Hans Journal of Data Mining 数据挖掘, 2023, 13(1), 46-54
Published Online January 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.131005

Hans汉斯

基于双向LSTM模型的COVID-19预测分析

张秀子[1]，王佳英[1,2]，单　菁[1,2]

[1]沈阳建筑大学计算机学院，辽宁 沈阳
[2]沈阳工业大学软件学院，辽宁 沈阳

收稿日期：2022年12月3日；录用日期：2023年1月3日；发布日期：2023年1月11日

摘　要

2020年初，COVID-19疫情快速爆发，使得多个国家采取很多措施来控制疫情蔓延，疫情的出现对各国的医疗体系和经济造成了较大冲击，因此疫情信息的估计和预测对于政府和企业制定公共卫生防控措施具有重要的参考价值，对于这种快速传播的高致病性传染病来说，信息掌握的滞后会导致较为严重的后果。故提出了将ARIMA模型和LSTM模型等应用到疫情数据中，将无症状感染患者和症状感染患者统计为新增病例，依据2020年至2022年数据，预测短期内美国每日新增确诊病例数量，并提出了在模型中引入双向LSTM，以均方差误差(MSE)和平均绝对误差(MAE)来评价不同参数下模型预测精度，结果显示，提出的模型和参数获得的预测患病数量更接近实际患病数量，得到了较好的预测数据。

关键词

预测，COVID-19，时间序列，神经网络

Predictive Analysis of COVID-19 Based on Bidirectional LSTM Model

Xiuzi Zhang[1], Jiaying Wang[1,2], Jing Shan[1,2]

[1]School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang Liaoning
[2]School of Software, Shenyang University of Technology, Shenyang Liaoning

Received: Dec. 3rd, 2022; accepted: Jan. 3rd, 2023; published: Jan. 11th, 2023

Abstract

At the beginning of 2020, the rapid outbreak of the COVID-19 epidemic caused many countries to take many measures to control the spread of the epidemic. At the same time, the emergence of the epidemic had a great impact on the medical systems and economies of various countries. Therefore, the estimation and prediction of epidemic information has important reference value for the

文章引用: 张秀子, 王佳英, 单菁. 基于双向 LSTM 模型的 COVID-19 预测分析[J]. 数据挖掘, 2023, 13(1): 46-54.
DOI: 10.12677/hjdm.2023.131005

government and enterprises to formulate public health prevention and control measures. For this fast-spreading highly pathogenic infectious disease, the delay in information acquisition will lead to more serious consequences. Therefore, it is proposed to apply the ARIMA model and LSTM model to the epidemic data, and count asymptomatic infected patients and symptomatic infected patients as new cases. Based on the data from 2020 to 2022, we predict the number of new confirmed cases in the United States every day in the short term. A bidirectional LSTM was introduced into the model, and the mean square error (MSE) and mean absolute error (MAE) were used to evaluate the prediction accuracy of the model under different parameters. The results show that the predicted disease number obtained by the proposed model and parameters is closer to the actual disease number, and better prediction data are obtained.

Keywords

Forecast, COVID-19, Time Series, Neural Network

1. 介绍

2020 年以来，世界多地陆续出现新冠疫情爆发的情况，如美国、印度等多个国家和地区都受到了新冠疫情的影响[1] [2] [3]。截止到 2022 年 5 月 2 日，全球确诊病例较前一日增加 640,707 例，达到 513,581,249 例，死亡病例增加 2505 例，达到 625,602 例。

随着大数据时代的到来和人工智能领域的发展，出现了各种各样针对传染病预测的智能模型，这些模型被用来帮助人们了解传染病未来信息和为卫生决策提供了合理建议。如：2002 年出现的非典[4]，2014 年的埃博拉病毒，都对人类生产生活造成了严重影响。如果当时存在传染病预测的智能算法，可以在一定程度上帮助公共卫生机构能进一步做好准备和措施。因此，高效的传染病预测模型可以指导民众进行疾病的预防，降低疾病预防和控制成本以及协助部门在相关疾病的防控工作中提升效率。

估计确诊病例的数量也为了解疫情的发展提供了非常宝贵的意见，在预测方面，时间序列预测是传染病预测中一个比较常见的方法，许多研究人员运用时间序列对 COVID-19 进行预测[5]，通常是使用时间序列中差分自回归滑动平均模型(ARIMA) [6] [8]、传播动力学模型(SIR)等对未来感染人数情况进行预测，但这些方法还存在一定的局极限性。因此，本研究中讨论多个模型，应用了 ARIMA 模型以及基于 LSTM 的神经网络来预测短期内美国的每日新增确诊病例，提出引用研究双向长短期记忆神经网络(LSTM)，通过对比各个模型，发现通过双向 LSTM 预测得到的结果较为准确。

