



基于神经网络和自回归模型的网络流量预测

熊皓^{1*}, 刘嘉勇¹, 王俊峰²

(1. 四川大学 网络空间安全学院, 成都 610065;

2. 四川大学 计算机学院, 成都 610065)

(*通信作者电子邮箱 xionghao961001@163.com)

摘要: 互联网的急速发展在给人带来了巨大便利的同时, 也使网络中的网络流量出现了爆炸性的增长, 预测网络流量对于网络的研究、管理和控制都具有很高的现实指导意义。为了减小网络流量数据的预测误差, 提出了一种基于神经网络和自回归模型的网络流量预测模型——卷积神经网络(CNN)-长短期记忆(LSTM)网络+自回归(AR)。通过卷积神经网络和长短期记忆网络结合来同时获取数据的短期局部依赖特征和长期发展趋势, 添加历史连接组件将网络流量的周期性考虑在预测中, 完成对网络流量中非线性项的处理, 利用自回归模型预测线性项, 将两部分结果结合得到最终预测值。实验结果表明, 对比传统的网络流量预测模型, 在最好情况下所提出模型的均方误差(MSE)、均方根误差(RMSE)和平均绝对误差(MAE)分别减少了1.5604、0.1468和0.1405, 这说明该模型有更好的预测表现, 预测值与实际值的差距更小。

关键词: 网络流量; 卷积神经网络; 长短期记忆网络; 自回归模型

中图分类号: TP393

文献标志码: A

Network traffic prediction based on neural network and autoregressive model

XIONG Hao^{1*}, LIU Jiayong¹, WANG Junfeng²

(1. College of Cybersecurity, Sichuan University, Chengdu Sichuan 610055, China;

2. College of Computer Science, Sichuan University, Chengdu Sichuan 610055, China)

Abstract: The rapid development of Internet not only brings great convenience to human beings, but also makes the network traffic in the network appear explosive growth. The prediction of network traffic is of great practical guiding significance for the research, management and control of the network. In order to reduce the network traffic data prediction error, a network traffic prediction model based on neural network and autoregressive model — Convolutional Neural Networks (CNN)-Long Short-Term Memory(LSTM) Network+AutoRegressive (AR) was proposed. By combining CNN and LSTM network, the short-term local dependence characteristics and long-term development trend of the data are obtained at the same time, add a historical connection component to consider the periodicity of network traffic into the forecast to complete the processing of non-qualitative items of network traffic, autoregression model is used to predict linear terms, and the final predicted value is obtained by combining the results of the two parts. The experimental result shows that, compared with the traditional network traffic prediction model, in the best case, the Mean Square Error (MSE), the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) of the proposed model is reduced by as much as 1.5604, 0.1468 and 0.1405, respectively, which indicates that the model has a better prediction performance and the difference between the predicted value and the actual value is smaller.

Keywords: network traffic; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM) network; AutoRegressive (AR) model

收稿日期: 2020-10-19; 修回日期: 2021-01-11; 录用日期: 2021-01-11。

基金项目: 四川省科技厅重点研发项目(2020YFG0374)。

作者简介: 熊皓(1996-), 男, 贵州贵阳人, 硕士研究生, 主要研究方向: 网络安全、深度学习; 刘嘉勇(1962-), 男, 四川成都人, 教授, 博士, 主要研究方向: 网络通信、网络安全、信息安全; 王俊峰(1976-), 男, 安徽芜湖人, 教授, 博士, 主要研究方向: 空间信息网络、软件安全、安全云计算。



0 引言

随着互联网的飞速发展,网络规模的不断增大,网络当中的流量数据也急速增加,并且还有持续增多的趋势。网络流量是反映网络状态的重要参数,对网络流量进行分析和准确的预测,可以有效管理网络,提高网络的利用率,也可以通过对异常网络流量的监控来完成入侵检测。对网络流量预测的研究对于合理分配网络资源、保证网络服务质量、设计网络结构等都具有重大的意义,网络流量分析的重要性也变得越来越

高。Sang 的研究^[1]证明了网络流量的可预测性,而网络流量的可预测性也给诸多领域带来了收益,比如网络安全、网络规划、动态带宽分配和拥塞控制等。而网络流量具有时效性、非线性性和随机性等特性,这使得对网络流量进行准确预测有很高的难度。近年来随着机器学习等技术的不断更新,也推动了网络流量预测的发展,研究人员针对网络流量数据的预测已经提出了大量模型,主要可分为线性时间序列模型、非线性时间序列模型和混合模型等。

