

基于 ConvLSTM 的南京地区共享单车需求预测研究

王 俊¹, 于爱荣²

(1.南京信息职业技术学院人工智能学院, 江苏 南京 210044;
2.中国人民解放军陆军工程大学指挥控制工程学院, 江苏 南京 211117)
✉ intraweb@163.com; yu_alice@163.com



摘 要:针对传统单车需求预测研究在处理不同地区无桩共享单车需求预测中存在的特征提取不精细和地区需求预测匹配度不精准等问题,在对南京地区两个自然年度、三类典型地区共享单车的使用需求数量、天气状态、季节周期等多类数据进行采集和特征工程的基础上,提出了一种兼顾时空序列的基于 ConvLSTM(Convolutional Long Short Term Memory)深度学习预测模型,该模型通过卷积操作后能够提取数据中隐含的更多空间信息,将其应用于测试集中并与经典时序 LSTM(长短期记忆网络)和 CNN(卷积神经网络)进行对比发现, RMSE 分别提升 0.05 和 0.04,最大误差分别提升约 0.86 和 0.3。

关键词:共享单车;需求预测;深度学习;ConvLSTM;交通

中图分类号: TP183 **文献标志码:** A

Research on the Prediction of Shared Bicycle Demand in Nanjing Based on ConvLSTM

WANG Jun¹, YU Airong²

(1.School of Artificial Intelligence, Nanjing Vocational College of Information Technology, Nanjing 210044, China;
2.College of Command Control Engineer, Army Engineering University, Nanjing 211117, China)
✉ intraweb@163.com; yu_alice@163.com

Abstract: In traditional shared bicycle demand prediction research, there are issues of imprecise feature extraction and inaccurate matching of regional demand prediction for predicting free floating shared bicycles in different regions. Aiming at these issues, this paper proposes a deep learning prediction model that takes into account the spatiotemporal sequences based on ConvLSTM (Convolutional Long Short Term Memory). The model is proposed on the basis of the collection and feature engineering of multiple types of data, such as the usage demand quantity, weather conditions, and seasonal cycles of shared bicycles in two natural years and three typical regions in Nanjing. This model can extract more spatial information hidden in the data after convolutional operations. Compared with the classical time-series LSTM network and CNN (Convolutional Neural Network), the RMSE of the test set improves by 0.05 and 0.04, respectively, and the maximum error improves by 0.86 and 0.3, respectively.

Key words: shared bicycles; demand prediction; deep learning; ConvLSTM; traffic

0 引言(Introduction)

随着共享经济的发展,共享单车作为一种新型的出行方式,它的便捷性、环保性和经济性受到越来越多人的青睐。在

国内,共享单车的运营模式主要有有桩共享单车(Station-Based Bike Sharing, SBBS)和无桩共享单车(Free-Floating Bike Sharing, FFBS)^[1-2]。当前,很多大型城市的交通节点和人口

相对集中,如何根据不同城市或城市不同区域的差异化需求精准预测共享单车需求量,合理规划布局共享单车的停放点和区域,是社会交通发展过程中面临的棘手问题。国内学者对共享单车需求预测和规划的研究主要集中于用车频率与周期特点分析、停放区域规划、城市活力分析和交通枢纽影响力分析等不同角度,但是较少针对城市特点和区域特点开展共享单车需求测算的设计。常见的长短期记忆网络(LSTM)、自动回归移动平均(ARMA)等需求预测模型在分析城市人口布局和区域特征分布时存在提取特征能力不足的问题,导致其预测结果与实际相比有较大偏差。因此,采用新的模型或技术手段挖掘用户使用共享单车的年度、季节性和周期性等方面的重要规律和特点,以此优化单车投放数量计划,并提升用户对共享单车使用的满意度,这成为当前的迫切需求^[3-5]。

1 时间序列预测与深度学习技术(Time series prediction and deep learning techniques)