2. 相关模型

时间序列中存在许多模型，如 SIR、ARIMA、RNN、LSTM 等，SIR 模型是传染病领域较为常用且基础的模型，本研究中之所以不采用该模型，是因为该模型对人群的分类不够细致，模型中也没有反馈机制，进而对模型预测准确性有一定的影响。而 RNN 循环神经网络模型，虽然每个神经元的输出都可以在下一个时间段直接作用到自身，但是该模型会出现梯度消失的问题，且不能解决，故也会对模型预测有不小的影响。ARIMA 模型[7]是时间序列中较为典型的模型，常用于传染病预测，而 LSTM 模型是 RNN 模型的一种特殊形式，是一种较为先进的技术，目前也已成功应用于时间序列分析和传染病病毒预测方

面，因此本研究中选取 ARIMA 模型以及 LSTM 模型，并提出将双向 LSTM 加入到模型中，以提高模型的预测准确率。

2.1. ARIMA 模型

ARIMA 模型(p, d, q)，差分整合移动平均自回归模型，是时间序列预测方法之一。ARIMA 模型是包括自回归 AR(p)模型，移动平均 MA(q)模型和自回归-滑动平均混合模型 ARMA(p, q)模型[9] [11]，该模型适用的时间序列必须满足平稳且非白噪声，若不平稳则进行差分处理进行平稳检验直至平稳为止，差分的次数就是 ARIMA(p, d, q)的阶数 d，但并非差分次数越多越好，每一次的差分训练都会造成一定程度上的信息损失。模型表示为：

$$\Delta^d y_t = \theta_0 + \sum_{i=1}^{p} \varnothing_i \Delta^d y_{t-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-1} \tag{1}$$

式中，y_t 为原始时间序列，$\Delta^d y_t$ 表示 y_t 经 d 次差分后的平稳序列，ε_t 表示零均值的白噪声随机误差序列，$\varnothing_i\left(i=1,2,\cdots,p\right)$ 和 $\theta_j\left(i=1,2,\cdots,q\right)$ 为模型估计参数， p 和 q 为模型的阶。

ARIMA 模型的基本思想是：将预测对象随时间推移而形成层的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列，这个模型一旦被识别后就可以从时间序列的过去值以及现在值去预测未来值[10]。

ARIMA 模型步骤如图 1 所示，获取到序列后，需要进行平稳性检验，若平稳则继续，不平稳要进行差分运算，然后对序列进行白噪声检验，若为非白噪声开始拟合模型，获得模型的参数，进行模型检验，直至最后得到预测结果，以对得到的结果进行分析，以对得到的结果进行分析，ARIMA 模型的参数可由自相关函数(ACF)图和偏自相关(PACF)图估计得到。

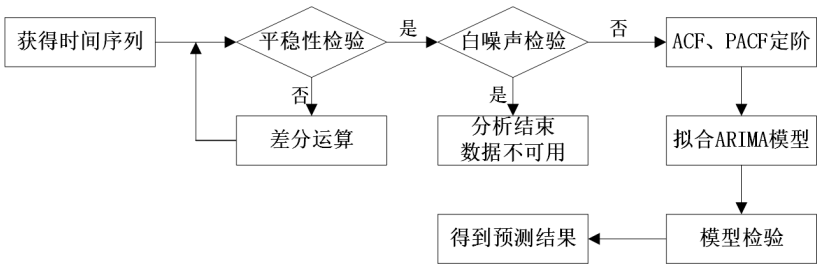


Figure 1. ARIMA model process
图 1. ARIMA 模型过程

2.2. 双向 LSTM 模型

LSTM 模型是一种特殊的 RNN [12]模型，能够学习长期的规律，在 1997 年被 Hochreiter & schmidhuber 提出，能够应用于各个领域，长短期记忆神经网络主要是可以解决长期依赖问题，它的默认行为是长时间记住实际信息，传统的神经网络中，模型不会关注上一时刻的处理会有什么信息可以用于下一时刻，每一次都只会关注当前时刻的处理。所有的递归神经网络都是由重复神经网络模块构成一条链，可以清晰的看出它的各个处理层，通常是一个单 tanh 层，通过当前输入及上一刻的输出来得到当前输出，但 LSTM 结构[11]不再单是一个单 tanh 层，它拥有四层结构，加了三个门状态(遗忘门，输入门，输出门)和一个记忆单元，通过门控单元对往数据进行筛选，可以有选择的决定信息是否通过，过滤干扰信息，减轻记忆负担。

