用 LSTM 对市级周交通事故量预测方法研究

孙振华[1,2],王转转[2],肖 鑫[2]
(1. 绍兴市交通建设有限公司,浙江 绍兴 321000;
2. 长安大学 信息工程学院,陕西 西安 710064)

摘 要:市级交通事故量时间序列的波动是影响对其准确预测的关键因素。提出的预测方法针对市级日交通事故量时间序列,采用长短期记忆神经网络(Long Short-Term Memory,LSTM)捕捉序列当前观测值与前序观测值的时序依赖关系,通过寻找最优窗口长度的 LSTM 市级日粒度交通事故量预测模型使拟合数据对训练集误差最小,对验证集的预测结果在转为周粒度时取得了较为准确的预测效果。提出的预测方法解决了影响市级周交通事故量准确预测的问题,该方法发现基于交通事故量训练的用于捕获观测值时序依赖关系的 LSTM 模型对数据基本趋势准确性的表达远好于对数据波动性的表达。为此,提出最优窗口算法来确定 LSTM 模型最优窗口长度,以确保对训练集基本趋势表达的准确性,再根据所发现的细粒度下的预测结果对交通事故量基本趋势的准确描述可转化为粗粒度下对波动性准确描述的事实,将日粒度预测结果转为周粒度后就取得了较为准确的预测效果。

关键词:交通事故;神经网络;长短期记忆;时间序列;最优窗口
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2023)02-0195-08
doi:10.3969/j.issn.1673-629X.2023.02.029

Approach of Predicting Number of Citywide Traffic Accidents Using Long Short-term Memory Neural Network

SUN Zhen-hua[1,2],WANG Zhuan-zhuan[2],XIAO Xin[2]
(1. Shaoxing Transportation Construction Co.,Ltd.,Shaoxing 321000,China;
2. School of Information Engineering,Chang'an University,Xi'an 710064,China)

Abstract:The fluctuation of the number of citywide traffic accidents is a key to affect the accuracy of prediction. The proposed approach,in the time-series observations at the day level coming from the statistics of urban traffic accidents,adopted the Long Short-Term Memory (LSTM) to capture the dependent relationship between the current observations and the preceding observations. The LSTM model for predicting the number of citywide traffic accidents per day corresponding to the optimum window length of the input sequences at the day level was developed to achieve the minimum error between the fitting data and the observations of the training set. As a result,the fluctuation of the validation set at week level can be depicted accurately when the predicted data are aggregated at week level. The key point to compromise the accuracy of prediction has been addressed in the proposed approach,in which,the LSTM model for capturing the dependent relationships between observations,trained by the number of citywide traffic accidents,is found out to perform the description of the trend far more accurate than the fluctuation of data. Then an algorithm was proposed to determine the optimum window length for the LSTM model to ensure the accuracy of depicting the trend of the training set. According to the fact that the predicted data perform the accurate description of the trend of the number of citywide traffic accidents in fine-grained time intervals will result in the accurate description of the fluctuation in coarse-grained time intervals,the more accurate prediction effects have been gained when the predicted data at the day level are aggregated at the week level.

Key words:traffic accidents;neural network;LSTM;time series;optimum window

0 引 言

城市交通事故量是衡量一个城市特定时间内道路安全水平的重要指标。交通事故的发生虽然微观上具有一定的不确定性和随机性,但在宏观空间层面上还

收稿日期:2022-04-07 **修回日期**:2022-08-10
基金项目:浙江省交通运输科技计划项目资助(2020026)
作者简介:孙振华(1982-),男,博士研究生,工程师,研究方向为交通数据建模;通讯作者:王转转(1997-),女,硕士研究生,研究方向为交通数据建模。

呈现一定的规律性,因而具有可预测性。当前,预测的方式主要可以分为实时预测和短周期(如周和月)预测。实时预测通常应用深度学习技术结合获取的空间多源交通数据进行模型训练并进行交通事故量的预测[1-2];短周期交通事故量预测通常将交通事故量按时间先后顺序构成序列,对序列数据进行模型训练并进行预测。短周期交通事故量预测是市一级交警部门制定交通决策和交通措施的重要参考。对于经验丰富的交警而言,即使交通事故量时间序列基本趋势表现为明显的季节周期性特征,但叠加在周期性规律中的波动也会使他们也很难对交通事故进行较为准确的估计,因而如何使预测达到可接受的准确度是需要解决的问题。短周期城市交通事故量的准确预测需要捕获交通事故量时间序列内在的时序依赖关系。常用的基于时间序列的典型模型有自回归模型、神经网络模型以及组合模型等。

时间序列自回归模型是一种能对时间序列观测值内在时序依赖关系进行线性表征的一类模型。该类模型应用前需要对时间序列样本数据的平稳性(数据的均值、方差、协方差指标是与时间无关的常数)进行校验以决定模型类型选择并对参数定阶来反映时序依赖关系。平稳的数据适合选用自回归滑动平均模型(Auto - Regressive and Moving Average Model, ARMA),模型阶数 p 与 q 分别表示序列观测值由过去的 p 个序列观测值和 q 个随机扰动的线性组合来表示。这两个参数可通过计算相关系数和偏自相关系数并通过模型参数优化方法最终确定,并由残差序列是否为与时间序列无关的白噪声来评估其有效性。谢华为[3]用具有平稳特性的 2003 至 2015 年的全国交通事故数样本确定 ARMA 参数并对 2011 至 2015 年的交通事故数进行拟合。如果样本数据具有非平稳特点,则需要采用差分自回归移动平均模型(Auto - Regressive Integrated Moving Average Model, ARIMA)进行差分处理,差分次数 d 是该模型的一个输入参数。张杰等[4]发现 1970 至 1997 年全国交通事故十万人口死亡率时间序列样本数据具有非平稳特点,因此采用 ARIMA 进行差分处理并确定模型参数 p 、q 和 d ,并对 1993 至 1997 年的死亡率进行预测。张艳艳等[5]采用 ARIMA 模型对非平稳的 2011 至 2014 年福建海域水上交通的月事故量进行差分处理并确定参数,对 2015 年各月水上交通事故量进行预测并评估误差。季节性差分自回归滑动平均模型(Seasonal Auto - Regressive Integrated Moving Average, SARIMA)在 ARIMA 基础上引入季节性因子来表征数据的周期性特征,并从趋势性、季节性变动以及随机变动三个维度对时间序列数据内在时序依赖关系进行度量。Halim 等[6]观察到印度尼西亚孟加锡市 2016 至 2019 年间的交通事故量具有明显的季节性特征外,再引入了 2020 年新冠病毒流行期间事故量有明显下降趋势的数据,建立 SARIMA 预测模型对 2021 全年的交通事故量的变化趋势进行了预测。

