基于门控循环单元的铁路客票业务流量数据预测

谢　泽[1],朱建生[2],李　雯[2]
(1. 中国铁道科学研究院,北京 100081;
2. 中国铁道科学研究院电子计算技术研究所,北京 100081)

摘　要:铁路客票业务流量数据是反映系统业务运行状态的重要记录,为加强流量数据异常预警,针对流量数据具有历史规律性及突变性的特点,选用适于解析数据时间序列依赖度高的门控循环单元神经网络模型(GRU),对流量数据实现时序拟合及趋势预测。GRU 采用不同时间步长对流量数据进行拟合的结果在整点或半点周期时间步长具有局部最小特征,该特征与铁路售票时刻规则形成的时间序列依赖规律相一致。在相同数据条件下,使用 GRU 算法与自回归模型等主流预测算法进行拟合准确度对比,结果证明 GRU 在解析铁路客票业务流量数据依赖方面具备较高的准确性。经过对异常流量数据趋势预测及拟合,在数据异常区间,预测结果与真实数据的拟合近似验证了 GRU 算法能够为铁路客票业务流量数据异常预警提供可行性策略。

关键词:门控循环单元;流量数据;时序拟合;趋势预测;数据预警
中图分类号:TP391.9　　**文献标识码**:A　　**文章编号**:1673-629X(2021)10-0209-06
doi:10.3969/j.issn.1673-629X.2021.10.035

Network Traffic Data Forecast of Railway Passenger Ticket Service System Based on Gated Recurrent Unit

XIE Ze[1],ZHU Jian-sheng[2],LI Wen[2]
(1. China Academy of Railway Sciences,Beijing 100081,China;
2. Institute of Computing Technologies,China Academy of Railway Sciences,Beijing 100081,China)

Abstract:The traffic data of railway passenger ticket service is important record to reflect the operation status of the system. In order to strengthen the early warning of abnormal traffic data,according to the characteristics of historical regularity and mutation of traffic data, the gated recurrent unit neural network model (GRU) which is suitable for analyzing data with high dependence on time series is selected to realize time series fitting and trend prediction. The results of GRU fitting the flow data with different time steps have the local minimum feature in the whole point or half point cycle time, which is consistent with the time series dependence rule formed by the railway ticketing time rule. Under the same data conditions,GRU algorithm and auto regression model and other mainstream prediction algorithms are used to compare the fitting accuracy. The results show that GRU has high accuracy in analyzing the traffic data dependence of railway passenger ticket service. After the trend prediction and fitting analysis of abnormal traffic data, the difference between the predicted results and the real data is obvious in the data abnormal interval, which verifies that GRU algorithm can provide feasible strategies for the early warning of abnormal traffic data of railway passenger ticket service.

Key words:gated recurrent unit;network traffic data;time series fitting;trend prediction;data early warning

0　引　言

中国铁路客票系统作为国内唯一的官方火车票销售渠道,自 2011 年应用互联网售票模式以来,用户数已达 5.4 亿。在节假日旅客出行高峰期,中国铁路客票系统日售票量可达 1 000 万张。当系统运行出现异常时不仅会影响乘客的出行体验,还会给社会造成巨大损失。因此,对国内铁路客票系统的异常状态监测显得尤为重要。当前铁路客票系统已实现了全链路监控,其中关键指标阈值监测是最常用的方法。如周期时间内互联网用户登录人数的阈值监控,周期时间内用户支付购票数的阈值监控,以及服务器集群关键性能指标的阈值监控等。但是,在一些异常情况下,如某些服务器响应超时导致的集群崩溃、网络流量激增等,以上独立的关键指标监控可能产生滞后效应,导致预

收稿日期:2020-11-03　　**修回日期**:2021-03-04
基金项目:中国国家铁路集团有限公司科技研究开发计划课题(K2019X008)
作者简介:谢　泽(1992-),男,博士研究生,研究方向为信息技术;朱建生,研究员,研究方向为铁路信息化建设。