1.1 时间序列预测

时间序列是按照时间先后顺序对随机变量 $X_1, X_2, \dots, X_T$ 进行排列的一种序列化表现,记为 $\{X_t, t \in T\}$,并用 $x_1, x_2, \dots, x_n$ 或者 $\{x_t, t \in 1, 2, \dots, N\}$ 表示该随机序列的 N 个有序观测值。其中,对时间序列应用最广泛的就是预测,其核心思想是利用过往的历史数据预知未来的走势。与回归或分类预测相比,时序预测的核心原理大相径庭,虽然它们都需要使用数据进行预测,但是回归或分类预测通常是使用其他数据点预测目标数据点,而时间序列预测则是利用历史数据预测未来数据点。这种预测方式通常基于时间序列的趋势性、季节性、周期性等因素,并考虑其他因素,如天气、假期等^[6]。

在时间序列分析中,通常可以将时间序列数据分解为趋势性、季节周期性和随机误差三大部分。趋势性主要描述了时序数据在长周期中隐含的有规律性的单向升高或下降的变化特点,该部分通常可以通过线性或非线性回归模型拟合。季节周期性描述了时序数据在一定时间周期内出现的可复现的规律性波动,该部分可采用季节性分解方法分离出季节性成分。随机误差也称为残差,是在上述两个特征之外出现的随机性的不规则波动,随机波动所导致的时间序列变化,可以通过残差分析检查模型的拟合程度。时间序列通常包含趋势性、季节周期性和随机误差的特征,可以采用加法模型或乘法模型进行表达。加法模型将时间序列分解为趋势性、季节性和残差的加和,乘法模型将时间序列分解为趋势性、季节性和残差的乘积。加法模型如公式(1)所示,乘法模型如公式(2)所示。

$$X_t = m_t + s_t + e_t \quad (1)$$

$$X_t = m_t * s_t * e_t \quad (2)$$

其中:趋势性特征用 m_t 表征,季节周期性特征用 s_t 表征,随机误差用 e_t 表征。在时间序列分析中,需要选择适当的模型描述数据中的趋势性和季节性变化。如果趋势性特征与季节周期性特征之间的关系相对独立,加法模型可以很好地表达它们相互之间的影响。如果趋势性特征的变化对季节周期性特征的效应有明显的影响,那么乘法模型将是更好的选择。因此,在选择模型时需要考虑数据中不同因素之间的相互关系,以便

能够更准确地描述数据的变化趋势。

针对时间序列数据的预测问题,目前主要有三种主流方法可供选择。经典的 ARMA 和 ARIMA(自回归综合移动平均)模型是以往进行时序数据建模与处理的常用模型,该类模型能够有效地处理特征平稳的时序数据,并且具备对单一变量数据序列的回归能力,但其对复杂特征数据的处理能力较弱,效果不佳;经典的机器学习模型往往将时序数据处理转化为无监督学习任务,在进行有效的数据特征工程、预处理和模型训练后进行时序数据的未来预测,该方法能够处理对复杂特征数据的分析与建模,并且能够较好地解决时序数据中多变量特征的协同性回归和非线性等问题,该类方法的典型模型有 Lightgbm 和 Xgboost;深度学习模型是近年来比较流行的时序数据处理方法,该方法在神经网络的基础上加入了大规模的隐藏层和感知单元,能有效地提取时序数据中的各类特征,该类模型主要包括 CNN 模型、GRU 模型和 LSTM 模型。具体来说,选择合适的时间序列预测方法应该考虑具体问题的复杂程度和数据量大小^[7]。

1.2 深度学习技术

深度学习技术在时间序列预测中的应用始于全连接神经网络(FCNN)。然而,FCNN 存在过拟合和泛化能力不足等问题。随着卷积神经网络(CNN)在图像领域取得的新突破,CNN 在特征提取中表现出的优越性使其成为深度学习技术中的热门模型,CNN 通过不同的卷积核和卷积操作可有效地对高维度的数据进行特征提取。但是,CNN 只能处理短期预测问题,对于处理长期预测问题表现不佳。在 CNN 之后出现的循环神经网络(RNN)可以通过内部隐藏状态很好地保留历史信息,其采用方向传播的参数训练模式,能高效推导网络中的权重参数,能自动高效地学习序列数据的特征。上述模型虽然能力突出,但是在训练过程中易出现梯度消失或爆炸等问题,这些问题限制了其在长序列预测方面的应用。为了解决这些问题,长短期记忆网络(LSTM)被提出,它通过门控机制控制历史信息的保留和遗忘,从而更好地处理长期依赖性问题。最近,变换器和深度强化学习(DRL)等技术也被应用于时间序列预测中^[8]。