基于神经网络的时间序列模型是能对时间序列观测值变化进行自学习的一类模型。有别于时间序列自回归模型参数的定阶依赖样本数据特征或先验知识,该类模型能自动捕获时间序列样本观测值内在的时序依赖关系并能进行样本外预测,实现这一点的前提往往需要样本数据量足够丰富,如果有同样时空相关的截面数据辅助则更好。安杰等[7]为了预测 2011 年交通事故中的事故数、死亡人数、受伤人数及综合死亡率,选取 1997 到 2010 年时间序列相关数据的同时,还引入了同时期的国内生产总值(GDP)、人口数、公路里程等维度数据,评估它们与同年全国交通事故量的相关性进而形成截面数据,将当年截面数据作为输入以及将来年的年交通事故量作为期望输出值,训练得到基于误差反向传播(Back Propagation, BP)神经网络的道路交通安全预测模型。李兴兵[8]等用 BP 神经网络对城市 2011 至 2016 年的每日数据作为时间序列训练样本,结合机动车年保有量、日天气因素、节假日类型等构成截面数据,并对 2017 年每日数据作为验证样本进行预测。由于 BP 神经网络模型有收敛速度慢、训练时间长、容易陷入局部极小点等缺点,因而张志豪等[9]针对 1998 至 2012 年全国交通事故死亡人数时序数据以及 GDP、国民总收入、人均 GDP 等维度的截面数据,采用长短时记忆神经网络模型(Long Short - Term Memory, LSTM)进行模型参数的训练,并对 2013 至 2016 年全国交通死亡人数进行预测,取得了较好的预测效果。

组合模型是将多个类型模型融合起来对交通事故进行预测的一类模型,与时间序列自回归模型一样都具有很强的数据特性决定模型选择的特征。孙秩轩等[10]发现城市 2006 年 1 月至 2013 年 6 月道路交通事故月度受伤人数为非平稳数据,提出基于 ARIMA 模型和支持向量回归机(Support Vector Regression, SVR)的组合预测模型。该模型确定 ARIMA 参数来拟合 2006 至 2012 年的 84 个数据。由于残差波动具有明显的季节性特征,继而构造含有残差模糊粒子的子序列对 SVR 进行参数寻优回归拟合。这样得到的组合模型的预测准确度相比于单一 ARIMA 模型有明显提高。谢学斌[11]基于 ARIMA 和 XGBoost (Extreme Gradient Boosting,极端梯度提升)组合模型对 1951 至 2010 年全国交通事故量进行拟合。XGBoost 是一种基于决策树的分布式高效梯度提升算法,在该研究中

实现对 ARIMA 模型拟合值残差进行预测。张志豪等[12]提出 LSTM-GBRT(Gradient Boosted Regression Trees, 梯度提升回归树,是一种基于决策树的分布式高效梯度提升算法)组合预测模型,针对 1998 年至 2012 年全国交通事故中的死亡人数以及包含 GDP、国民总收入、人均 GDP 等维度的截面数据,采用 LSTM 进行模型参数的训练,用 GBRT 实现对 LSTM 拟合值残差进行预测,从而提升了 2013 至 2016 年全国交通死亡人数预测的准确度。王臻[13]和张兴强提出了 ARIMA 和模糊神经网络模型(Fuzzy Neural Network, FNN)组合模型,用 ARIMA 来拟合 1980 至 2004 年全国道路交通事故量,用模糊神经网络模型(Fuzzy Neural Network,FNN)以当年的截面数据(公路里程、机动车拥有量、客运周转量、货运周转量以及 GDP、事故起数)为输入,对来年的事故数作为输出量进行监督拟合学习,再通过最优加权方法确定两个模型的权重形成组合模型,利用该组合模型对全国 2005 至 2007 年道路交通事故量进行预测并取得了较好的效果。

研究工作选用 LSTM 对国内某市周交通事故量进行预测的原因是:

(1)对交通事故量进行预测需要捕获特定时空交通事故量时间序列内在的依赖关系,并假定依赖关系保持不变从而能进行预测。然而这种依赖关系会因交通环境、道路、车辆数量、交通参与者随着时间的变化而演变,因而采用 LSTM 进行参数的自学习调整对模型随时间演化就显得很有必要;

(2)当前对交通事故量预测研究所用的数据多为全国级别并以年为单位,尚没有诸如对市级周交通事故量时空下不使用截面数据的相关预测研究,这种短周期城市级的预测对交警的实际决策工作更有意义。