警监测不及时。铁路客票系统业务流量数据能够较为实时地反映系统整体运行状态,为了从整体上对铁路客票系统进行监控,业务流量数据的时间序列预测和趋势预测成为加强异常预警的主要手段。

中国铁路客票系统的业务流量数据主要由用户登录服务数据、车次余票信息查询服务数据、购票服务数据等组成,不同的业务在实现方案中具有一定的串行和并行关系。除此之外,由于售票规则规定小时整点或半点开始售票,业务流量数据在时间维度上具有长短相关性和自相似性。因此,铁路客票系统业务流量数据的预测不同于传统的流量数据预测,如何通过挖掘具有特殊相关性的网络流量数据来提高预测精度成为一个亟待解决的问题。

1 相关工作

流量数据拟合算法已在交通行业[1-2]、互联网金融[3-4]以及临床医学[5-6]等领域广泛使用。基于回归方程的传统统计算法在以上领域很难通过复杂模型实现时序数据的高精度拟合,虽然机器学习算法通过反复训练可以得到比传统统计模型更加精确的预测结果[7-9],但是当其被用于分析输入变量之间具备强序列依赖关系数据时,性能下降[10]。而具备特殊“重置门”和更新门的 GRU 神经网络可以通过设定不同的时间步长挖掘滑动窗口内的序列依赖,且该算法易于解构,可依据不同条件改良为适于实际的预测模型,并有高拟合准确度[11-13]。与以上领域建模任务不同,由于铁路客票在售票时刻上存在整点、半点等起售的业务特点,流量数据呈现出历史规律性及突变性。因此选择一个适用于具有复杂、随机特点的铁路客票业务流量数据预测模型,充分解析序列间依赖关系是至关重要的。

文中针对铁路售票时刻特点,选用 GRU 对铁路客票业务流量数据进行拟合,当该算法时间步长在 1 至 100 范围内,步长值分别为 1、30、60、90 时,对应流量数据拟合结果为局部最优。局部最优的规律与客票售票时间规则形成数据依赖规律相一致,证明 GRU 适于解析铁路客票业务流量数据。

2 流量数据特性

铁路客票业务流量数据主要由余票查询、用户登录、购票等业务形成的网络流量构成,其中余票查询业务量占全部业务量的 60% 以上,在单日内其余业务在风控层请求总数都具备明显且规律的凹凸性。如图 1 所示,该规律与在铁路售票时刻前后有大量用户进行登录、购票等操作的社会行为一致。由于余票查询不需要用户登录,因此余票查询业务在铁路客运业务每日的服务时间内都被用户大量访问,图中展现出余票查询业务对于流量数据总量的变化会有减弱规律特征的作用。

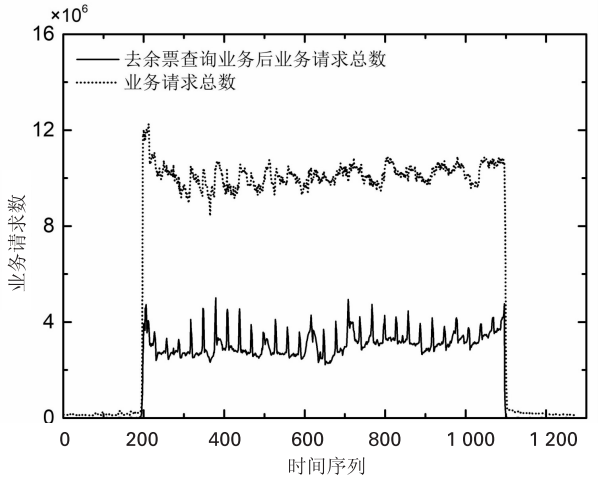


图 1 风控层客票业务请求次数统计

图 2 为多日流量数据统计,从中可以看出售票规则导致流量数据在数理结构上具有直观的长短相关性、自相似性,短相关性取决于铁路售票时刻规则,长相关性取决于售票规则在长时间段内保持不变。文中选用的 GRU 神经网络可以通过设定不同的时间步长挖掘滑动窗口内的序列依赖,适于挖掘铁路客票业务流量数据内含的长短相关性及自相似性。

