Computer Science and Application 计算机科学与应用, 2020, 10(10), 1834-1842
Published Online October 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1010193

Hans汉斯

基于ARIMA和LSTM的网络流量预测研究

孙远航

同济大学电子与信息工程学院，上海
Email: 475626064@qq.com

收稿日期：2020年10月6日；录用日期：2020年10月21日；发布日期：2020年10月28日

摘　要

网络流量预测是网络安全领域重要的研究方向之一，精准预测网络流量的趋势和峰值，并针对现有信息安全系统发现网络中可能存在的安全问题做出提前预警。随着各传感器的大量部署，系统已拥有大量可用数据，但是缺乏行之有效的分析方法，为此本文通过深度学习的方式对网络流量预测建立模型，提出一种基于LSTM神经网络的流量预测模型，并与ARIMA模型比较验证LSTM网络模型具有更好的性能，验证了该模型在网络流量预测中的适用性与更高的准确性。

关键词

LSTM，ARIMA，流量预测

Research on Network Traffic Forecast Based on ARIMA and LSTM

Yuanhang Sun

College of Electronics and Information Engineering, Tongji University, Shanghai
Email: 475626064@qq.com

Received: Oct. 6th, 2020; accepted: Oct. 21st, 2020; published: Oct. 28th, 2020

Abstract

Network traffic prediction is one of the important research directions in the field of network security. It accurately predicts the trend and peak value of network traffic, and provides early warning for existing information security systems that may find security problems in the network. With the large-scale deployment of various sensors, the system has a large amount of available data, but lacks effective analysis methods. For this reason, this paper establishes a model for network traffic prediction through deep learning, and proposes a traffic prediction model based on LSTM neural

文章引用: 孙远航. 基于 ARIMA 和 LSTM 的网络流量预测研究[J]. 计算机科学与应用, 2020, 10(10): 1834-1842.
DOI: 10.12677/csa.2020.1010193

network , and compared with the ARIMA model to verify that the LSTM network model has better performance, and verify the applicability and higher accuracy of the model in network traffic prediction.

Keywords

LSTM, ARIMA, Traffic Forecast

1. 引言

随着网络技术的高速发展和互联网技术的普及，网络已经成为人们工作、生活必不可缺的一部分[1]。传统的网络流量预测主要凭借主观经验，根据历史数据和个人经验进行流量的预测判断，该类方法误判的几率较大，不能够对流量变化趋势做出准确判断，预测结果较实际数据存在明显偏差。如果基于大量历史数据并且实时掌握目前网络状况，进行流量预测，其结果会更加可靠，因此提出了网络流量预测分析模型。

近年来，许多流量预测模型被提出，其中包括统计与回归方法[2]，基于流量的方法[3]和机器学习方法[4]。目前被广泛认可的自回归积分滑动平均模型(Autoregressive integrated moving average model，简称 ARIMA)，已于 20 世纪 70 年代应用在网络安全领域，目的是预测网络系统的状况，并且逐渐成为新的预测模型的比较对象。ARIMA 在数据规则变化时表现出良好性能，但是当数据不规则变化时，预测的误差较大。同时随着计算性能逐渐提高，传感器技术的成熟，能够感知采集大量数据，由此引发的“数据爆炸”的问题越来越受到重视，传统的方法由于维度的局限性难以适应现在预测需求，因此目前急需新的处理大数据的预测模型。

上述研究方法虽然能初步地解决流量预测的问题，但是没有考虑到内部深层细粒度特征。随着机器学习技术的发展，深度学习以其强大的特征提取能力，基于深度学习的预测模型逐渐成为新趋势。到目前为止，深度学习技术已经在自然语言处理、语音识别和计算机视觉领域取得显著成功，并向其他领域应用广泛扩展，其中时间序列分析为新的扩展方向之一。网络流量预测分析可以抽象成时间序列问题，目前循环神经网络模型(Recurrent neural network，简称 RNN)被广泛认可，但是根据以往研究表明，传统的 RNN 方法性能难以满足需求，对 10 分钟后的数据难以预测，为此本文提出将长短期记忆网络模型(Long short-term memory，简称 LSTM)用于网络流量的预测，该方法具有较高的准确性，与 ARIMA 模型对比性能提高近 70%。