1　长短时记忆神经网络模型

LSTM 模型是递归神经网络(RNN)的一个变种。它通过时间反向传播算法对数据进行训练以解决 RNN 网络存在的梯度消失及无法反映某些长期影响的问题。图 1 是 LSTM 的结构,它的水平方向由一系列彼此相连标注为 A 的存储单元构成,采用了门限机制来控制数据的记忆和遗忘,其中 c_t 代表长时累计信息, a_t 代表短时的存储单元的输出, $\tilde{c}_t$ 代表调整后更新的内容;在垂直方向有输入参量 x_t 和输出参量 h_t 的单向流动。标准的 LSTM 可以表述为:输入节点数量为 n 时,输入序列可表示为 $\{x_{t-n-1},\cdots,x_{t-1},x_t\}$,输出节点数量为 m 时,输出序列可表示为 $\{h_{t-m-1},\cdots,h_{t-1},h_t\}$。每个存储单元有输入门 i_t、遗忘门 f_t 和输出门 o_t。输入门决定更新哪些信息到存储单元;输出门决定存储单元将输出哪些信息;遗忘门决定存储单元中要忘记哪些信息。这 3 个门控制了从输入序列到输出序列的信息流,能体现前序观测值对当前观测值的长短期影响。公式如下:

$$i_t = \delta(\boldsymbol{W}_i \cdot [a_{t-1}, x_t] + \boldsymbol{b}_i) \tag{1}$$

$$o_t = \delta(\boldsymbol{W}_o \cdot [a_{t-1}, x_t] + \boldsymbol{b}_o) \tag{2}$$

$$f_t = \delta(\boldsymbol{W}_f \cdot [a_{t-1}, x_t] + \boldsymbol{b}_f) \tag{3}$$

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

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{5}$$

$$h_t = a_t = o_t * \tanh(c_t) \tag{6}$$

其中,δ 和 tanh 分别代表激活函数 Sigmoid 和双曲线正切函数这两类非线性函数, $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 表示相应的权重系数矩阵和偏置向量,“ * ”表示点乘。LSTM 根据输入序列计算输出序列并与设定的期望值进行误差分析,通过迭代更新系数的学习方式使误差最小化或收敛,从而具有逼近可表征观测值函数的能力,即捕获当前序列观测值和前序观测值的时序依赖关系,最终完成系数调整的 LSTM 神经网络就具有了对训练样本的拟合以及对验证样本的预测能力。正因为如此,LSTM 神经网络及其扩展型已应用到具有时间序列特征但时序观测值关系复杂的交通预测当中,除交通事故量预测外,还应用于短时交通流预测[14-15]、异常驾驶行为检测[16]、道路交通速度预测[17]、铁路货运量预测[18]、交通流状态预测[19]、船舶航迹预测[20]、公交行程时间预测[21]等。

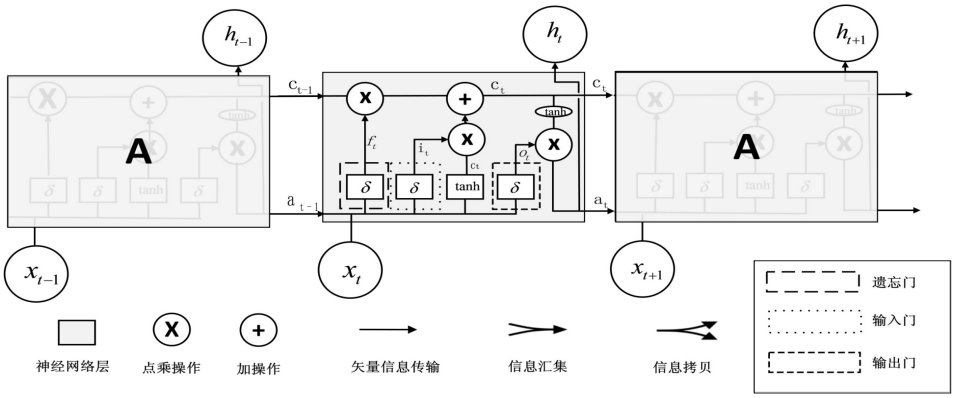

图 1　LSTM 模型逻辑结构

2 用 LSTM 对市级周交通事故量预测方法

2.1 数据预处理

(1)构造日粒度交通事故量时间序列。

对一个城市特定的时间段内按照每日发生的事故起数进行统计,就可构造出日粒度交通事故量时间序列 $\text{Seq\_}X_1 = \{x_1, x_2, \cdots, x_t\}$ 。

(2)时间序列观测值的离差标准化。

由于序列中的交通事故观测值波动较大,这会影响到 LSTM 训练的速度和精度。为消除这种影响,需针对观测值序列进行如式(7)的离差标准化[22]处理。

$$x_i = \frac{(x_i' - \min\{x_j'\})}{(\max\{x_j'\} - \min\{x_j'\})} \tag{7}$$

经过离差标准化处理后的时间序列表示为:

$\text{Seq\_}X_1' = \{x_1', x_2', \cdots, x_t'\}$ 且 $x_t' \in [0,1]$。

2.2 构建对应最优滑动窗口的 LSTM 市级日交通事故量预测模型

(1)构造训练集。

训练集 $\boldsymbol{X}_1$ 与 $\boldsymbol{Y}_1$ 如式(8)所示:

$$\boldsymbol{X}_1 = \begin{bmatrix} x_1' & x_2' & \cdots & x_w' \\ x_2' & x_1' & \cdots & x_{w+1}' \\ \vdots & \vdots & \cdots & \vdots \\ x_{t-w}' & x_{t-w+1}' & \cdots & x_{t-1}' \end{bmatrix}, \boldsymbol{Y}_1 = \begin{bmatrix} x_{w+1}' \\ x_{w+2}' \\ \vdots \\ x_t' \end{bmatrix} \tag{8}$$

其中,$\boldsymbol{X}_1$ 表示训练集的输入,是一个 $(t-w)\times w$ 的矩阵,每一行代表训练 LSTM 的一个输入序列;$\boldsymbol{Y}_1$ 表示训练集的期望输出值,是一个 $(t-w)\times 1$ 的矩阵;w 称为滑动窗口,表示输入序列长度,它将 $\text{Seq\_}X_1'$ 划分为 $\boldsymbol{X}_1$ 与 $\boldsymbol{Y}_1$。