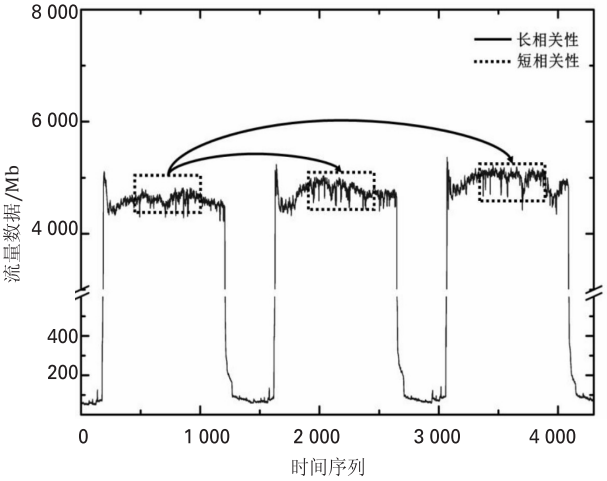


图 2 流量数据长短相关性

3 流量数据预测

3.1 GRU 算法简介

GRU 基于循环神经网络模型(RNN),解决了 RNN 求导过程链中连乘导致的梯度消失问题,并且该模型由长短时记忆神经网络进化。图 3 所示为 GRU 隐藏层细胞单元的具体展开,R_t、Z_t、U_t、X_t、h_t 分别为重置门、更新门、候选隐藏状态、输入以及隐藏状态,虚线边框 GRU 重置门模块的结构特性决定该算法适于解决长跨度依赖。由于铁路客票起售规则,所以每

30 分钟整数倍时间内统计的流量数据之间存在较强依赖关系,GRU 时间步长的调节可以改变解析数据间依赖跨度的大小[14]。

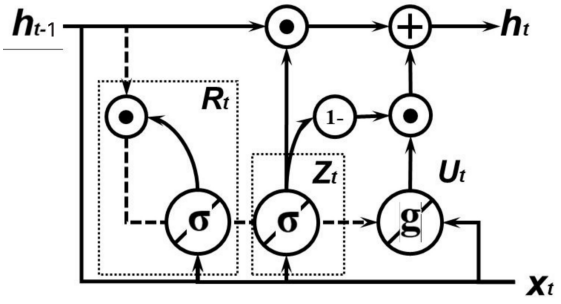


图 3　GRU 隐藏层细胞单元

GRU 隐藏层细胞单元模型的前向计算如下, $\boldsymbol{W}$ 和 b 分别为相应的权重系数矩阵和偏置项, σ 为 sigmod 激活函数,在重置门模块中决定了对之前序列数据的记忆程度, g 为 tanh 双曲正切函数:

$$R_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + b_i) \tag{1}$$

$$z_t = \sigma(W_{xf}X_t + W_{hi}h_{t-1} + b_f) \tag{2}$$

$$d_t = \sigma(W_{xc}X_t + W_{hc}(h_{t-1} \odot R_t) + b_c) \tag{3}$$

$$U_t = g(d_t) \tag{4}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot U_t \tag{5}$$

GRU 模型训练步骤如下:

Step1 处理样本集:假设经过小波前置分析后得到近似信号的时间序列为 $F = \{f_1, f_2, \cdots, f_n\}$,那么可以将该序列划分为 GRU 模型训练集与测试集,分别用 $F_{\text{tr}} = \{f_1, f_2, \cdots, f_v\}$, $F_{\text{te}} = \{f_{v+1}, f_{v+2}, \cdots, f_n\}$ 表示,其中 $v < n$ 且 v , n 为正整数。