线性时间序列模型有自回归(AutoRegressive,AR)模型和移动平均(Moving Average,MA)模型两种比较流行的子模型,它们可以组合成自回归移动平均模型(AutoRegressive Moving Average,ARMA)模型,并且它有很多的变体,线性时间序列是网络流量预测的传统方法。Sadek 等^[2]提出了基于 K 因子 ARMA 模型的时间序列模型,用于预测多尺度高速网络流量。Burney 等^[3]提出了基于小波滤波器的 SARIMA(Seasonal AutoRegressive Integrated Moving Average)方法,使用 Daubechies db4 小波滤波方法分解原始信号,来降低数据中的噪声。Hag 等^[4]提出了新的调整自回归综合移动平均(Adjusted AutoRegressive Integrated Moving Average,ARRIMA)模型用于网络建模和预测具有长距离依赖的互联网流量,AARIMA 模型给出的 H 参数值比 ARIMA(AutoRegressive Integrated Moving Average)模型给出的 H 参数值更准确。

非线性时间序列模型中有一个经典的广义自回归条件异方差模型 (Generalized AutoRegressive Conditional Heteroscedasticity,GARCH),另外神经网络也常用于对网络流量进行预测。Abdelouhab 等^[5]提出了基于多层感知(Multilayer Perceptron,MLP)的神经网络来分析和预测 IP(Internet Protocol)网络上的互联网流量。Chabaa 等^[6]提出了基于时间序列的自适应神经模糊推理系统(Adaptive Neuro-Fuzzy Inference System,ANFIS),用于建模和预测互联网流量。Wang Peng 等^[7]采用反向传播小波神经网络(BackPropagation Wavelet Neural Network,BPWNN)技术来改善反向传播神经网络(BackPropagation Neural Network,BPNN)技术在网络流量预测中的缺陷,在多步预测中取得更好效果。Hong Zhao^[8]提出了一种基于快速小波变换的最小均方差方法来预测自相似网络,在更低的计算复杂度下取得了更高的预测精度。

很多研究人员也考虑通过使用混合模型来提高网络流量预测的准确度。Bo Zhou 等^[9]结合使用了线性 ARIMA 和非线性 GARCH 时间序列模型,该模型主要用于预测长距离依赖和自相似性网络流量。Zeng 等^[10]提出了线性时间序列 ARIMA 和非线性时间序列多层人工神经网络(Multilayer Artificial Neural Network,MLANN)的混合模型来预测短期网络流量,混合模型的预测准确度高于单一模型。Li 等^[11]将小波和 ARIMA 模型与时间序列相结合进行互联网流量分析,利用小波技术将原始信号分解成不同的频率,用近似系数函数衡量小波方法的性能。

但现有的网络流量数据预测研究中存在的一个问题是,研究者们仅仅考虑到了网络流量中的时间特性,而忽视了区域间的空间相关性。另外网络流量中会有短期和长期重复模式存在,这会导致网络流量中通常会出现呈非线性关系的周期项数据和线性关系的趋势项数据,而很多研究中并没有将这部分内容考虑在内。将这些因素考虑在网络流量预测的处理过程中,将会有利于预测准确度的提高。

为了解决以上问题,本文提出了一种基于神经网络和自回归模型的网络流量数据预测模型,将网络流量数据预测任务分解为线性和非线性两个部分,使用卷积神经网络



(Convolutional Neural Network,CNN)和长短期记忆网络(Long Short-Term Memory,LSTM)结合来作为非线性组件,并添加历史数据连接组件处理网络流量时间序列中的周期项,而线性组件则采用自回归模型来实现,对网络流量数据中的趋势项进行预测,并融合两部分来完成对网络流量数据的预测。同时 CNN 可以发现数据的空间相关性,而 LSTM 的作用则是提取数据的长期依赖模式,使网络流量数据的预测更准确。