2. 流量预测

流量预测作为网络安全的重要研究方向之一可以抽象为时间序列预测问题，时间序列预测虽然作为机器学习的重要领域之一，但与其它机器学习问题不同，关键区别在于固定的数据序列约束以及提供的约束条件。时间序列预测问题涉及许多时间分量，而这往往是问题解决的难点，并且不能简单地使用机器学习算法解决。时间序列预测为根据当前和过去采集到的数据，分析预测出未来几分钟到几个小时的可能状况。此类问题与网络流量预测问题、交通预测问题十分类似，两类都可以归结为时间序列预测的

研究问题，随着深度学习的高速发展，为两类问题的解决提供新的研究思路和方法。

人工神经网络(Artificial neural network，简称 ANN)自 20 世纪 80 年代以来被广泛应用于时间序列预测问题中，众所周知的 ANN 网络模型有以下几种：多层感知器(Multi-layer Perceptron，简称 MLP)、循环神经网络、径向基网络(Radial basis function network，简称 RBFN)，以及其他多种模型。ANN 出现之前，线性统计方法为预测的主流方法，但是该方法存在严重局限性，不能够捕获到复杂现实世界中的非线性关系[5]，同时 Makridakis 组织的一次大规模的预测竞赛的结果也表名，数据中存在不同程度上的非线性，不能通过线性统计方法处理。然而 ANN 作为广义非线性预测模型，能一定程度上捕获数据中的非线性关系，因此得到广泛的普及。随着科技进步，机器计算能力大幅度提升，Hinton 于 2006 年提出深度学习概念[6]，深度学习提供的高维度空间以及低维嵌套结构，有效解决了非线性降维方法无法解决的问题。Enzo Busseti 等[7]创新性的将深度学习应用于时间序列预测中，探索出不同模型预测精度不同的结论。

RNN 同样作为目前解决时间序列预测的主流方法之一，也可应用于网络流量预测问题中，由于 RNN 的提出，目前许多模型都是基于 RNN 的改进。但是 RNN 在面对长期时间序列表现出来的性能欠缺误差较大的问题，迫使我们寻找更好的方法，由此提出将 LSTM 模型应用在时间序列预测问题中。李高盛等人于 2019 年对 LSTM 模型进行相关研究[8]，LSTM 神经网络模型由三层构成，其中隐藏层由内存块组成，并且能够通过训练方法自动确定参数，与现有的方法相比无疑是一种创新点。

3. 相关工作

流量预测具备时空复杂性，未来流量的预测需要大量数据作为基础，但是因为设备流动性大的原因，数据的采集往往较为困难。本节将首先介绍数据集获取的方法，然后详述涉及的 RNN 和 LSTM 模型的相关知识。

3.1. 数据的获取

网络数据的获取是指以一定时间为周期，通过采集设备获取特定区域内的流量数，并进行处理和存储。由于设备流动性大的问题，采集的数据仅能在一定程度上反应此时的网络流量，但是已然具有参考价值。目前有多种方式获取网络流量数据，本数据集的获取方式为通过 WiFi 和移动网络信息获取实时流量。通过移动设备连入网络的数量评估区域内流量，该技术不需要昂贵的基础设施，成本低廉。

3.2. RNN 模型

传统神经网络相邻层之间的节点完全连接，而同一层的节点没有任何联系，由于节点之间需要相互交互，导致这种类型的网络模型在处理时间序列预测问题上性能欠佳。RNN 中的隐藏单元可以接收到先前状态对当前状态的反馈。图 1 展示了一个基本的 RNN 架构。