$\hat{\boldsymbol{Y}}_1 = \{x_{w+1}', x_{w+2}', \cdots, x_t'\}$ 是在输入为 $\boldsymbol{X}_1$ 训练 LSTM 逼近 $\boldsymbol{Y}_1$ 的拟合输出结果,即 $\boldsymbol{Y}_1$ 与 $\boldsymbol{X}_1$ 表达了交通事故量的时序依赖关系。

(2)确定 LSTM 超参数。

LSTM 隐含层存储单元个数 N:

隐含层数目设为 1,其存储单元个数按经验公式(9)来确定。

$$N = \sqrt{n+m} + a \tag{9}$$

其中,$n = w$,$m = 1$,a 可取 1 到 10 中的一个值,此处 a 取 10。

损失函数:

损失函数选用平均绝对误差(MAE),表示训练集拟合结果与期望输出值的偏离程度 loss,如式(10)所示。

$$\text{MAE} = \sum_{i=1}^{i=t} |\hat{x}_i' - x_i'| / t \tag{10}$$

其中,$\hat{x}_i'$ 代表训练集离差标准化后的拟合值,x_i' 代表训练集离差标准化后的观测值,t 代表训练集样本数。

迭代次数 epochs:

该参数表示用 $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ 训练 LSTM 并使式(10)的误差 loss 趋向收敛的次数,可通过观察来确定。

(3)确定最优滑动时间窗口算法。

滑动窗口 w 决定了 $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ 构成,不同的 $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ 训练出的 LSTM 预测模型的拟合结果误差有所不同,即 w 间接决定了误差值,因而需要确定一个最优长度的 w',从而使训练出的 LSTM 预测模型拟合误差最小,相应的算法如图 2 所示。

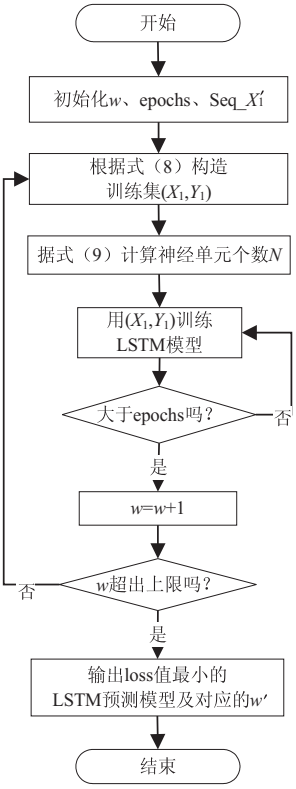

图 2 最优滑动窗口算法流程

2.3 评估对应最优滑动窗口 LSTM 市级日交通事故量预测模型预测结果对验证集的周粒度拟合效果

(1)构造验证集。

经过离差标准化处理后的可用于验证的时间序列表示为 $\text{Seq\_}X_2' = \{x_{t+1}', x_{t+2}', \cdots, x_{t+s}'\}$,$s < t$。根据滑动窗口 w',生成的验证集 $\boldsymbol{X}_2$ 与 $\boldsymbol{Y}_2$ 见式(11)。

$$\boldsymbol{X}_2 = \begin{bmatrix} x_{t-w'+1}' & x_{t-w'+2}' & \cdots & x_t' \\ x_{t-w'+2}' & x_{t-w'+3}' & \cdots & x_{t+1}' \\ \vdots & \vdots & \cdots & \vdots \\ x_{t+s-w'}' & x_{t+s-w'+1}' & \cdots & x_{t+s-1}' \end{bmatrix}$$

$$\boldsymbol{Y}_2 = \begin{bmatrix} x_{t+1}' \\ x_{t+2}' \\ \vdots \\ x_{t+s}' \end{bmatrix} \tag{11}$$

其中, $\boldsymbol{X}_2$ 表示验证集的输入,是一个 $s \times w'$ 的矩阵,每行表示输入 LSTM 预测模型的序列;验证集的期望输出 $\boldsymbol{Y}_2$ 就是 Seq_X_2'。预测模型根据 $\boldsymbol{X}_2$ 输出的预测结果为 $\hat{\boldsymbol{Y}}_2 = \{\hat{x}'_{t+1}, \hat{x}'_{t+2}, \cdots, \hat{x}'_{t+s}\}$。

(2)预测结果对验证集的周粒度误差评估。

对 $\hat{\boldsymbol{Y}}_2$ 进行逆离差标准化后得到的 $\hat{Y}_3 = \{\hat{x}_{t+1}, \hat{x}_{t+2}, \cdots, \hat{x}_{t+s}\}$ 是预测的日粒度交通事故量,对应的观测值 $Y_3 = \{x_{t+1}, x_{t+2} \cdots, x_{t+s}\}$。将 $\hat{Y}_3$ 和 Y_3 转为周粒度 $\hat{Y}_4 = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_l\}$, $l < s$ 和 $Y_4 = \{y_1, y_2, \cdots, y_l\}$, $\hat{Y}_4$ 对 Y_4 的误差可用平均绝对百分比误差(Mean Absolute Percentage Error)指标来评估。

$$\mathrm{MAPE} = \frac{100\%}{l} \sum_{i=1}^{i=l} \frac{|\hat{y}_i - y_i|}{y_i} \tag{12}$$

其中, $\hat{y}_i$ 表示验证集的周粒度统计口径预测值, y_i 表示验证集的周粒度统计口径观测值, l 是验证集的周粒度统计口径样本数。

(3)预测结果对验证集的周粒度拟合效果评估。

$\hat{Y}_4$ 对 Y_4 的拟合效果可用式(13)所示的 R-Square 指标来度量。

$$R^2 = 1 - \frac{\sum_{i=1}^{i=l} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{i=l} (y_i - \bar{y}_i)^2} \tag{13}$$