1 基于线性和非线性组件的混合预测模型

本文提出了一个卷积神经网络(Convolutional Neural Networks, CNN)-长短期记忆(Long Short-Term Memory,

LSTM)网络+自回归(AutoRegressive, AR)模型来完成对网络流量数据的预测,它是基于神经网络和自回归模型的混合模型,模型的结构如图 1 所示。所提出的模型由线性组件和非线性组件两部分共同组成,神经网络模型作为非线性组件,自回归模型则作为线性组件,分别各自处理网络流量数据中的线性和非线性部分。其中神经网络模型由卷积神经网络和长短期记忆网络共同构成,CNN 用于获取变量的短期局部依赖模式,解决当前网络流量预测模型中存在的忽视区域间的空间相关性的问题,LSTM 用于抓取数据的长期发展趋势,历史数据连接组件将以往数据加入预测过程中,提高网络流量预测的准确率,而自回归模型则用于处理网络流量的线性部分,同时增强该混合模型对输入规模发生变化的网络流量的鲁棒性。

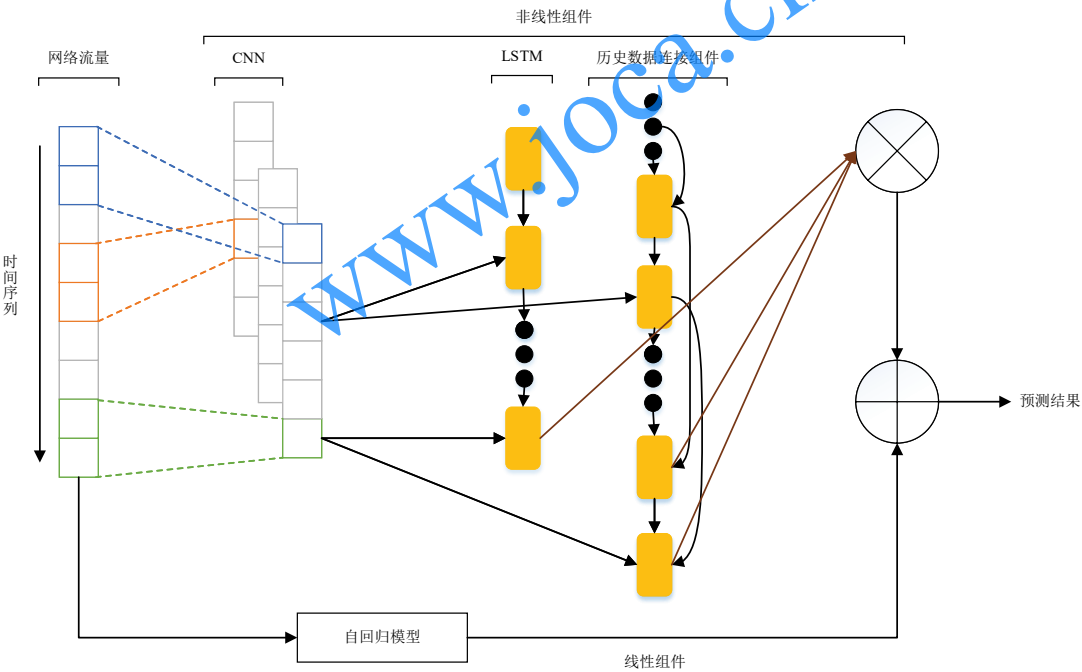


图 1 混合模型结构

真实的网络流量数据中存在一定的周期性,因此可以利用网络流量数据的这个特点来提高预测精度。例如当想要预测当天下午五点时的网络流量值时,一个经典的方法是除了将最近的数据记录考虑在内,还可以将以往历史时期下午五点的网络流量也应用到预测中。因此在模型的非线性部分中

还创新性的添加了一个历史数据连接组件,将当前的隐藏单元与相邻时间段中相同历史时期的隐藏单元连接起来,把历史时期中的网络流量数据考虑到预测过程中,能使网络流量预测更准确。

通过结合 CNN 和 LSTM 的优势来完成网络流量数据中短期和长期依赖模式的获取,同时考虑网络流量的时间特性



和空间相关性,加入历史连接组件以便在预测中引入历史数据,并使用自回归模型完成对数据中线性部分的处理,结合线性和非线性部分的输出得到最后的预测结果,使得该混合模型在网络流量预测任务上表现良好。