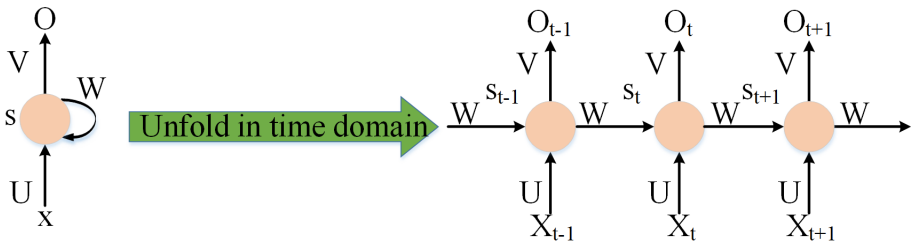

Figure 1. Structure of RNN
图 1. RNN 架构

在这种结构中，输入向量逐次送入 RNN 中，而不像传统神经网络一样采用固定数量的输入方式，同时能够根据实际情况确定 RNN 的深度。最终的输出结果不仅受当前输入值的影响，还受先前隐藏层输出结果的影响。以下用数学模型表示：

$$\begin{aligned} t_i &= W_{hx}x_i + W_{hh}x_{i-1} + b_h \\ h_i &= \sigma(t_i) \\ s_i &= W_{oh}h_i + b_y \\ \hat{o} &= g(s_i) \end{aligned} \tag{1}$$

公式中 x_i 表示输入变量，W_{hx}、W_{hh} 和 W_{oh} 为权重矩阵，b_h 和 b_y 是偏差向量，σ 和 g 为 Sigmoid 函数，t_i、h_i 和 s_i 为临时变量，$\hat{o}$ 为输出变量。损失函数如下：

$$f = \sum_i \left(\left\| \hat{o}_i - o_i \right\| / 2 \right) \tag{2}$$

其中 o_i 表示真实输出。$t+1$ 处的输出由 $t+1$ 处的输入和历史数据共同决定。由于梯度的问题，RNN 模型的精度随时间的推移下降，最终导致输出误差较大。

3.3. LSTM 模型

LSTM 模型在 RNN 模型的基础上进行改进，通过将隐藏层作为记忆单元，能够解决短期和长期时间序列的相关性问题。图 2 给出记忆单元的结构图，存储单元位于整个记忆单元的核心处，用红色圆圈表示。输入为已知数据，而输出则是预测的结果。在记忆单元中存在三个门，分别为输入门、遗忘门、输出门，通过绿色圆圈在图中标识。每个单元的状态用 S_t 表示，每个门的输入由预处理数据 X_t 和先前 S_{t-1} 状态组成。蓝色点代表汇聚点，虚线是前一状态函数。以记忆单元中信息的流动为基础，状态的更新和输出可以归纳为公式 3：

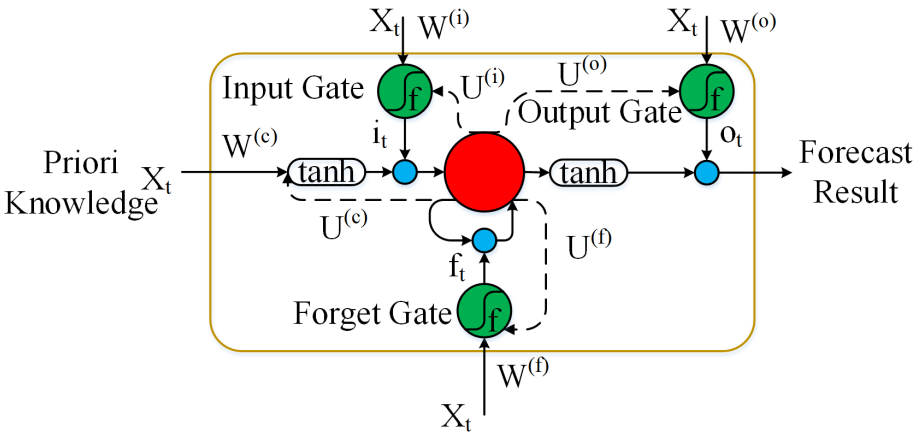

Figure 2. Design of the memory unit of LSTM
图 2. LSTM 记忆单元设计

$$\begin{aligned} i_t &= \sigma\left(W^{(i)}X_t + U^{(i)}S_{t-1}\right) \\ f_t &= \sigma\left(W^{(f)}X_t + U^{(f)}S_{t-1}\right) \\ o_t &= \sigma\left(W^{(o)}X_t + U^{(o)}S_{t-1}\right) \end{aligned}$$