而上述网络还存在一个问题，它是从以前的时间步中学习表示，可能存在一些不确定性，为了更好

的进行预测，理解前后关系，可以从将来的时间步中学习表示，因此双向 RNN 由此被提出来[13]，双向 RNN 的结构和连接如图 2 所示。

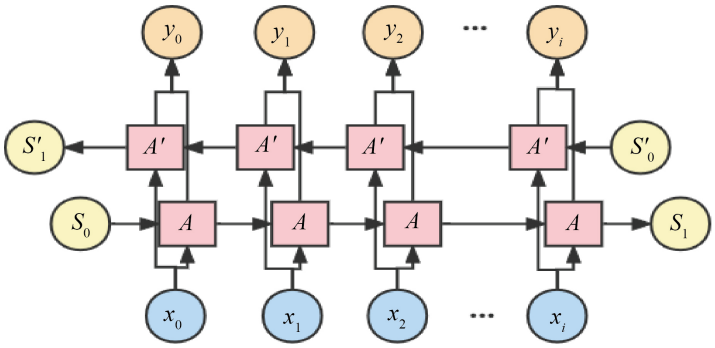


Figure 2. Bidirectional RNN model structure
图 2. 双向 RNN 模型结构

有两种类型的联系，一种是向前的时间联系，这有助于我们从以前的表述中学习，另一种是向后的时间联系，这有助于我们从未来的表述中学习。正向传播分为两个步骤:从左到右移动，从初始时间步长开始计算直到我们到达最后的时间步长，从右向左移动，从最后一个时间步长开始计算直到到达初始时间步长。因此本文中提出将 RNN 换成 LSTM，提出使用一层 LSTM + 双向 LSTM 模型结构进行预测，以达到更好的预测效果，如果是单一层的双向 LSTM 结构，模型在预测时，他对于前后文的关系上还有进步的空间，所以在此基础上多加一层 LSTM 网络。

3. 实验和结果

3.1. 实验评价标准

在对于时间序列预测中，均方差误差(MSE)、均方根误差(RMSE)和平均绝对误差(MAE)是在预测中常用指标。所用到的公式如下，其中 y_i 和 $\hat{y}_i$ 分别为第 i 个真实值和预测值，n 为测试数据的数量。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{2}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{3}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{4}$$

MSE 是真实值与预测值之差平方的期望值，它可以评价数据的变化程度，它的值越小，说明预测模型描述实验数据具有更好的精确度。RMSE 是均方误差的算术平方根，衡量预测值与真实值之间的偏差，常用来作为机器学习模型预测结果衡量的标准。MAE 是绝对误差的平均值，平均绝对误差能更好的反映预测值误差的实际情况，它的值越小，说明误差越低，同时训练的时候也不能过小，会造成过拟合现象。

3.2. 对美国地区疫情进行分析

3.2.1. 数据集

为了进行预测分析，本研究中使用的数据取自美国约翰·霍普金斯大学，从 2020 年 1 月 22 日至 2022 年 5 月 2 日，收集的数据集为 COVID-19 每日新增确诊病例数量，具体如图 3 所示。本研究中将总数据

分成两部分训练集和测试集，并将时间序列数据转换为包含输入和输出组件的结构，以便后续应用到模型中去。

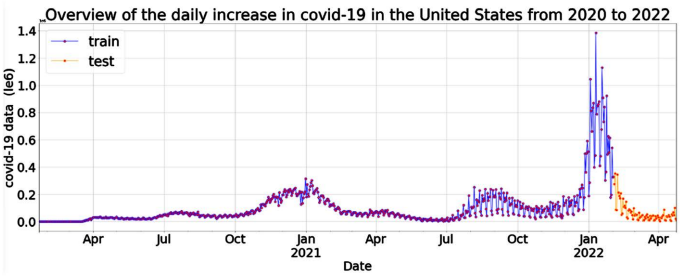


Figure 3. The daily increase in confirmed cases in the United States
图 3. 美国每日新增确诊病例情况