1.1 非线性组件

1.1.1 卷积神经网络

卷积神经网络是一种深度人工神经网络,它目前常被用于计算机视觉的工作中,主要组件有卷积层、池化层、全连接层等,卷积层和池化层交叉出现,并连接到一个或多个全连接层^[12]。卷积神经网络通过反向传播算法来自适应学习特征,它本身包括一个卷积计算的步骤,这个步骤也是 CNN 的核心部分,它的主要特征是权值共享,即在一个特征图谱上的神经元拥有相同的权值。

卷积层由一个或多个特征图谱组成,每一个特征图谱中包含多个神经元,不同特征图谱上的神经元通过卷积核连接起来。前一层的特征图谱和一个具备学习能力的卷积核完成卷积,然后经过一个激活函数处理,就获得了一个输出特征图谱,它的计算过程如式(1)所示:

$$y = f\left(\sum_{i \in G_m} x_i^{p-1} * w_{im}^p + b_m^p\right) \quad (1)$$

f 是激励函数, y_m^p 是第 p 层第 m 个特征图谱的输出, x_i^{p-1} 是第 $p-1$ 层第 i 个特征图谱的输出, G_m 是前一层的特征集合, w_{im}^p 是卷积核, b_m^p 是偏置项的值, $*$ 是卷积操作。

池化层的作用主要是降低特征维数,保持局部不变性,并在提取显著特征的同时减少模型的参数,这可以在一定程度上缓解过拟合现象的产生。在全连接层中,每个输入都通过一个可学习的权值连接到每个输出,这使得它可以学习到卷积层或池化层中的分类信息,并将学习到的所有局部特征整合为整体特征,并且全连接层的每个神经元都会在一个激励函数处理之后传递给输出层。

1.1.2 长短期记忆神经网络

原始的循环神经网络(Recurrent Neural Network,RNN)在训练中,由于训练时间过长以及网络层数过多,容易出现梯度爆炸或消失的现象,这使得它没法处理长序列的数据,从而无法提取长距离数据中的隐藏的信息,LSTM 就是针对 RNN 难以捕获长期依赖关系的缺点所提出的^[13]。图 2 为 LSTM 记忆单元的结构示意图,它主要包含了单元状态、遗忘门、输入门和输出门。

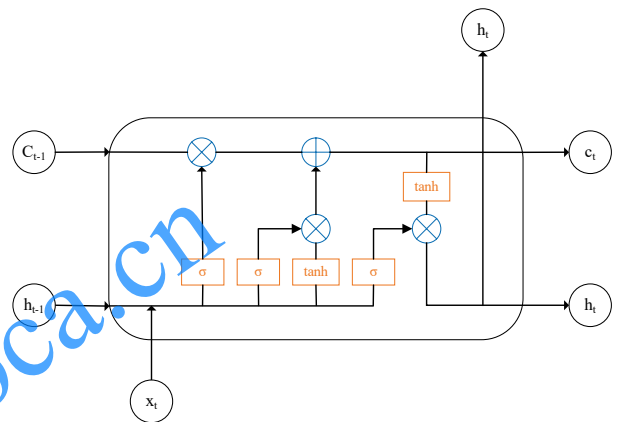


图 2 LSTM 记忆单元

在 t 时刻向前传播的有隐藏状态 h_t ,以及增加的一个单元状态 c_t ,它代表了 LSTM 的记忆,并且通过遗忘旧记忆和增加新记忆来完成自身的更新。

第一步是确定单元状态中要丢弃的信息,主要通过遗忘门 f_t 来实现这步操作,它按照一定的概率来选择遗忘掉上一层的单元状态与否, f_t 的计算过程可由下式(2)所表示:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

其中 W_f 是权值矩阵, b_f 是遗忘门偏置项的值, $[h_{t-1}, x_t]$ 表示把 h_{t-1} 和 x_t 连接成一个更长的向量。

第二步是修改单元状态中存放的信息,这一步通过两个步骤完成: 第一个步骤是使用 sigmoid 函数确定要更新哪部



分值,第二个步骤是通过 \tanh 函数创造一个新的单元值 \tilde{c}_t 并将其加入到状态中。

修改过程可表示成式(3)、(4):

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

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

第三步是更新单元状态,即将 c_{t-1} 更新为 c_t , c_t 就是新的候选值,它集合了通过遗忘门的旧记忆和输入门的新记忆。

更新过程如式(5)所示:

$$C_t = f_i * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

最后一步就是确定输出了,这个输出将会基于单元状态。首先经过一个 sigmoid 函数处理获得一个原始输出值,再把单元状态使用 \tanh 进行计算,把它缩放成一个-1 到 1 之间的值,再和前面的原始输出值逐对相乘,就得到了 LSTM 的输出。