Step2 设定 GRU 参数进行训练:通过设定时间步长 L 对序列进行窗口分割,分割后模型输入为 $X = \{X_1, X_2, \cdots, X_{v-L}\}$, $X_q = \{f_q, f_{q+1}, \cdots, f_{v-L+q}\}$ 。该输入经过隐藏层后形成的输出为 $P = \{P_1, P_2, \cdots, P_{v-L}\}$,对应的理论输出为 $Y = \{Y_1, Y_2, \cdots, Y_L\}$ 。选择公式(6)损失函数最小化作为优化目标,遍历学习率及训练步数范围,使用 adam 作为优化算法不断更新网络权重,得到在训练集范围及相关参数范围内最优神经网络。

$$\text{loss} = \sum_{i-1}^{L(v-L)} (p_i - y_i)^2 / [L(v - L)] \tag{6}$$

3.2　数据拟合准确度选择

平均绝对误差 mean absolute error (MAE)、平均绝对百分误差 mean absolute percentage error (MAPE)和均方根误差 root mean square error (RMSE)是最常用的数据拟合准确度预测评价指标。MAE 主要用于测量实验数据集的预测值与实际值之间的平均绝对误差。MAE 定义为:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |a(t) - f(t)| \tag{7}$$

MAPE 主要用于测量拟合数据与真实数据的百分比误差,MAPE 数学定义为:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{a(t) - f(t)}{a(t)}\right| \times 100\% \tag{8}$$

RMSE 主要用于衡量拟合数据与真实数据的均方根差,RMSE 数学定义为:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} |a(t) - f(t)|^2} \tag{9}$$

式(7)、(8)、(9)中, $a(t)$ 表示第 t 个时间序列真实数据, $f(t)$ 表示第 t 个时间序列拟合数据, n 为时间序列总数。从数理角度分析,MAE 和 MAPE 在数学形式上都属于 L1 范数,而 RMSE 在数学形式上属于 L2 范数,数学表达式的幂级数约高,那么这个表达式的输出结果则对异常值越敏感。这也代表当拟合数据集合中出现一个异常大或异常小的数据值时,RMSE 的计算结果将比 MAE 和 MAPE 大。假设有如下两个数据集合:

set_1 = [5,10,5,10,5,10,5,10,5,10, 5,10,5,10, 5,10,10]

set_2 = [5,10,5,10,5,10,5,10,5,10,5,10,5,10,5, 10,1 000]

集合 set_1 和 set_2 的区别为末尾数据分别是 10 和 1 000,集合 set_2 的末尾数据 1 000 可以看作为拟合数据的异常大值。将 set_1 集合作为真实数据, set_2 集合作为拟合数据,由此计算 MAE 为 58.2,RMSE 为 240.1, MAPE 作为 MAE 的数学变换,敏感幅度低于 MAE。由此可证明,对于时间序列拟合结果的异常值,RMSE 比 MAE 和 MAPE 更加敏感,因此选择 RMSE 作为拟合准确度的定义表达。

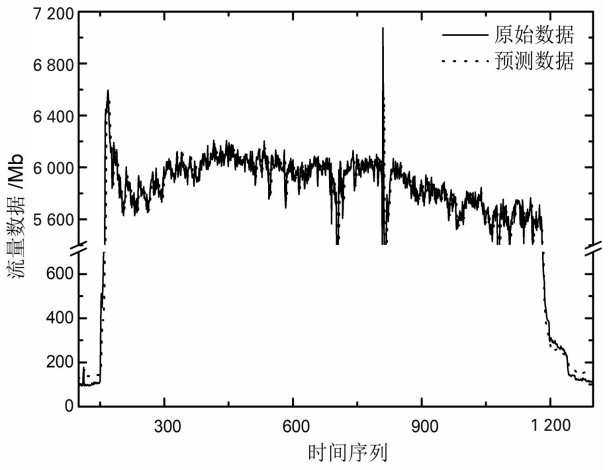


图 4　流量数据与拟合结果对比

对 2017 年 7 月 10 日流量数据进行 GRU 拟合, GRU 优化算法为 adam 函数,时间步长为 30,学习率 $\eta = 0.05$。使用 7 月 8、9 日两天数据作为训练集,拟合结果与测试集数据对比如图 4 所示,数据拟合准确