$$\tilde{S}_t = \tanh\left(W^{(c)}X_t + U^{(c)}S_{t-1}\right) \tag{3}$$

$$S_t = f_t \circ S_{t-1} + i_t \circ \tilde{S}_t$$

$$O_t = o_t \circ \tanh\left(S_t\right)$$

公式中的∘表示 Hadamard 乘积，i_t、f_t和o_t表示三种不同的门，$\tilde{S}_t$为新状态的记忆单元，S_t为最终状态的记忆单元，O_t是最终输出的存储单元。$W^{(i)}$、$W^{(f)}$、$W^{(o)}$、$W^{(c)}$、$U^{(i)}$、$U^{(f)}$、$U^{(o)}$和$U^{(c)}$为系数矩阵。

经由不同功能门后，LSTM 记忆单元可以捕获短期和长期时间序列中复杂的相关特性，相比 RNN 模型性能明显提高。

4. 实验分析

4.1. 数据集分析

实验数据集来源于 2018 年某火车站整个候车室的设备流量信息，候车室提供 WiFi 接入点，每个数据采集点以一分钟为周期采集连入的设备量，通过该数据反应此时网络数量。整个数据集由 1000 组数据组成，时间从 2018 年 9 月 10 日 18 时 55 分开始到 2018 年 9 月 11 日 11 时 34 分结束，图 3 为数据集的折线图，反应出数据的变化趋势。从图中反应的网络流量信息中可以看出，设备数量从开始到 9 月 10 日 20 时 13 分达到峰值，随后数量逐渐减少，至 9 月 11 日 4 时 26 分到达谷底，后又呈逐渐升高的趋势，由此可以看出数据集不是固定的，在此进一步分析序列的自相关性如图 4 所示，图中表现了前 210 个点之间存在明显的正相关性，210 到 630 的数据存在负相关性，剩下数据相对而言相关性较弱。