深度学习技术在时间序列预测中的应用非常广泛,包括股票价格预测、交通流预测、气象预测、能源需求预测等。例如,在股票价格预测方面,深度学习技术可以通过学习股票价格历史数据预测未来的股票价格走势。在交通流预测方面,深度学习技术可以通过分析交通历史数据预测未来的交通状况,为交通管理和规划提供支持。在气象预测方面,深度学习技术可以通过学习天气历史数据预测未来的天气情况,为气象预警和防灾减灾提供支持。在能源需求预测方面,深度学习技术可以通过学习用电量历史数据预测未来的能源需求量,为能源规划和管理提供支持。

1.3 ConvLSTM

LSTM 模型是由 Hochreiter 等人对 RNN 进行改进而提出的一种新型模型,其设计初衷是为了解决长期依赖性问题。21 世纪,LSTM 被一代代学者逐渐完善,形成了较为系统的 LSTM 框架,在自然语言处理(NLP)、自动驾驶汽车等领域均

有应用。在 LSTM 模型中,ConvLSTM (Convolutional Long Short Term Memory)是将 CNN 和 LSTM 进行结合的一种特殊形态,它继承了 CNN 的卷积特点,能够有效地提取高维数据信息的特征,同时又集合了 LSTM 的记忆能力,能够应对较为复杂的时序数据。该模型对于语音、视频等多特征的时序数据有很好的处理效果^[9]。同时,ConvLSTM 还被广泛应用于如天气、股票、能量场等多尺度的数据预测。ConvLSTM 的网络结构图如图 1 所示。

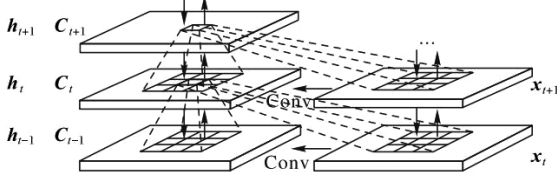


图 1 ConvLSTM 的网络结构图

Fig. 1 Network structure diagram of ConvLSTM

全连接的 LSTM,其输入和输出状态都是一维的张量,它主要的优势在于梯度会被保存在单元中,从而避免梯度消失的问题。LSTM 单元结构如图 2 所示。

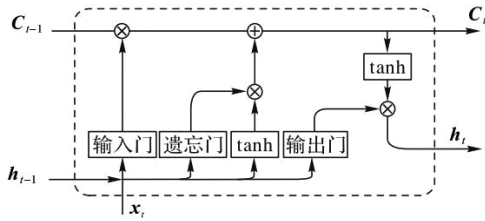


图 2 LSTM 单元结构

Fig. 2 LSTM unit structure

LSTM 的具体计算如公式(3)至公式(7)所示:

$$i_t = \sigma(W_{xi}x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \quad (4)$$

$$C_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc} * h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho} * h_{t-1} + W_{co} \circ c_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \circ \tanh(C_t) \quad (7)$$

其中:输入门、遗忘门和输出门分别用 i 、 f 和 o 表示, W_i 、 W_f 、 W_o 是其对应的权重, b_i 、 b_f 、 b_o 为遗忘门的偏置项, sigmoid 函数为 σ , $\circ$ 为点积操作, C_t 、 C_{t-1} 分别为 t 和 $t-1$ 时刻的长期记忆, o_t 为神经元的中间记忆^[10]。虽然全连接的 LSTM 已经可以用来解决时间关联性问题,但是它无法捕捉空间特征。