其中, $\bar{y}_i$ 表示 $\{y_i\}$ 均值。R^2 值通常介于 0 ~ 1 之间,越接近 1 表示预测值的整体拟合效果越好,反之越差。

3　实验结果验证

城市交通事故发生后,执勤交警会记录每个事故人、车、路、环境的详细信息形成一条交通事故记录,包括事故发生的时间、发生路段地点、车辆、驾驶员、环境、财产损失、伤亡人数、雨雪天气等。对来自国内某市 2011 至 2015 年 20 988 条交通事故记录按日统计就形成日粒度交通事故量时间序列 1 824 条。以 2014 年为界形成训练集和验证集,将这些日统计记录以月为单位分成四组,前三组七天为一个单位,后一组为剩余天数,统计形成周粒度交通事故量时间序列 240 条(以下周粒度默认指此类型),将 2011 年至 2014 年的时间序列(其中日粒度为 1 459 条,周粒度为 192 条)作为训练集,2015 年的时间序列作为验证集(其中日粒度 365 条,周粒度 48 条)。经过验证交通事故量是平稳数据,因而也可采用 ARMA 和 SARMA 进行预测对比实验。

在 Keras 框架中,用 Python3.8 对周粒度时间序列训练集按图 2 算法流程计算得到最优滑动窗口 $w' = 45$ 的 LSTM 市级周粒度交通事故量预测模型,其对训练集的拟合结果和验证集的预测拟合结果能很好表达数据的基本趋势,但不能很好地匹配波动的数据,因而图 3 中度量拟合效果和预测拟合效果的 R-square 指标值都欠佳。采用 ARMA 模型和 SARMA 模型也出现类似效果,这说明基于交通事故量训练的用于捕获观测值时序依赖关系的模型对数据基本趋势准确性的表达远好于对其波动性的表达。

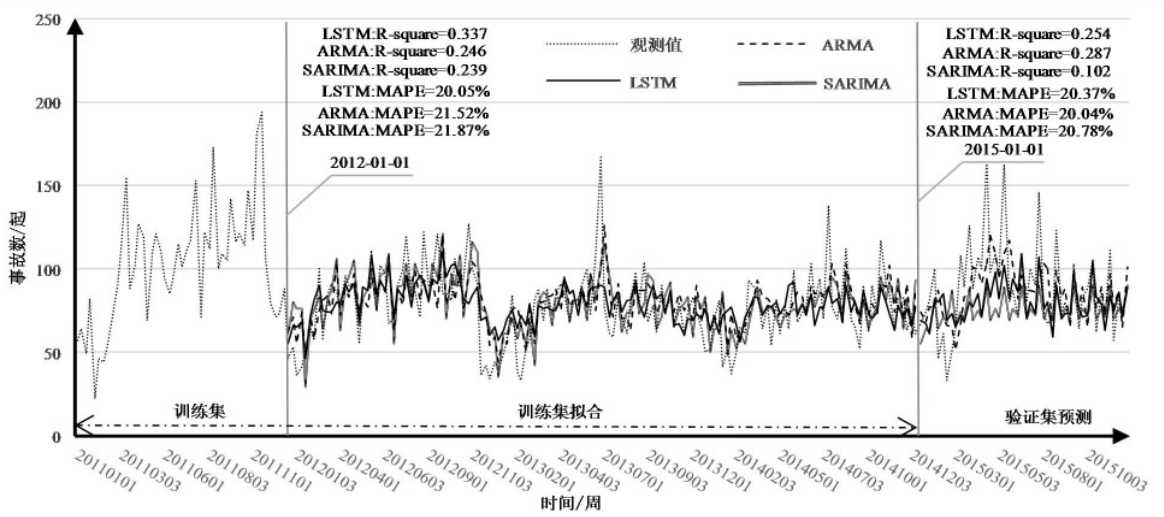

图 3　最优滑动窗口 LSTM 市级周级组粒度交通事故量预测模型对训练集拟合效果以及对验证集预测的拟合效果

图 4 展示了最优滑动窗口 $w' = 72$ 的 LSTM 市级日交通事故量预测模型对训练集的拟合结果和对验证集的预测结果与图 3 类似,因而度量拟合效果和预测拟合效果的 R-square 指标值也都欠佳。需要注意的是,细粒度下的预测结果对交通事故量基本趋势的准确描述可转化为粗粒度下对波动性准确描述的事实,如周粒度下显著的波动在日粒度下则表现为较为平缓的变化趋势叠加小规模的波动。将这三类模型的拟合结果

和预测结果从日粒度转为周粒度统计口径后,如图 5 所示,三个模型相对于图 3 各自都提升了拟合效果和预测拟合效果,尤其是 LSTM 和 ARMA 模型较为明显,度量预测拟合效果的 R-square 指标值分别为 0.817 和 0.832,这意味着预测结果整体上与实际结果吻合程度分别达到了 81.7% 和 83.2%。而 SARIMA 模型拟合效果相对较差的原因可解释为数据季节周期性相对不突出所致。需要明确的是,LSTM 预测模型对验证集较为准确的预测能力实际上是来自对训练集交通事故量时间序列内在依赖关系的学习与量化(ARMA 也是如此),然而图 5 中描述验证集的预测拟合效果的 R-square 指标值 0.817 稍高于训练集拟合效果 0.719,这可解释为正常的交通事故量波动引起。对于市级周交通事故量预测而言,平均绝对百分比误差在 15% 以内是可接受的准确度,图 5 中 LSTM 和 ARMA 模型对训练集的拟合结果和对验证集的预测结果的平均绝对百分比误差都在这一范围内。