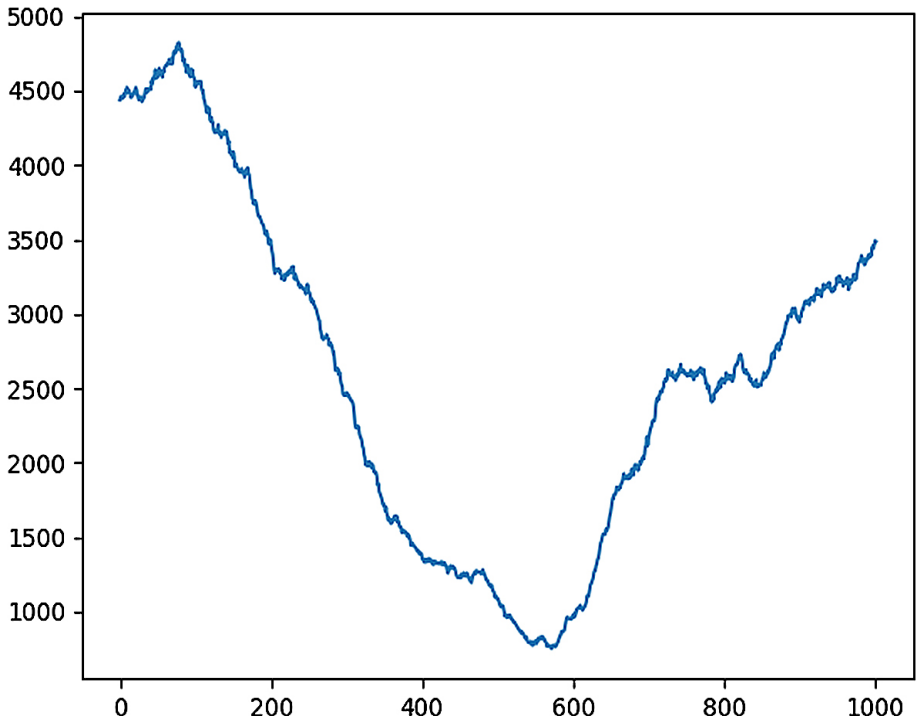

Figure 3. Line chart of dataset
图 3. 数据集折线图

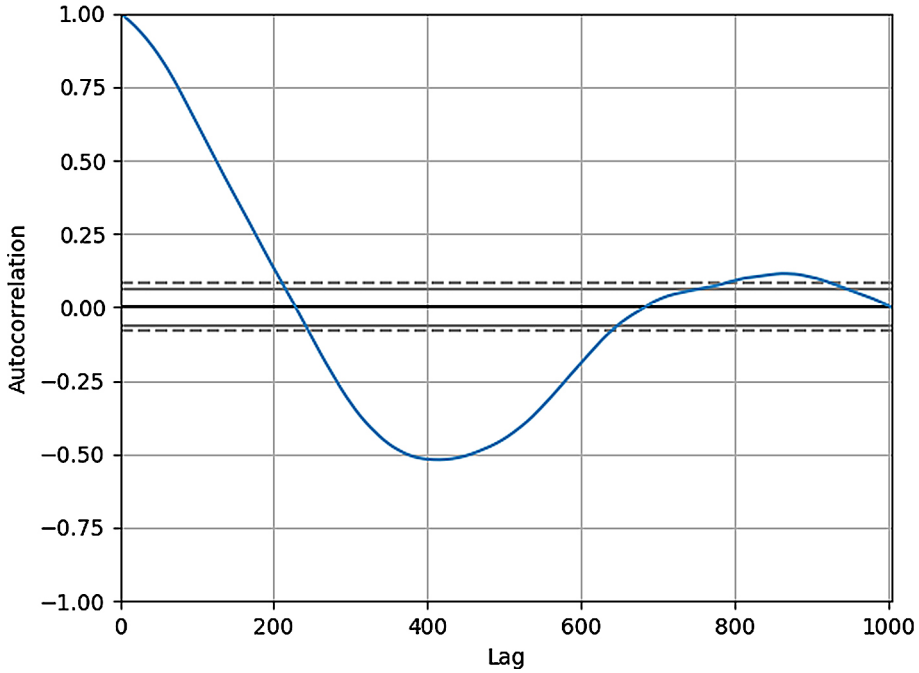

Figure 4. Autocorrelation plot of data set
图 4. 数据集自相关图

4.2. 评估标准

本实验采用均方误差(Mean squared error，简称 MSE)和均方根误差(Root mean square error，简称 RMSE)评估模型的精确度，预测模型越准确 MSE 和 RMSE 的值越小，因此许多预测模型会以 MSE 或 RMSE 的值作为衡量模型好坏的标准。

$$\mathrm{MSE} = \frac{1}{N}\sum_{t=1}^{N}\left(observed_t - predicted_t\right)^2$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(observed_t - predicted_t\right)^2} \tag{4}$$

4.3. 实验结果

本次实验的实验环境为：CPU 为 Intel core i5-9400F 2.90 GHz，操作系统 Windows10 16 位，内存为 16GBDDR。本次实验首先用 LSTM 模型在不同迭代次数参数下的结果，然后列举 ARIMA 模型在不同(p, q, d)下 MSE 的误差值，选取最优 LSTM 模型和最优 ARIMA 模型进行比较。

设置不同迭代次数，LSTM 预测出模型的准确度不同，若迭代次数少 LSTM 模型可能还未到平稳状况就被迫中止，相反若迭代次数多，不止程序运行时间过长，并且会影响最后模型的准确度。图 5 给出在 50、100、150、200、250、300、350、400 迭代次数下 LSTM 模型的 RMSE 误差盒式图，从盒式图中可以看出，当迭代次数为 150 时 RMSE 的平均值最小为 19.992，由此将该参数设定为 LSTM 模型的最优参数。