输出结果为式(6)、(7):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

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

LSTM 神经网络中不但拥有 RNN 中的存在于各隐藏层单元中的外部循环,还包含了单元内部的自循环,这使得 LSTM 不仅对于时间序列数据历史信息的挖掘更完整,同时也更加完备的考虑了时间序列的长期依赖性^[14]。

1.2 线性组件

1.2.1 自回归模型

自回归模型常用在处理时间序列当中,它的基本思想是使用同一变量 x 的之前各时间点的值,来对当前时间点的值进行预测。自回归模型来源于回归分析里的线性回归,并对其进行了修改,在线性回归中是使用 x 预测 y ,而自回归模型中使用 x 来预测 x ,所以它被叫作自回归。

如果一个时间序列 $\{X_t\}$ 具有平稳、正态、零均值的特点,

并且在 t 时刻的值 X_t 能用它前 n 步的值 $X_{t-1}, X_{t-2}, X_{t-n}$ 来线性表示,则根据多元线性回归思想,AR 模型可定义为式(8):

$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + \xi_t \quad (8)$$

其中 p 和 ϕ_i 分别为阶数和系数,而 ξ_t 为白噪声序列,则可以

将自回归模型简单理解为 X 的当前值等于一个或多个过去值的线性组合再加上一个随机误差^[15]。

由于卷积和递归分量的非线性特性,神经网络模型的一个主要缺点是对输入数据的规模不敏感。而在特定的真实数据集中,输入信号的范围往往不断非周期地变化,这大大降低了神经网络模型的预测精度,而使用自回归模型可以很好地解决这个问题。

2 实验与结果分析

2.1 实验数据

本文所使用的网络流量数据来源于 MAWI Working Group Traffic Archive,采集了 samplepoint-F 上从 2020 年 7 月 20 日到 2020 年 8 月 19 日的网络流量数据来进行实验,数据的收集间隔为 10 分钟,每小时采集 6 个数据,一天采集 144 个数据。实验中共使用了 3447 条数据来完成模型的训练和验证,859 条数据被用于测试模型性能。实验中所使用的原始网络流量数据如图 3 所示。为了防止梯度消失,使网络加快收敛,使用最大最小值归一化来对数据进行预处理,归一化过程如式(9)所示:

$$\bar{x} = \frac{x - \min}{\max - \min} \quad (9)$$

其中 \max 和 \min 分别为样本数据的最大和最小值,经过归一化处理后原始数据的值都在[0,1]内。

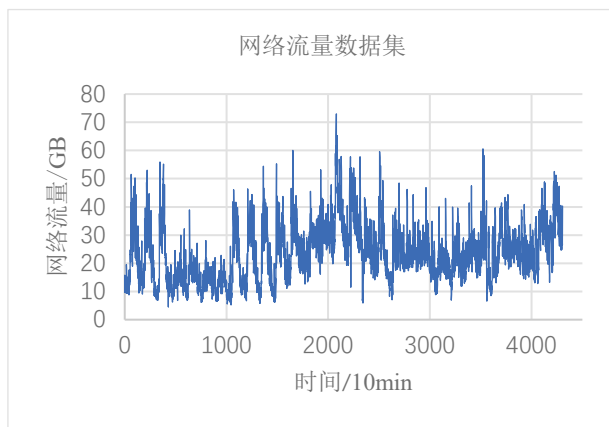


图3 实验中使用的网络流量数据集

2.2 实验模型

本文使用了三个模型与所提出的模型来进行性能比较,分别是两个经典模型 LSTM、门控循环单元(Gate Recurrent Unit,GRU)^[16]和 Lv 等提出的堆栈自动编码器(Stacked Autoencoder,SAE 模型^[17]),其中 SAE 模型使用了一个逻辑回归层来完成预测任务。使用所采集的网络流量数据来完成对 LSTM、GRU、SAE 模型以及所提出的 CNN-LSTM+AR 模型的训练和测试,并根据预测结果来比较它们的模型性能。

