLSTM与S-TDL模型1月20日05:10~1月21日05:00预测结果,如图4~图6所示.

表2 预测精度评价

Table 2 Evaluation of models' accuracy

模 型	性能评价指标		训练时间/min
	MAE	RMSE	
S-TDL	7.09	8.92	146.32
ConvLSTM	7.33	10.45	138.20
LSTM	13.58	20.03	149.61
CNN	11.26	15.78	90.95
BPNN	22.02	33.14	28.21

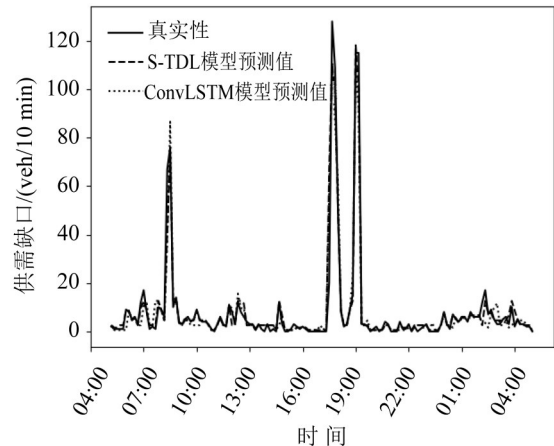


图4 区域(2, 4)的S-TDL和ConvLSTM预测结果

Fig. 4 Comparison of prediction results on the S-TDL and ConvLSTM in area (2, 4)

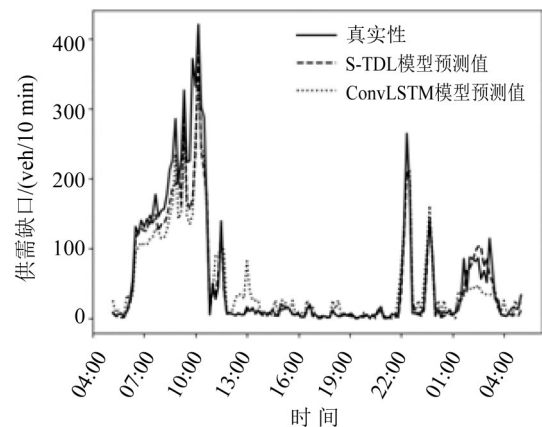


图5 区域(6, 8)的S-TDL和ConvLSTM预测结果

Fig. 5 Comparison of prediction results on the S-TDL and ConvLSTM in area (6, 8)

图4中S-TDL和ConvLSTM预测精度无明显差异,预测值与真实值曲线拟合程度高,可准确预测早、晚高峰变化.图5中,07:00-10:00时段ConvLSTM预测精度下降,无法准确预测高峰时期供需缺口变化,也无法捕捉凌晨时(01:00-4:00)的小高峰;而S-TDL预测值与真实值曲线仅在高峰期有

较小差异,其他时段拟合程度高.图6中区域(4,3)供需缺口变化复杂,波动性大,ConvLSTM预测结果有明显滞后现象,预测精度低,而S-TDL预测精度较高.与ConvLSTM相比,S-TDL考虑了区域特征对网约车供需缺口的影响,可有效捕捉区域差异性,因此保证了各区域供需缺口准确预测,避免出现某些区域预测精度高而个别区域预测精度低的现象.S-TDL模型更高的预测精度证明了土地利用模式对该区域网约车供需缺口预测的重要性.

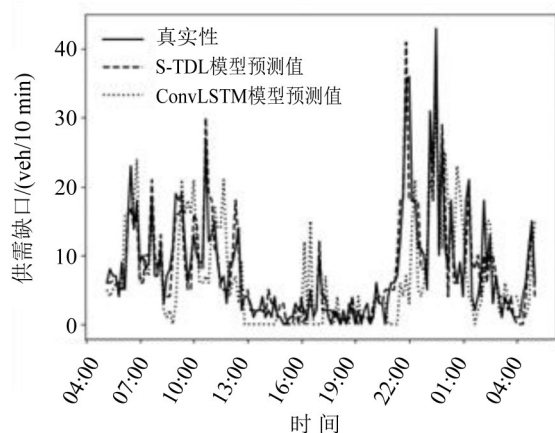


图6 区域(4,3)的S-TDL和ConvLSTM预测结果

Fig. 6 Comparison of prediction results on the S-TDL and ConvLSTM in area (4, 3)

(2) 不同天气预测结果比较.

为探究不同天气下S-TDL预测结果,将测试样本分别按照天气类型和空气质量等级进行分类,计算各类别预测结果的RMSE.其中,空气质量依据PM2.5指数划分为6个等级:优、良、轻度污染、中度污染、重度污染和严重污染^[12](实际样本数据无严重污染等级).

观察图7,与晴天、多云天气相比,ConvLSTM、CNN等模型雨天预测误差显著增大,说明雨天供需缺口变化规律更加复杂,难以准确预测.而S-TDL仅在大雨天气时预测误差增大,其他天气下误差无明显波动,且不同天气下预测精度均优于其他模型.观察图8,空气质量为优、良、轻度污染和轻度污染时,各模型预测误差保持稳定,而在重度污染天气下,各模型预测误差有不同程

度增长.但无论空气质量等级如何,S-TDL预测精度始终优于其他模型.综上所述,S-TDL在不同天气类型和空气质量等级下,预测效果稳定且精度较高,因此,可适用于不同天气下的供需缺口预测.