度定义为 RMSE,拟合结果与测试集数据基本一致,拟合准确度的值为 88.23。

3.3 不同时间步长拟合结果对比及趋势预测

GRU 模型中的时间步长 L 决定了在流量数据拟合中使用 L 个序列来预测第 $L+1$ 个序列值。图 5 展示了 L 为 1 至 100 时,GRU 模型对 2017 年 8 月 8 日流量数据的拟合准确度变化,黑色标点代表 L 为 5 的整倍数拟合均方根误差,当 L 分别为 1、30、60、90 时均方根误差处于局部极小,该结果证明 GRU 在解析流量数据过程中,对 1 分钟、30 分钟、60 分钟以及 90 分钟内流量数据间依赖敏感,验证了由铁路整点或半点售票时刻规则所决定的数据结构特性适于使用 GRU 算法解析。

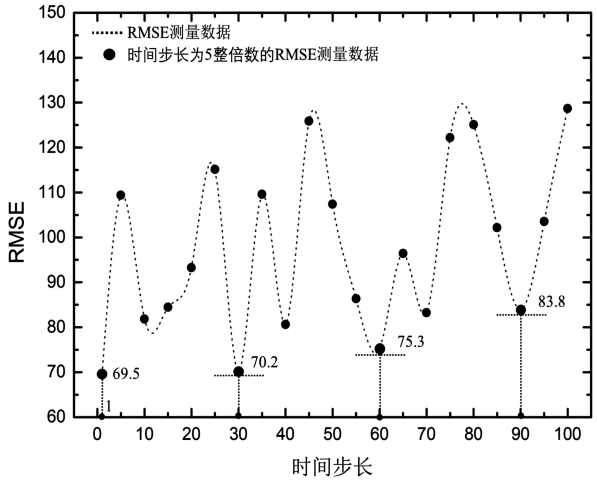


图 5 不同时间步长 RMSE 拟合统计

流量数据趋势预测也是异常预警的常用手段,文中采取的趋势预测基于 GRU 算法拟合数据。假设拟合数据长度为 L,通过设定窗口大小 U,滚动遍历 L,并且对窗口范围内拟合数据 P 使用最小二乘法得到线性拟合,$P=\{P_{n+1},P_{n+2},\cdots,P_{n+U}\}$,$n\in(1,L-U)$。对 2018 年 6 月 18 日去除服务开启与关闭时刻前后的流量数据进行 GRU 分析,如图 6 所示,在 GRU 拟合结果的基础上,实施滑动窗口大小为 100 的趋势预测,其他任意的窗口大小的预测趋势都与拟合结果基本吻合。

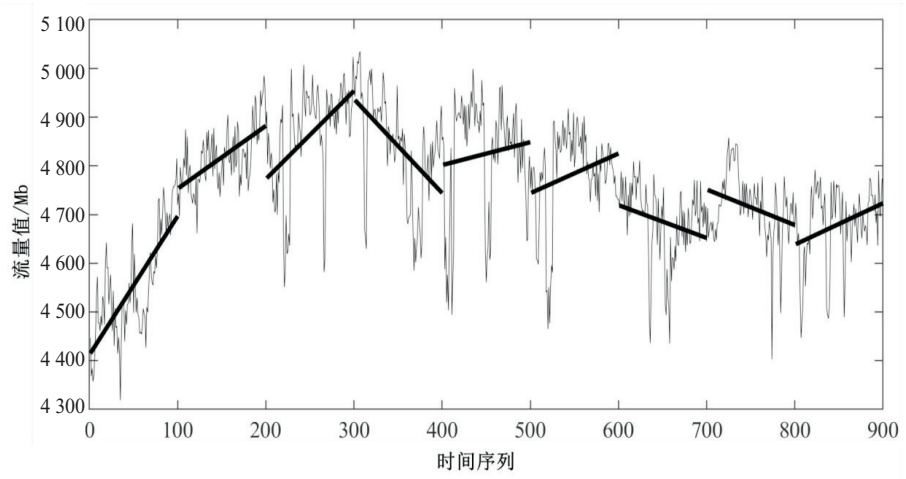


图 6 流量数据趋势预测

4 算法验证

4.1 对比模型

为了验证 GRU 算法的有效性,文中将 GRU 算法与以下 3 种时间序列预测模型进行对比。

(1)多元线性回归。

多元线性回归 MLR 是传统的统计学方法用于对多变量、多影响因素进行分析从而实现预测[15]。该模型函数如下:

$$y_t=a_0+a_1y_{t-1}+a_2y_{t-2}+\cdots+a_ky_{t-k}+b \quad (10)$$

其中,y_t 为 t 时序点预测目标,$y_{t-1},y_{t-2},\cdots,y_{t-k}$ 为 t 时刻前 k 个历史数据,a 为模型系数,b 为固定偏差。多元线性回归将不同时刻的历史数据作为不同的自变量,k 也称为数据窗口长度,在对比模型中选取 k 与测试集样本个数相同。

(2)自回归移动平均。

ARIMA 差分整合移动平均回归模型是自回归模型的经典算法,利用历史时序数据对当前时刻进行预测,模型本身采用不同时刻数据差分的原理解决对样本数据平稳性要求高的问题[16]。该模型函数可用 ARIMA(p,d,q)表示,p,d,q 分别为自回归项数、差分阶数及移动平均项数。差分 1 阶或 2 阶即可满足模型需求,p,q 需要自相关函数及偏自相关函数来确定。

(3)LSTM 长短时记忆神经网络模型。

LSTM 长短时记忆神经网络模型利用“遗忘门”结构实现了序列数据间依赖解析,并通过遗忘门决定历史时刻状态信息保留至当前状态的信息量大小[17-18]。GRU 作为 LSTM 神经网络模型的改进,在保证高拟合精度的情况下,降低了模型训练耗时。文中使用相同时间步长及训练批次对 LSTM 与 GRU 算法

进行对比。

4.2　对比结果

使用以上四种算法分别对 2017 年 12 月 5 日至 12 月 9 日 5 天流量数据进行拟合,10 月 23 日至 10 月 27 日流量数据作为训练集。文中所有计算结果使用计算机的配置相同:处理器为 Intel i5-7300;内存为 8 GB;显卡为 GTX 1050Ti;操作系统为 Windows 10(64 位);实现计算机开发环境为 PyCharm 2018.2.4;实现语言为 Python 3.6.9;程序开发 GRU 框架使用 Python Tensorflow 程序包。图 7 为 5 天流量数据拟合准确度变化,GRU 的平均拟合准确度最优 91.8,虽然 LSTM 拟合准确度与 GRU 接近,但是在时间步长与训练批次相同的训练过程中,GRU 耗时相对于 LSTM 具备极大优势。