ConvLSTM 由两种组件组成,即卷积层和 LSTM 层。其中,卷积层的作用在于以卷积方式对多维数据提取有效特征, LSTM 层的作用在于通过门控机制将需要记忆的信息进行迭代保存,将需要遗忘的数据丢弃。ConvLSTM 的各个时刻状态计算方法如公式(8)至公式(12)所示:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (8)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (9)$$

$$C_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (10)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o) \quad (11)$$

$$H_t = o_t \circ \tanh(C_t) \quad (12)$$

其中: $*$ 代表卷积, $\circ$ 代表哈达玛积, X_t 代表输入, C_t 代表单元输出, H_t 代表隐状态, i_t 代表输入门, f_t 代表遗忘门, o_t 代表

输出门。比较 FC-LSTM 和 ConvLSTM 的重要等式可以看出, ConvLSTM 与 FC-LSTM 主体大致相同,仅提取特征的方式不同, ConvLSTM 是将所有普通乘法改为卷积操作,卷积核越大,就越能够提取到更快的动作,而卷积核越小,就越能够提取到更慢的动作。

2 数据集与特征分析(Dataset and feature analysis)

收集南京地区两个自然年度三类典型地区共享单车的使用需求数量信息作为数据集,数据集有 24 个数据采集点,每小时采集 1 次。数据集中记录了与共享单车需求相关的区域、气温、天气等影响因素,数据集特征描述如表 1 所示。

表 1 数据集特征描述

Tab.1 Dataset feature description

字段名称	描述	类型	样例
采集时间	以 1 h 为间隔	时间型	2022-12-22 12:00
需求量	需求的数量	数值型	31
天气	天气状况	数值型	0 表示晴,1 表示多云, 2 表示阴,3 表示雨, 4 表示雪
温度	温度数值	数值型	26.5
风速	风速大小	数值型	0 表示无风,1 表示小风, 2 表示中风,3 表示大风
位置类别	三类地区采集位置	数值型	0

为了有效地训练神经网络(如 LSTM),需要注意样本数据中数值较大的输入可能会对学习和收敛速度造成负面影响,甚至影响神经网络的学习精度。因此在进行训练之前,需要对样本数据进行预处理以达到更好的训练效果^[11]。其中,标准化作为一种常用的数据预处理方式,可将数据值拟合为均值为 0、方差为 1 的正态分布函数。需要注意的是,进行标准化操作之前需要对转换数据进行分布情况的分析,如数据不符合高斯分布,其进行标准化后的可靠性将有较大程度地降低,标准化的转换函数如公式(13)所示:

$$X^* = \frac{x - u}{\sigma} \quad (13)$$

标准化需计算样本数据的均值 u 和标准差 σ ,通常可以通过对训练数据进行估计得到。与归一化方式相比,标准化方法的优点在于其鲁棒性更强、训练效果更为稳定。对标准化后的特征通过相关性分析,使用斯皮尔曼秩相关系数衡量变量之间的线性相关,去除冗余的特征,该系数计算如公式(14)所示:

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n} \quad (14)$$

其中: ρ 为斯皮尔曼秩相关系数; d_i 为对应变量的秩之差,即 X_i 和 Y_i 两个变量分别排序后对应成对的变量位置(等级)差,即 $X_i - Y_i$; n 为观测对象的数量。斯皮尔曼秩相关系数在 $-1 \sim 1$ 变化。当其绝对值接近 1 时,表明两个变量的相关性更强,当系数值为 0 时,表明两个变量完全不相干(一个变量的变化不受另一个变量变化的影响)^[12]。一般而言,相关系数绝对值大于 0.8 的则判定两个变量为强相关,小于 0.2 的则判定