2.3 模型评价指标

本文使用预测模型性能评估中常用的均方误差(Mean-Square Error,MSE)、均方根误差(Root Mean Square Error,RMSE)和平均绝对误差(Mean Absolute Error,MAE)三个评价指标来评估所使用模型对时间序列数据的预测性能,MSE、RMSE 和 MAE 的值越小,预测值和实际值的差值就越小,预测误差就越低,MSE、RMSE 和 MAE 的计算公式为式(10)、(11)、(12):

$$RSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (10)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |(\hat{y}_i - y_i)| \quad (12)$$

其中 N 表示数据的数目, y_i 表示数据实际值, \hat{y}_i 表示模型预测值。

2.4 实验结果

表1给出了四个模型在所使用的网络流量数据集上的模型评价指标,从中可看出:在该数据集上,本文所提出的 CNN-LSTM+AR 模型具有更小的 MSE、RMSE 和 MAE 值,这表明所提出模型的预测值与实际值更加接近,预测性能优于其他三个模型。实验结果也表明在完成对真实的网络流量数据的预测任务时,与实验中所使用的几种网络流量预测模型相比,所提出的模型更加具有竞争力。

表1 预测结果

模型	MSE	RMSE	MAE
LSTM	29.0136	5.3864	4.1745
GRU	28.3957	5.3288	4.1142
SAE	28.4550	5.3343	4.1692
CNN-LSTM+AR	27.4532	5.2396	4.0340

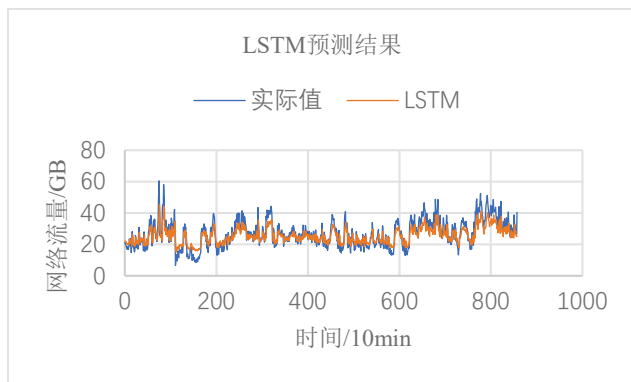


图4 LSTM 的网络流量预测结果

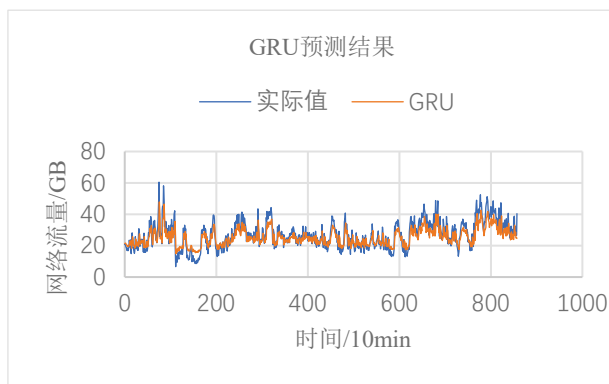


图5 GRU 的网络流量预测结果

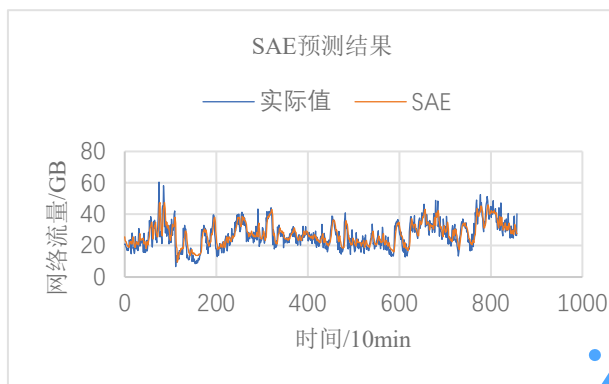


图6 SAE 的网络流量预测结果



图7 CNN-LSTM+AR 的网络流量预测结果

图 4-7 分别为 LSTM、GRU、SAE 和所提出的 CNN-LSTM+AR 模型对所使用的网络流量数据集的预测值以及实际数据值的对比图,从实验结果中可以得出,CNN-LSTM+AR 的模型性能优于 LSTM、GRU 和 SAE 模型,预测值与实际值的偏差较小,但仍然存在对于网络中变化尺度较大的突变流量的预测不够准确的问题。