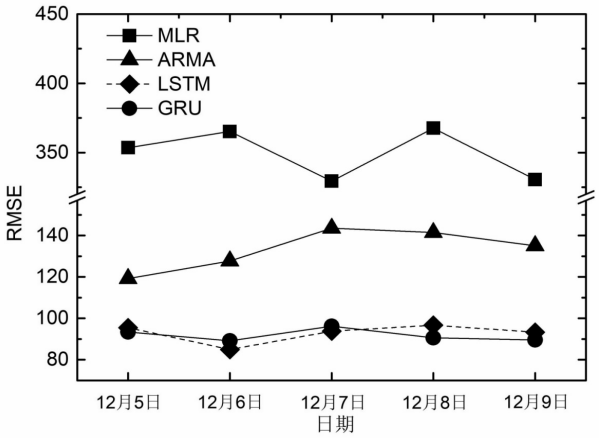


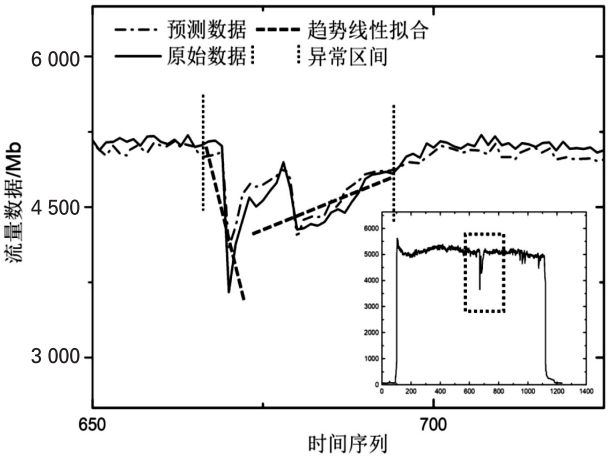
图 7　多种算法拟合对比统计

表 1　算法拟合对比 RMSE 数值统计

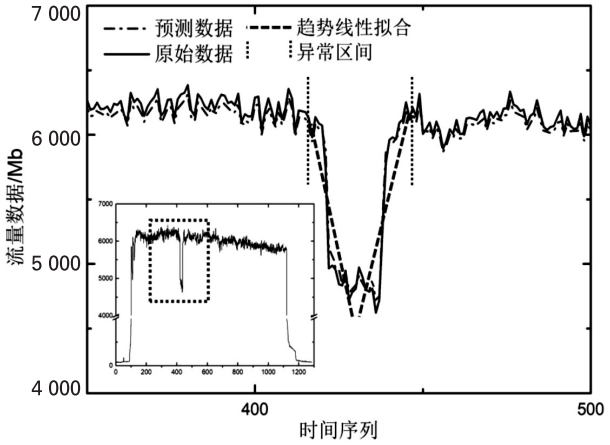
模型	12 月 5 日	12 月 6 日	12 月 7 日	12 月 8 日	12 月 9 日
MLR	353.5	365.2	329.4	367.6	360.5
ARMA	119.2	127.7	143.6	141.5	135.1
LSTM	95.5	84.9	93.8	96.7	93.3
GRU	93.4	89.1	96.2	90.6	89.6

4.3　异常数据检测

为了进一步验证 GRU 在铁路客票业务中异常预警的可行性,选取了 2018 年 1 月 7 日、2 月 11 日的流量数据进行验证,该 2 日内的流量数据分别在当天有大范围波动,与实际部分业务故障相符。选取以上测试集前 2 天流量数据作为训练集样本,故障点附近范围的趋势预测及流量数据拟合结果如图 8 所示。两日流量数据拟合 RMSE 分别为 114.5 和 99.3,GRU 拟合结果与实际流量数据的近似且线性趋势预测的准确。



(a)1 月 7 日流量数据



(b)2 月 11 日流量数据

图 8　异常流量数据检测

5　结束语

文中提出了基于 GRU 神经网络的流量数据时间序列预测方法,不同时间步长的拟合结果与铁路售票规则形成的数据依赖规律相一致,体现了 GRU 在长跨度客票业务流量数据依赖解析方面的优势。通过 3 种不同时间序列预测模型与 GRU 流量数据拟合对比,GRU 的最优拟合准确度证明了该算法具备较高的准确性。经过 GRU 对异常流量数据分析,在故障点处的拟合近似以及趋势预测准确,使得预测差值与趋势线性拟合斜率具备成为铁路客票业务流量数据预警监控指标的可能。

参考文献:

[1]　李　嘉,刘春华,胡赛阳,等. 基于交通数据融合技术的行程时间预测模型[J]. 湖南大学学报:自然科学版,2014

(1):33-38.
[2] 陆百川,舒 芹,马广露. 基于多源交通数据融合的短时交通流预测[J]. 重庆交通大学学报:自然科学版,2019,38(5):13-19.
[3] 黄有为,高 燕. 基于循环神经网络的金融数据预测系统[J]. 软件导刊,2019,18(1):28-33.
[4] 张栗粽,王谨平,刘贵松,等. 面向金融数据的神经网络时间序列预测模型[J]. 计算机应用研究,2018,35(9):2632-2637.
[5] 卓飞豹,宋 斌,雷 勇,等. 基于多预测器融合的医学时间序列数据预测[J]. 中国数字医学,2010,5(10):24-26.
[6] 陈 旭,刘鹏鹤,孙毓忠,等. 面向不均衡医学数据集的疾病预测模型研究[J]. 计算机学报,2019,42(3):596-609.
[7] JORDAN M I,MITCHELL T M. Machine learning:trends,perspectives,and prospects[J]. Science,2015,349(6245):255-260.
[8] WANG X,HAN M. Online sequential extreme learning machine with kernels for nonstationary time series prediction[J]. Neurocomputing,2014,145:90-97.
[9] SHI Z,HAN M. Support vector echo-state machine for chaotic time-series prediction[J]. IEEE Transactions on Neural Networks,2007,18(2):359-372.
[10] HUA Y,ZHAO Z,LI R. Deep learning with long short-term memory for time series prediction[J]. IEEE Communications Magazine,2019,57(6):114-119.
[11] 赵 兵,王增平,纪维佳,等. 基于注意力机制的 CNN-GRU 短期电力负荷预测方法[J]. 电网技术,2019,43(12):4370-4376.
[12] FU R,ZHANG Z,LI L. Using LSTM and GRU neural network methods for traffic flow prediction[C]//2016 31st youth academic annual conference of Chinese association of automation (YAC). Wuhan:IEEE,2016:324-328.
[13] SAJJAD M. A novel CNN-GRU-based hybrid approach for short-term residential load forecasting[J]. IEEE Access,2020,8:143759-143768.
[14] QU Z,SU L,WANG X,et al. A unsupervised learning method of anomaly detection using GRU[C]//2018 IEEE international conference on big data and smart computing (BigComp). Shanghai:IEEE,2018:685-688.
[15] QUIMING N S,DENOLA N L,SAITO Y,et al. Multiple linear regression and artificial neural network retention prediction models for ginsenosides on a polyamine-bonded stationary phase in hydrophilic interaction chromatography[J]. Journal of Separation Science,2008,31(9):1550-1563.
[16] HO S L,XIE M,GOH T N. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction[J]. Computers & Industrial Engineering,2002,42(2-4):371-375.
[17] TIAN Y,PAN L. Predicting short-term traffic flow by long short-term memory recurrent neural network[C]//International conference on wavelet active media technology and information processing. Chengdu,China:IEEE,2014:231-244.
[18] ZHOU J,LU Y,DAI H,et al. Sentiment analysis of chinese microblog based on stacked bidirectional LSTM[J]. IEEE Access,2019,7:38856-38866.

(上接第 208 页)

approach for URL based DUST removal by knowledge engineering systems[C]//2019 3rd international conference on computing methodologies and communication (ICCMC). Erode,India:IEEE,2019:699-701.
[16] 黄承慧,印 鉴,侯 昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011,34(5):856-864.
[17] LI Xiangdong,ZHANG Cheng. Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method[C]//2013 IEEE 4th international conference on software engineering and service science. Beijing:IEEE,2013:267-270.
[18] 李对红,王裴岩,张桂平,等. 基于字簇的多模型中文分词方法研究[J]. 计算机应用研究,2020,37(2):355-359.
[19] ZHANG H,LIU Q,CHENG X,et al. Chinese lexical analysis using hierarchical hidden Markov model[C]//Proceedings of the second SIGHAN workshop on Chinese language processing. [s. l.]:Association for Computational Linguistics,2003:63-70.
[20] 韩冬煦,常宝宝. 中文分词模型的领域适应性方法[J]. 计算机学报,2015,38(2):272-281.
[21] AIZAWA A. An information theoretic perspective of TF-IDF measures[J]. Information Processing & Management,2003,39:45-65.
[22] 但唐朋,许天成,张姝涵. 基于改进 TF-IDF 特征的中文文本分类系统[J]. 计算机与数字工程,2020,48(3):556-560.