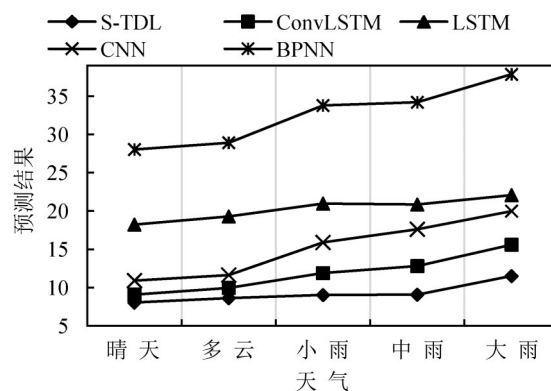


图7 不同天气类型预测结果

Fig. 7 Forecast results for different weather types

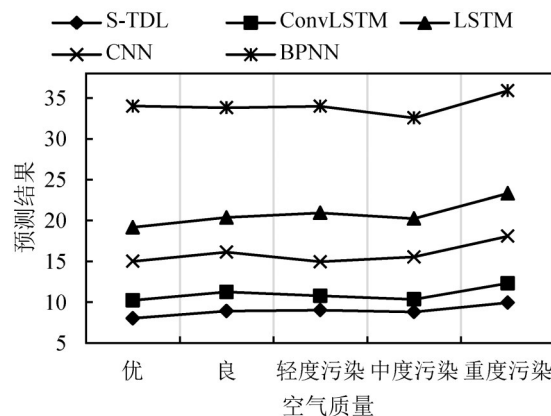


图8 不同空气质量等级预测结果

Fig. 8 Comparison of different air quality level prediction results

(3) 特征选择前后S-TDL预测结果比较.

为探究FC-MIC对模型预测结果影响,利用全部特征训练S-TDL并预测,与特征选择后的S-TDL进行比较.观察表3,输入全部特征的S-TDL测试误差比训练误差增加24.19%,说明存在明显过拟合现象,即在训练数据上拟合效果很好,测试数据上却不能较好拟合的现象.特征选择后的S-TDL测试误差比训练误差增加9.72%,与特征选择前相比降低14.47%,证明FC-MIC方法可有效减少过拟合,提高预测精度,同时缩短训练时间.

表3 特征选择前后S-TDL训练时间及误差

Table 3 Comparison of training time and error before and after feature selection

模 型	是否特征选择	特征数量	训练时间/min	RMSE		误差增加 百分比/%
				训 练	测 试	
S-TDL	否	55	204.67	7.98	9.91	24.19
	是	20	146.32	8.13	8.92	9.72

5 结 论

利用S-TDL实现城市各区域网约车供需缺口短时预测,为网约车调度策略提供支持.S-TDL模型由时空变量模型、空间属性变量模型和环境变量模型融合而成,可有效捕捉时空关联性、区域差异性和环境变化对供需缺口的影响.实验证明S-TDL模型预测精度显著优于BPNN、LSTM和CNN.提出FC-MIC两阶段特征选择方法,用于筛选冗余性小、与预测变量相关性大的有效特征.实验结果证明该特征选择方法可减少过拟合,提高运算效率.

由于数据获取限制,本文网约车供应量实为需求满足量,S-TDL模型仅能预测供给小于需求时的需求缺口,无法预测需求小于供给时的剩余供给.此外,拟在后续研究中深入分析各因素对供需缺口的影响程度,进一步完善预测模型.

参考文献:

- [1] KE J, ZHENG H, YANG H, et al. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach[J]. Transportation Research Part C: Emerging Technologies, 2017(85): 591-608.
- [2] JAMIL M S, AKBAR S. Taxi passenger hotspot prediction using automatic ARIMA model[C]//Science in Information Technology (ICSITech), 2017 3rd International Conference on. IEEE, 2017.
- [3] MOREIRA-MATIAS L, GAMA J, FERREIRA M, et al. Predicting taxi-passenger demand using streaming data[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1393-1402.
- [4] LIU J, CUI E, HU H, et al. Short-term forecasting of emerging on-demand ride services[C]//Transportation Information and Safety (ICTIS), 2017 4th International Conference on. IEEE, 2017.
- [5] LI Y, LU J, ZHANG L, et al. Taxi booking mobile app order demand prediction based on short-term traffic forecasting[J]. Transportation Research Record: Journal of the Transportation Research Board, 2017 (2634): 57-68.
- [6] XU J, RAHMATIZADEH R, BÖLÖNI L, et al. A sequence learning model with recurrent neural networks for taxi demand prediction[C]//Local Computer Networks (LCN), 2017 IEEE 42nd Conference on. IEEE, 2017.
- [7] 段宗涛, 张凯, 杨云, 等. 基于深度CNN-LSTM-ResNet组合模型的出租车需求预测[J]. 交通运输系统工程与信息, 2018, 18(4): 215-223. [DUAN Z T, ZHANG K, YANG Y, et al. Taxi demand prediction based on CNN-LSTM-ResNet hybrid depth learning model[J]. Journal of Transportation Systems Engineering and Information Technology, 2018, 18(4): 215-223.]
- [8] ZHANG X, WANG X, CHEN W, et al. A Taxi gap prediction method via double ensemble gradient boosting decision tree[C]// IEEE, International Conference on Big Data Security on Cloud. IEEE, 2017.
- [9] LI J, WANG Z. Online car-hailing dispatch: Deep supply-demand gap forecast on spark[C]// IEEE, International Conference on Big Data Analysis. IEEE, 2017.
- [10] WANG D, CAO W, LI J, et al. DeepSD: supply-demand prediction for online car-hailing services using deep neural networks[C]//2017 IEEE 33rd International Conference on Data Engineering (ICDE). IEEE, 2017.
- [11] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances in Neural Information Processing Systems, 2015.
- [12] 国家质量监督检验检疫总局. GB3095-2012 环境空气质量标准[S]. 北京: 中国环境科学出版社, 2012. [General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. GB3095-2012 Ambient Air Quality Standard [S]. Beijing: China Environmental Science Press, 2012.]