ARIMA 作为经典的时间序列预测模型，算法和参数调整方面已较为成熟，在此列举不同(p, q, d)参数下 ARIMA 模型 MSE 误差值，见表 1。表中可以看出，针对此数据集，ARIMA 参数在(2, 1, 1)情况下性能最优，因此选择参数(2, 1, 1)的 ARIMA 模型作为最优 ARIMA 模型。

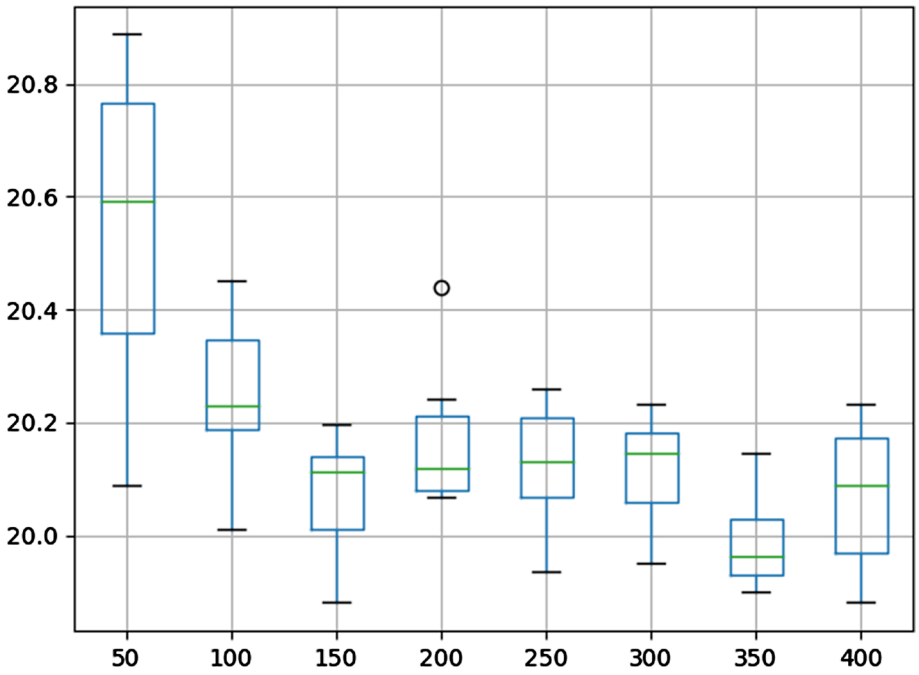

Figure 5. Box and whisker plot summarizing epoch results
图 5. 不同迭代次数误差值的盒式图

Table 1. The result of ARRIMA model
表 1. ARRIMA 模型运行结果

p	q	d	MSE
0	0	0	1,354,174.552
0	0	1	348,440.530
0	1	0	1357.026
0	1	1	1357.585
0	1	2	1369.117
0	2	0	2532.257
0	2	1	1319.250
1	0	0	1355.588
1	1	0	1363.889
1	2	0	2195.089
2	1	0	1381.750
2	1	1	1316.503
2	2	0	2041.721
4	1	0	1394.358
4	1	1	1324.233

上面分别介绍了 LSTM 最优模型的 RMSE 误差和 ARRIMA 最优模型的 MSE 误差，由于 RMSE 和 MSE 之间存在以下规则，

$$RMSE^2 = MSE \tag{5}$$

在此将 LSTM 模型取得的 RMSE 误差值转化成 MSE 误差值，结果为 399.640，由此可以算出 LSTM 模型相较 ARRIMA 模型，性能提高 69.64%。图 6 给出 LSTM 模型的预测曲线，蓝色为原始数据集，橙色为预测模型给出的预测数据集。