3.2.2. ARIMA 模型

A. 实验设置

使用美国地区的数据集经过平稳测试结果显示序列并不平稳，进而执行差分运算，当 p 值小于 0.01，序列平稳，故差分次数为 1。如图 4 所示，经过 ACF 和 PACF 得到 ARIMA 的参数(4, 6)、(4, 8)。

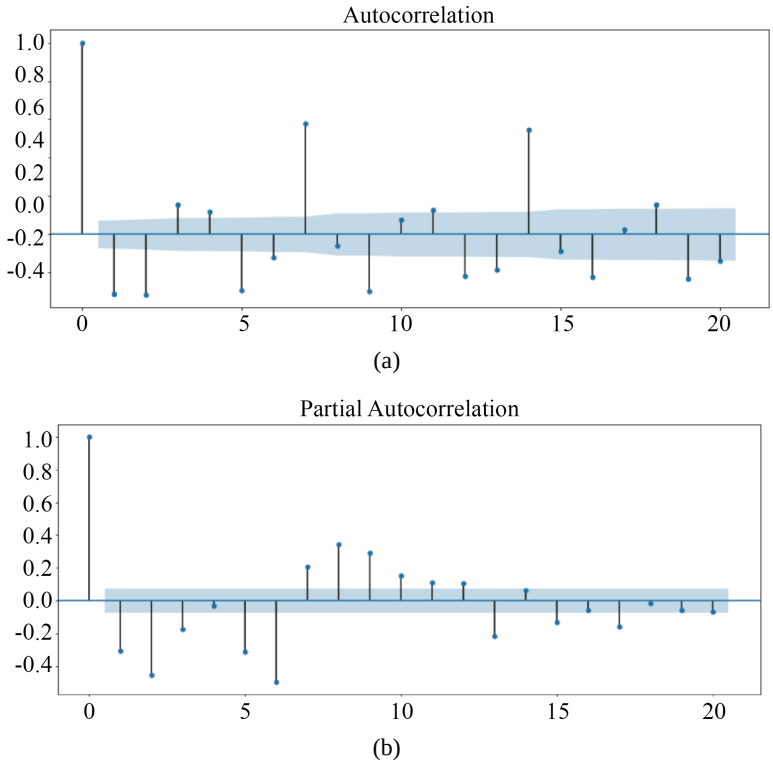


Figure 4. ARIMA model parameter. (a) ACF function parameter; (b) PACF function parameter
图 4. ARIMA 模型参数。(a) ACF 函数参数；(b) PACF 函数参数

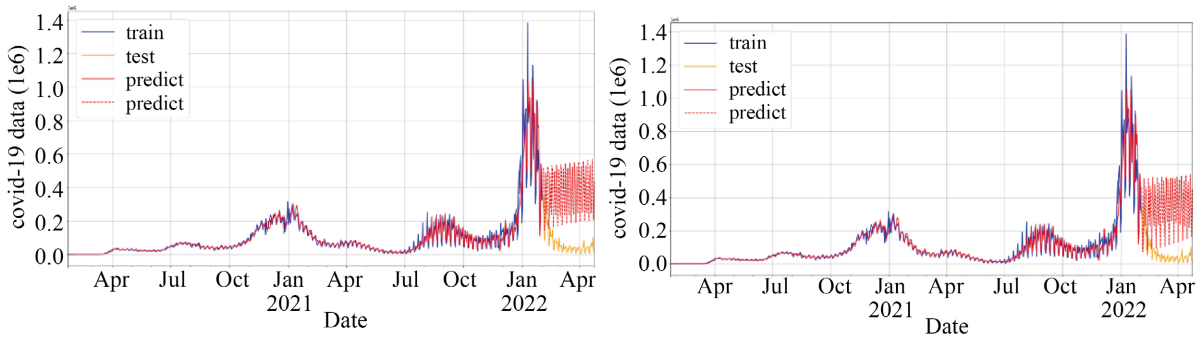
B. 实验结果

根据 ACF 和 PACF 图给出的参数建立模型，通过模型评估，部分对比如表 1 所示，可以看出参数(4,

1, 8)获得的均方误差值最小，因此该参数可作为模型的最佳参数，实际数据与预测数据之间部分对比如图 5 所示，可以看出，前半段模型预测较优，后半段预测值与真实值之间还存在一定差距，因此对于这种数据波动较大的数据集，ARIMA 模型的预测结果还有待提高。