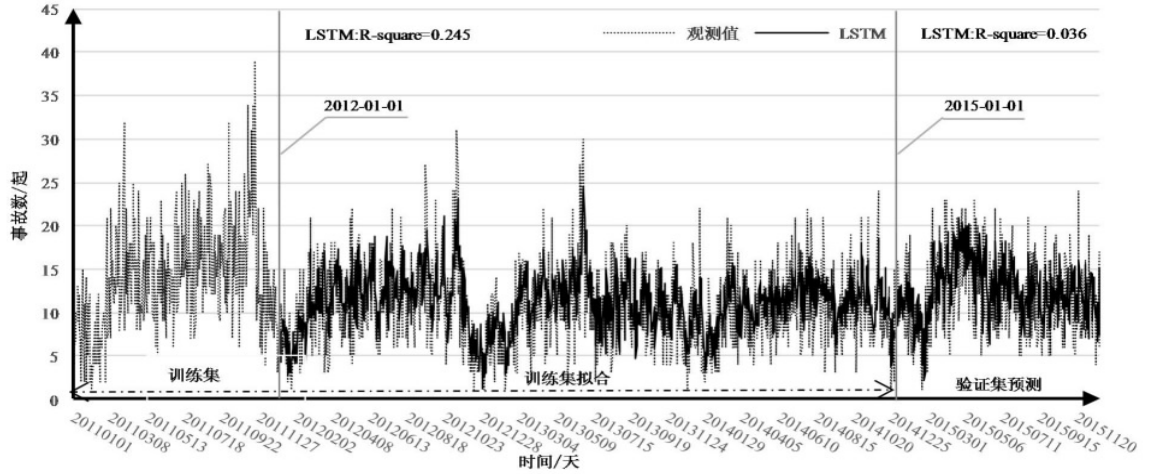

图 4 最优滑动窗口 LSTM 市级日交通事故量预测模型对训练集拟合效果以及对验证集预测的拟合效果

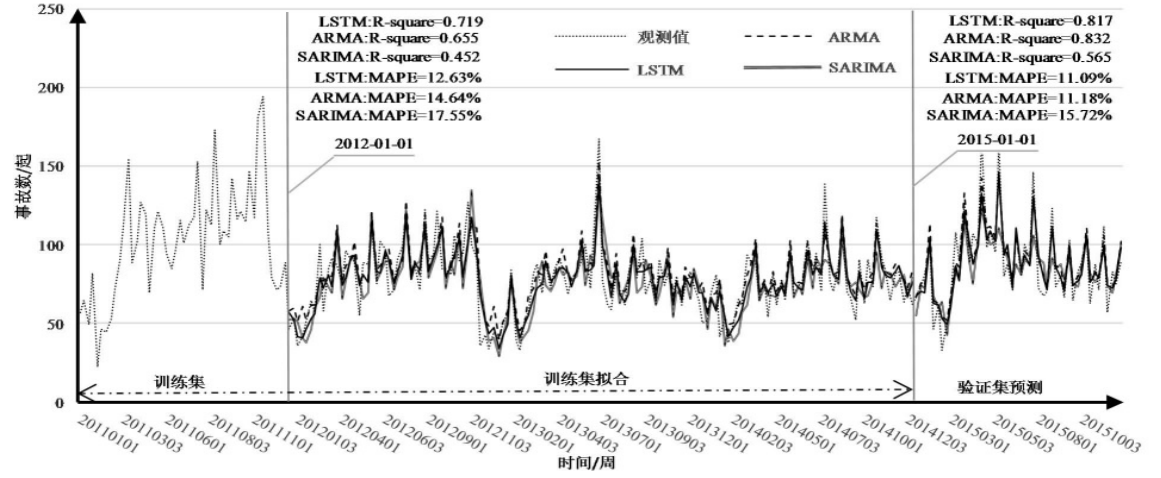

图 5 最优滑动窗口 LSTM 市级日交通事故量预测模型的输出结果转为周粒度时对训练集拟合效果以及对验证集的预测拟合效果

图 6 展示了 2011 ~ 2014 年期间七天为一周的统计口径得到的周交通事故量以及划分的训练集和验证集,与图 5 对比可发现,虽然交通事故量曲线有了明显变化,但是 LSTM 和 ARMA 则保持了同样好的拟合效果和预测拟合效果,从另一个侧面也印证了细粒度下预测结果对交通事故量基本趋势的准确描述可转化为粗粒度下对波动性的准确描述。虽然 LSTM 和 ARMA 都取得了较好的预测效果,但 LSTM 不像 ARMA 那样需要人工辅助来进行参数定阶,这个优点有利于 LSTM 随时间而滚动更新参数以保证预测的准确性,毕竟市级周交通事故时间序列内在依赖关系会随时间有所变化。然而需要注意的是,滑动时间窗口的长度对基本趋势准确描述有直接影响。图 7 展示了滑动窗口 $w=306$ 的 LSTM 市级日交通事故量预测模型输出结果转为周粒度后对训练集的拟合和对验证集的预测拟合效果,两个 R-square 指标值的显著下降可解释为 $w=306$ 滑动窗口首先会造成日粒度下对验证集的基本趋势描述准确度下降,进而影响了周粒度下对数据

波动描述的准确度。图 8 展示了滑动窗口长度 w 与周粒度拟合效果和预测拟合效果的 R-square 指标值的关系——呈现了先增大后减小趋势,这个变化过程说明采用最优窗口算法为 LSTM 市级日交通事故量预测模型确定最优窗口长度很有必要。