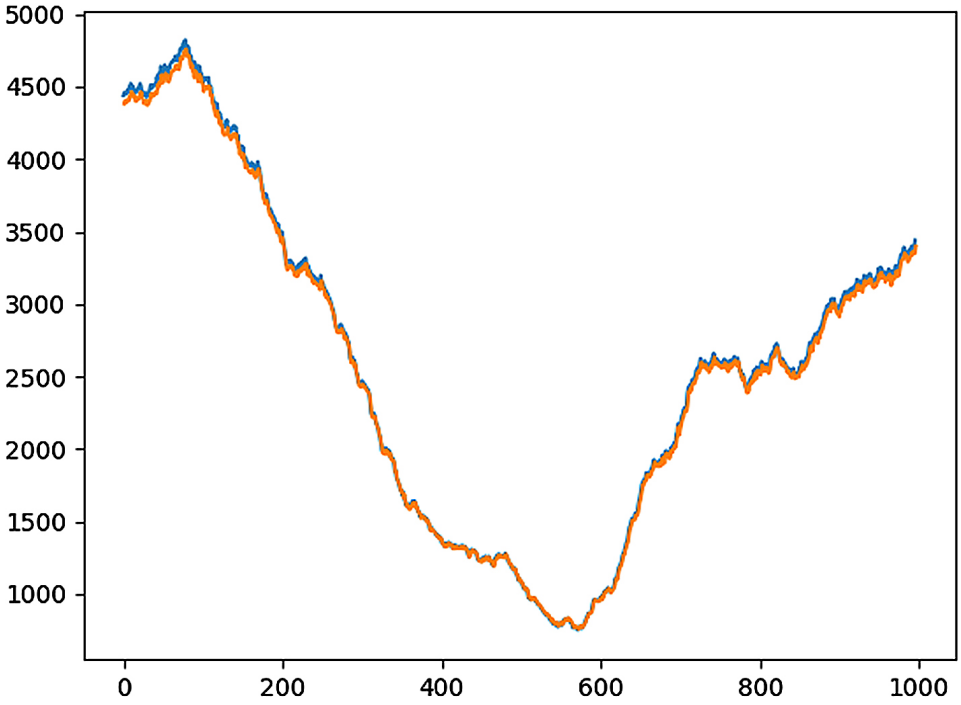

Figure 6. Line chart of LSTM model prediction
图 6. LSTM 模型预测图

5. 总结与展望

本文创新性地提出了一种网络流量预测模型，实验表明，方法可以很好地拟合数据集的流量曲线，验证出 LSTM 模型相比 ARIMA 模型性能提高明显，并能准确地预测出未来一段时间的变化趋势。网络流量预测模型的建立受各种因素影响，因此在未来工作中，作者将在更为广泛的数据集上应用，构建更加完善准确的网络流量预测模型。

参考文献

[1] 陈虹, 王闰婷, 肖成龙, 等. 基于 DBN-XGBDT 的入侵检测模型研究[J]. 计算机工程与用, 1-13[2020-06-07].

[2] Vlahogianni, E.I., Karlaftis, M.G. and Golias, J.C. (2014) Short-Term Traffic Forecasting: Where We Are and Where We're Going. *Transportation Research Part C: Emerging Technologies*, **43**, 3-19. https://doi.org/10.1016/j.trc.2014.01.005

[3] 李士宁, 闫焱, 覃征. 基于 FARIMA 模型的网络流量预测[J]. 计算机工程与应用, 2006(29): 148-150.

[4] Nagel, K. and Schreckenberg, M. (1992) A Cellular Automaton Model for Freeway Traffic. *Journal de Physique I*, **2**, 2221-2229. https://doi.org/10.1051/jp1:1992277

[5] Granger, C.W.J. and Terasvirta, T. (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford.

[6] Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Networks. *Science*, **313**,

504-507. https://doi.org/10.1126/science.1127647

[7] Busseti, E., Osband, I. and Wong, S. (2012) Deep Learning for Time Series Modeling. Standford University, USA.

[8] 李高盛, 彭玲, 李祥, 吴同. 基于 LSTM 的城市公交车站短时客流量预测研究[J]. 公路交通科技, 2019, 36(2): 128-135.