两个变量为弱相关,在实际特征分析过程中绝对值大于 0.6 或小于 0.4,即可判定为两个变量有一定的强弱相关性。各变量的特征相关性图如图 3 所示。

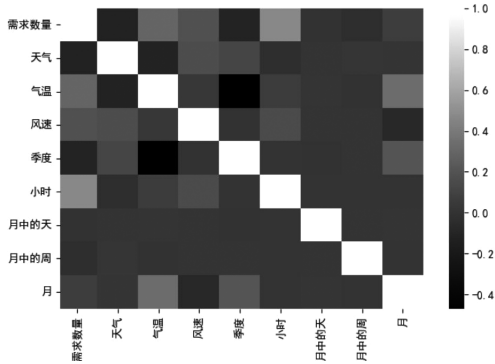


图 3 各变量的特征相关性图

Fig. 3 Feature correlation diagram of each variable

研究人员通过对三个区域内的共享单车的全部需求数据进行相关性分析发现,需求与气温和风速有一定的关系,并且与时间和月份存在关联性。通过对月度和小时的单车需求数量分析发现,大多数月份的需求呈现双峰分布,同时还发现工作时段的需求呈现增加态势,而下午非工作时段的需求相对稳定,月需求数量分布如图 4 所示,小时需求数量分布如图 5 所示。

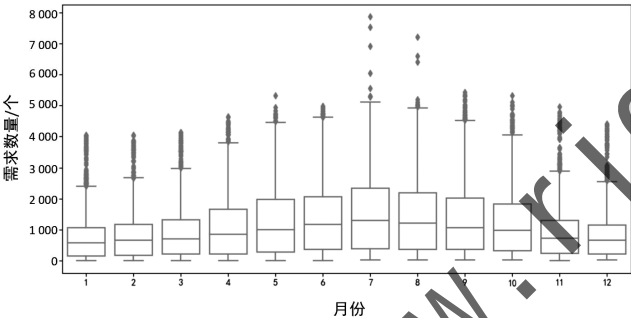


图 4 月需求数量分布

Fig. 4 Monthly demand quantity distribution

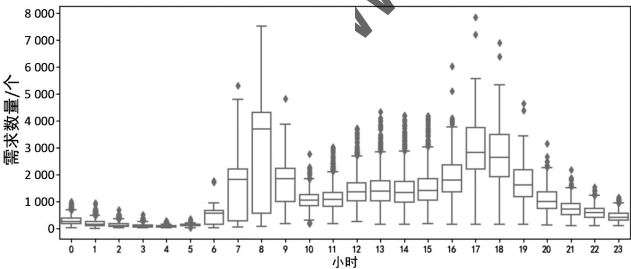


图 5 小时需求数量分布

Fig. 5 Hourly demand quantity distribution

从图 4 可以看出,南京地区的 7~9 月是共享单车需求量最高的月份,这与气温升高和暑假来临有密切关系。对图 5 中的每日数据进行对比分析发现,8 时、17 时和 18 时存在用车高峰,这与上下班时间点重合,在季节影响方面可以看出春、夏、秋季的需求旺盛,冬季的需求量较为低迷。

3 模型训练和评估(Model training and evaluation)

对于多变量时间序列预测问题,ConvLSTM 模型的输入数据形态为具有步长的多维数据,其表达方式为[samples, timesteps, rows, cols, channels],每一个时间步都包含有多个行列的数据。本文将 ConvLSTM 模型的步长设定为 14 d 的周期,这样便于拆解为两个子序列进行卷积,实现特征提取。每个子序列中包含 7 d 的特征数据,整个输入数据形状为[n, 2, 1, 7, 1]。为了验证 ConvLSTM 的训练和评估效果,研究人员将 ConvLSTM 模型与 LSTM、GRU 和 CNN 模型进行了横向对比。ConvLSTM 模型的参数信息如表 2 所示。

表 2 ConvLSTM 模型的参数信息

Tab.2 ConvLSTM model parameters

网络层	输出维度	参数量/个
time_distributed	$N, N, 7, 64$	640
time_distributed_1	$N, N, 3, 64$	0
time_distributed_2	$N, N, 192$	0
LSTM	$N, 50$	48 600
Dense	$N, 1$	51