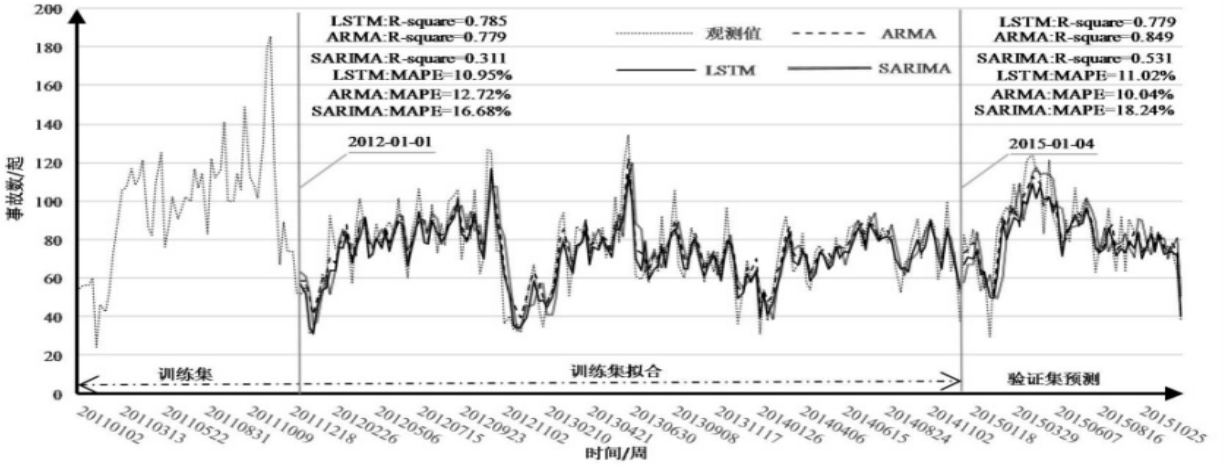

图 6　最优滑动窗口 LSTM 市级日交通事故量预测模型的输出结果转为七天一周统计口径时对训练集拟合效果以及对验证集的预测拟合效果

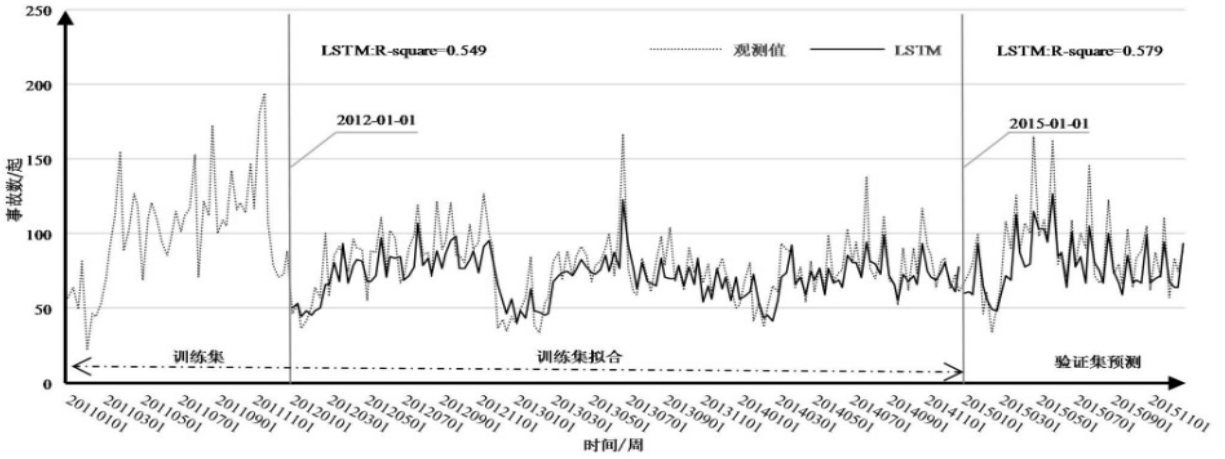

图 7　w =306 滑动窗口 LSTM 市级日交通事故量预测模型输出结果转为周粒度时对训练集拟合效果以及对验证集预测的拟合效果

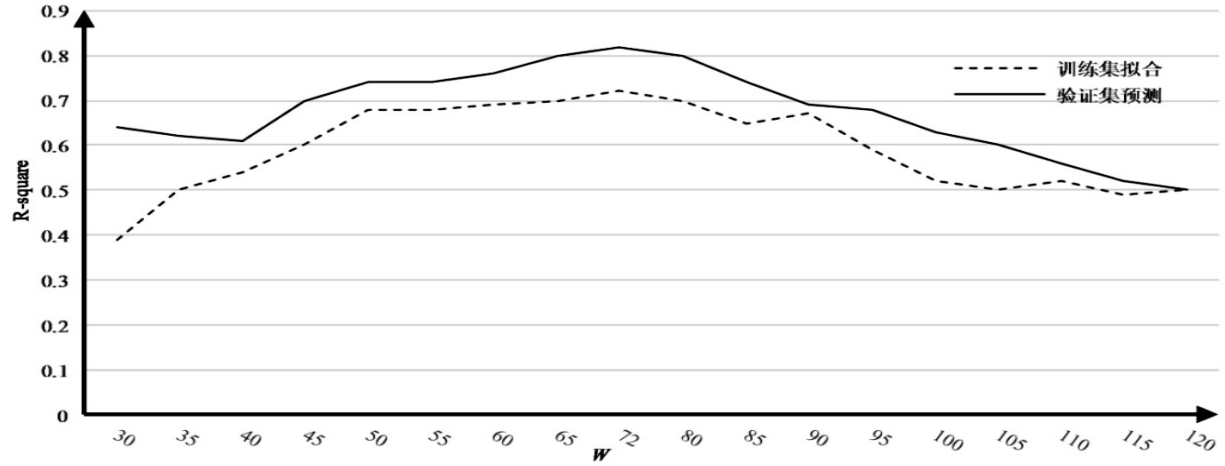

图 8　LSTM 市级日交通事故量预测模型输出结果转为周粒度时对训练集拟合以及对验证集预测的 R-square 值与滑动窗口长度关系

实验结果表明,提出的“用 LSTM 对市级周交通事故预测方法”可基于市级日粒度交通事故量时间序列对周交通事故量进行较为准确的预测。对交通事故量进行预测目前类似研究所用的数据多为全国每年的交通事故量时间序列,但由于该类数据量较少而很难发挥神经网络模型自我学习能力来捕获数据的时序依赖关系,往往需要补充多维截面数据,但截面数据会因涉及多个行业部门以及在统计上的滞后会影响交通事故量预测的时效性,而只基于交通事故量时间序列进行预测则会减少这方面的困难并增强预测模型的实用性。