3 结语

本文提出了一种基于 CNN、LSTM 和自回归模型的网络流量预测模型,该模型结合使用 CNN 和 LSTM 来提取变量之间的短期局部依赖模式,并且发现时间序列的长期变化趋势,同时引入新的历史数据连接组件将以往相关数据考虑在预测中,并使用自回归模型来增强神经网络的鲁棒性,使用线性组件和非线性组件相结合的方法来提高对网络流量数据的预测准确度。

实验结果表明,在同样的数据集上,与 LSTM、GRU 和 SAE 模型相比,CNN-LSTM+AR 模型在均方误差(MSE)、均方根误差(RMSE)和平均绝对误差(MAE)三个常用的预测准确率评价指标上,均有着不同程度的预测准确度提升,预测误差最好情况下分别降低了 1.5604、0.1468 和 0.1405,因此本文提出的模型表现更好,具有与实际值更接近的预测结果,且对于实际网络流量数据中出现的突变流量,预测表现好于其他三个模型。

参考文献

- [1] SANG A, LI S. A predictability analysis of network traffic[J]. Computer Networks, 2002, 39(4): 329-345.
- [2] SADEK N, KHOTANZAD A. Multi-scale high-speed network traffic prediction using k-factor Gegenbauer ARMA model[C]// Proceedings of the 2004 IEEE International Conference on Communications. Piscataway, IEEE, 2004, 4: 2148-2152.
- [3] SYED A R, BURNEY S M A, SAMI B. Forecasting network traffic load using wavelet filters and seasonal autoregressive moving average model[J]. International Journal of Computer and Electrical Engineering, 2010, 2(6): 1793-8163.
- [4] HAG H M, SHARIF S M. An adjusted ARIMA model for internet traffic[C]//AFRICON 2007. Piscataway, IEEE, 2007: 1-6.
- [5] CHABAA S, ZEROUAL A, ANTAR J. Identification and prediction of internet traffic using artificial neural networks[J]. Journal of Intelligent Learning Systems and Applications, 2010, 2(03): 147-155.
- [6] CHABAA S, ZEROUAL A, ANTAR J. ANFIS method for forecasting internet traffic time series[C]// Proceedings of the 2009 Mediterranean Microwave Symposium. Piscataway, IEEE, 2009: 1-4.
- [7] WANG P, LIU Y. Network traffic prediction based on improved BP wavelet neural network[C]// Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing. Piscataway, IEEE, 2008: 1-5.
- [8] ZHAO H. Multiscale analysis and prediction of network traffic[C]// Proceedings of the 2009 IEEE 28th International Performance Computing and Communications Conference. Piscataway, IEEE, 2009: 388-393.
- [9] ZHOU B, HE D, SUN Z. Traffic predictability based on ARIMA/GARCH model[C]// Proceedings of the 2006 2nd Conference on Next Generation Internet Design and Engineering. Piscataway, IEEE, 2006: 8 pp.-207.
- [10] ZENG D, XU J, GU J, et al. Short term traffic flow prediction using hybrid ARIMA and ANN models[C]// Proceedings of the 2008



Workshop on Power Electronics and Intelligent Transportation System. Piscataway, IEEE, 2008: 621-625.

[11] LI J, SHEN L, TONG Y. Prediction of network flow based on wavelet analysis and ARIMA model[C]// Proceedings of the 2009 International Conference on Wireless Networks and Information Systems. Piscataway, IEEE, 2009: 217-220.

[12] YAMASHITA R, NISHIO M, DO R K G, et al. Convolutional neural networks: an overview and application in radiology[J]. Insights into imaging, 2018, 9(4): 611-629.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[14] 刘承启,林振荣,黄文海. 基于 LSTM 的 WEB 服务响应时间大数据预测方法 [J]. 四川大学学报:自然科学版, 2019, 56: 71-77.

[15] 赵超,李东方,唐亚勇. 分位点门限自回归时间序列模型的贝叶斯方法 [J]. 四川大学学报: 自然科学版, 2016, 53: 748-752.

[16] CHO K, VAN B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL].[2014-09-03].<https://arxiv.org/pdf/1406.1078.pdf>

[17] LV Y, DUAN Y, KANG W, et al. Traffic flow prediction with big data: a deep learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 16(2): 865-873.