ConvLSTM 总计训练参数为 49 291 个,50 次的迭代效果如图 6 所示,CNN、LSTM、GRU 的迭代效果如图 7 至图 9 所示,对比图 7 至图 9 可以看出,ConvLSTM 在 10 次迭代后的残差就已经降至较为稳定的状态,并且在验证集上的应用效果也较好。

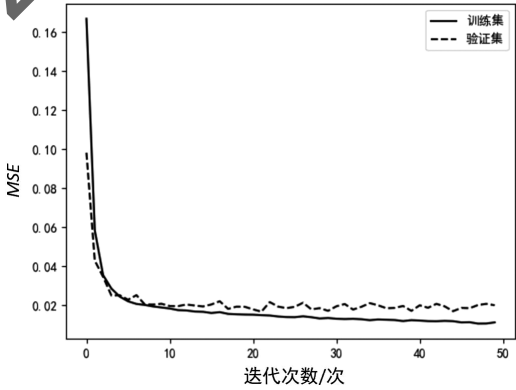


图 6 ConvLSTM 模型训练残差

Fig. 6 ConvLSTM model training residual

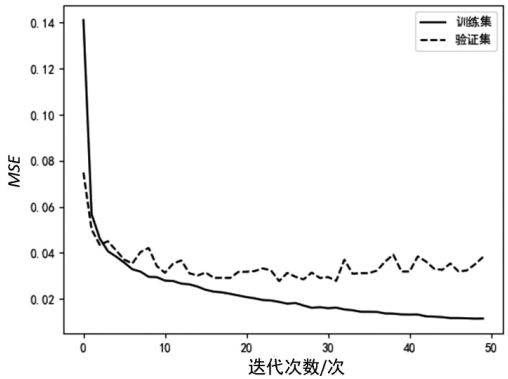


图 7 CNN 模型训练残差

Fig. 7 CNN model training residual

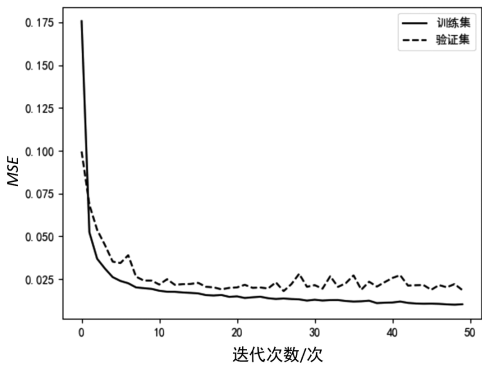


图 8 LSTM 模型训练残差

Fig. 8 LSTM model training residual

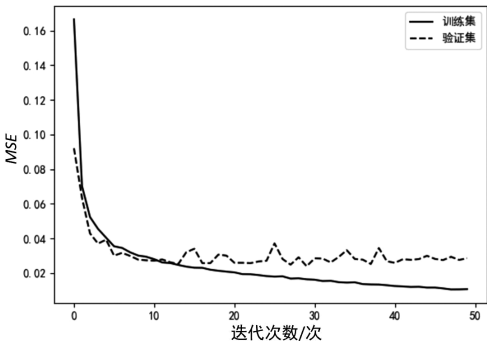


图 9 GRU 模型训练残差

Fig. 9 GRU model training residual

ConvLSTM 与 LSTM、GRU、CNN 模型效果对比如表 3 所示。

表 3 模型效果对比

Tab.3 Comparison of model effects

指标项	CNN	LSTM	GRU	ConvLSTM
训练集 RMSE	0.106 6	0.112 7	0.117 6	0.106 1
测试集 RMSE	0.178 5	0.169 0	0.154 8	0.124 6
MSE	0.031 9	0.028 6	0.024 0	0.015 5
解释性变量分值	0.922 6	0.980 3	0.942 2	0.963 0
最大误差	1.794 6	1.241 7	1.049 0	0.938 9

从表 3 可以看出,ConvLSTM 模型的解释性变量分值最高,明显优于单一使用 CNN 或 LSTM 的模型,其在训练集和测试集上的 MSE、RMSE 误差有一定的提升,利用 ConvLSTM 进行预测的结果如图 10 所示。