4 结束语

(1)提出了用 LSTM 对市级周交通事故量预测方法。该方法通过构建一个对应最优输入序列长度的 LSTM 市级日交通事故量预测模型捕获交通事故量时间序列中的当前观测值与前序观测值的时序依赖关系,当将预测结果转为周粒度统计口径后,就实现了对交通事故量较为准确的预测。该方法不需要相关截面数据,因而对市级交警预测交通事故量具有实用价值。

(2)市级交通事故量时间序列的波动是影响对其准确预测的关键因素,所提的预测方法解决了影响市级周交通事故量准确预测的问题。该方法发现基于交通事故量训练的用于捕获观测值时序依赖关系的 LSTM 模型对数据基本趋势准确性的表达远好于对数据波动性的表达,为此提出最优窗口算法来确定 LSTM 模型最优窗口长度,以确保对训练集基本趋势表达的准确性,再根据所发现的预测结果对细粒度交通事故量基本趋势的准确描述可转化为粗粒度下对波动性准确描述的事实,将日粒度预测结果转为周粒度后就取得了较准确的预测效果。

(3)用 LSTM 对市级周交通事故量预测方法能进行较为准确预测的前提是验证集和训练集保持相同的时序依赖关系。随着时间的推进,如果预测结果与训练集时间过久很难保证这种时序依赖关系不发生变化,因而下一步将研究市级日交通事故量 LSTM 预测模型的自我优化更新机制来保持预测的准确性。

参考文献:

[1] BAO J,LIU P,UKKUSURI S V. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data[J]. Accident Analysis & Prevention,2019,122:239-254.

[2] WANG J H,SONG H,FU T,et al. Crash prediction for freeway work zones in real time:a comparison between convolutional neural network and binary logistic regression model[J]. International Journal of Transportation Science and Technology,2022,11(3):484-495.

[3] 谢华为. 基于 ARMA 平稳时间序列的道路交通事故预测[J]. 宁德师范学院学报:自然科学版,2018,30(3):268-272.

[4] 张 杰,刘小明,贺玉龙,等. ARIMA 模型在交通事故预测中的应用[J]. 北京工业大学学报,2007,33(12):1295-1299.

[5] 张艳艳,刘晓佳,熊子龙,等. 基于 ARIMA 模型的水上交通事故预测[J]. 中国水运:下半月,2017,17(2):51-54.

[6] HALIM H,BUSTAM B,SAING Z. The Forecasting of a traffic accident in the pandemic of Covid-19[J]. Turkish Journal of Computer and Mathematics Education (TURCOMAT),2021,12(6):3664-3669.

[7] 安 杰,董龙洋. 基于 BP 神经网络的道路交通安全预测模型研究[J]. 公路与汽运,2014(3):63-67.

[8] 李兴兵,黄 力. 基于神经网络的区域交通事故数预测建模研究[J]. 信息系统工程,2020(5):139-142.

[9] 张志豪,杨文忠,袁婷婷,等. 基于 LSTM 神经网络模型的交通事故预测[J]. 计算机工程与应用,2019,55(14):249-253.

[10] 孙轶轩,邵春福,计 寻,等. 基于 ARIMA 与信息粒化 SVR 组合模型的交通事故时序预测[J]. 清华大学学报:自然科学版,2014,54(3):348-353.

[11] 谢学斌,孔令燕. 基于 ARIMA 和 XGBoost 组合模型的交通事故预测[J]. 安全与环境学报,2021,21(1):277-284.

[12] ZHANG Zhihao,YANG Wenzong,WUSHOUR S. Traffic accident prediction based on LSTM-GBRT model[J]. Journal of Control Science and Engineering,2020(2020):1-10.

[13] 王 臻,张兴强. 基于 ARIMA-FNN 的道路交通事故最优加权组合预测模型[J]. 交通信息与安全,2010,28(3):89-92.

[14] 温惠英,张东冉. 基于 Bi-LSTM 模型的高速公路交通量预测[J]. 公路工程,2019,44(6):51-56.

[15] 李明明,雷菊阳,赵从健. 基于 LSTM-BP 组合模型的短时交通流预测[J]. 计算机系统应用,2019,28(10):152-156.

[16] 惠 飞,郭 静,贾 硕,等. 基于双向长短记忆网络的异常驾驶行为检测[J]. 计算机工程与应用,2020,56(24):116-122.

[17] 阎嘉琳,向隆刚,吴华意,等. 基于 LSTM 的城市道路交通速度预测[J]. 地理信息世界,2019,26(5):79-85.

[18] 程肇兰,张小强,梁 越. 基于 LSTM 网络的铁路货运量预测[J]. 铁道学报,2020,42(11):15-21.

[19] 马焱棋,林 群,赵昱程等. 基于深度学习 LSTM 对交通流状态的预测[J]. 数学的实践与认识,2021,51(4):47-56.

[20] 李 永,成梦雅. LSTM 船舶航迹预测模型[J]. 计算机技术与发展,2021,31(9):149-154.

[21] 徐丸絮,沈吟东. 基于 Attention-LSTM 神经网络的公交行程时间预测[J]. 现代电子技术,2022,45(3):83-87.

[22] HU Zhenhua,LIU Xin,HU Pan. The impact of bike sharing on urban transportation[J]. The Frontiers of Society,Science and Technology,2019,1(4):288-298.