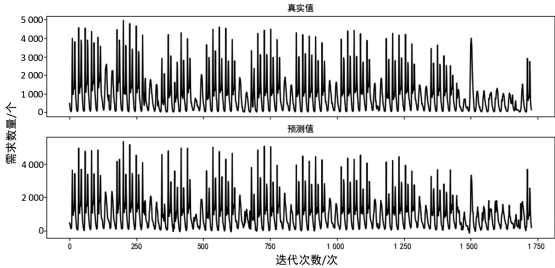


图 10 ConvLSTM 模型预测结果

Fig. 10 ConvLSTM model prediction results

从图 10 可以看出,预测结果的趋势吻合度较好,对于波动性变化的敏感度适应能力较强。

4 结论(Conclusion)

本文对南京地区两个自然年度三类型地区共享单车的使用需求数量、天气状态、季节周期等多类信息进行分析 and 特征提取,并在此基础上提出了一种同时考虑了时空序列的基于 ConvLSTM 的深度学习预测模型,该模型通过卷积操作提取数据中隐含的空间信息,设定序列长度周期为 14 d 时,模型拥有最优的预测效果。通过与经典的时序 LSTM 网络、卷积 CNN 等模型进行相比,本文设计的模型其 RMSE、MSE 和解释性变量分值等指标显示出更好的效果。在模型设计中,时间序列长度的设置至关重要,该参数设置过大,会导致短期需求预测准确度降低;该参数设置过小,会导致样本与模型的网络结构不匹配而出现无法有效拟合的问题,因此后续研究将围绕共享单车的数据特征工程展开,增加城市、地区、地形等方面特征的多模态数据,提升模型的预测精准度。

参考文献(References)

[1] 李浩,曹元密,涂辉招. 共享单车骑行需求预测研究[J]. 综合运输,2022,44(5):92-101.

[2] 蒋璇,徐铖铖,张靖,等. 校园无桩共享单车时空动态需求预测[J]. 长安大学学报(自然科学版),2022,42(5):105-115.

[3] 孙胜男,方润涛,叶波. 基于出行需求的共享单车总量预测分析[J]. 西部交通科技,2021(9):169-171.

[4] 刘恒孜,贺玉龙,宋太龙,等. 共享单车需求预测及调度优化[J]. 科学技术与工程,2021,21(35):15247-15254.

[5] 孙启鹏,曾开邦,张锴琦,等. 北京市共享单车出行的时空规律与需求预测研究[J]. 交通运输系统工程与信息,2022,22(1):332-338.

[6] 丁一丹,张福鼎. 城市公共自行车智能需求预测及调度管理系统[J]. 电子测试,2022,36(7):24-26.

[7] 刘本兴. 基于 LightGBM 的共享单车短时需求量预测[J]. 现代信息科技,2022,6(20):84-89.

[8] 张徐,聂文惠. 基于改进随机森林算法的共享单车需求量预测[J]. 计算机与数字工程,2021,49(9):1860-1865.

[9] 李福,徐良杰,朱然博,等. 基于 XGBoost 算法的共享单车借车需求量预测[J]. 武汉理工大学学报(交通科学与工程版),2021,45(5):880-884.

[10] 马云鹤,王玉玫,赵宇帆. 基于时空融合图的共享单车需求预测系统[J]. 计算机测量与控制,2023,31(2):97-103.

[11] ZHENG H F, LIN F, FENG X X, et al. A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction[J]. IEEE Transactions on intelligent transportation systems,2021,22(11):6910-6920.

[12] MA M, MAO Z. Deep-convolution-based LSTM network for remaining useful life prediction[J]. IEEE transactions on industrial informatics,2021,17(3):1658-1667.

作者简介:

王 俊(1979-),男,博士,高级工程师。研究领域:人工智能,机器学习。

于爱荣(1979-),女,硕士,讲师。研究领域:大数据,复杂系统架构。