Table 1. Prediction and evaluation using ARIMA Model
表 1. 使用 ARIMA 模型预测评价

model	MSE	RMSE	MAE
ARIMA(4, 1, 4)	2572650.1491	1593.4940	2447.1835
ARIMA(4, 1, 5)	2572650.1491	1603.9483	2599.8431
ARIMA(4, 1, 6)	1957310.6831	1399.0392	2143.7414
ARIMA(4, 1, 8)	1357596.1943	1165.1593	1783.4039
ARIMA(3, 1, 4)	2900335.3640	1902.1521	2976.1856



(a) ARIMA(4, 1, 6)　　(b) ARIMA(4, 1, 8)

Figure 5. Comparison of different parameters of ARIMA model
图 5. ARIMA 模型不同参数对比

3.2.3. LSTM 模型

A. 实验设置

本文中之所以选择长短期记忆网络，是因为它在普通的神经网络中增加了门的操作，可以有选择的对前面数据进行筛选以得出预测结果。LSTM 进行预测需要的是时序数据，根据前 timestep 步的长短来进行后面预测，即当 timestep 的数值设置为 n 时，模型根据前 n 个数据来预测后一个数据的值。模型训练之前，先对数据进行归一化，否则训练模型的损失函数下降的较慢，不利于模型训练。构建模型时，采用的 LSTM 是双向的 LSTM，可以考虑前后两个方向的数据，通过创建前向层和后向层串联计算下一预测状态。

本实验中 LSTM 模型的网络层数、网络节点参数等网络结果如图 6 所示，模型的损失函数图如图 7，本实验中选择的损失函数是均方误差损失。通过调整 timestep 数值和 epoch 次数，进行模型预测。

B. 实验结果

每轮预测有 n 个滞后观察的时间步长，以及不断变化的 epoch 次数，timestep 从 1 到 10，epoch 的数值从 100 到 500，以及改变神经网络层的参数数量也以获得不同的预测结果。通过多次实验测试，设置 timestep = 3，epoch = 500 时更合理，部分对比如表 2 所示，可以看出当第一层的神经网络参数为 64 时，第二层双层的 LSTM 参数为 32 时，epoch = 500 时，预测精度最优，实际数据与预测数据部分对比如图 8。

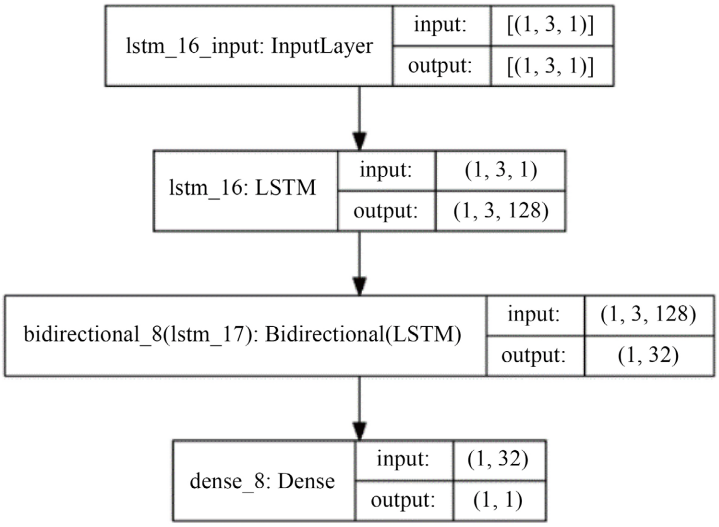


Figure 6. Model network structure
图 6. 模型网络结构

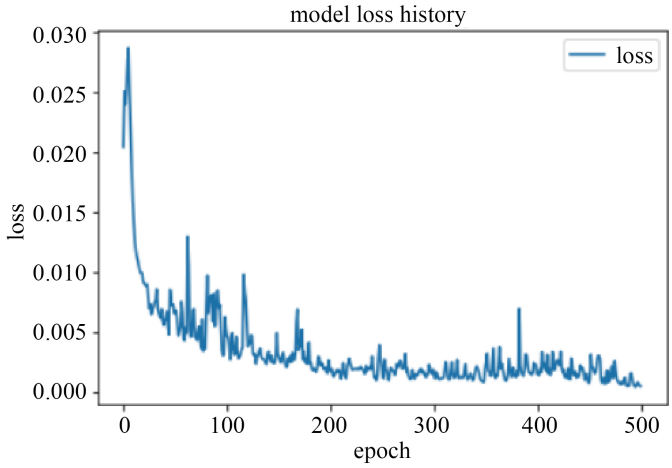
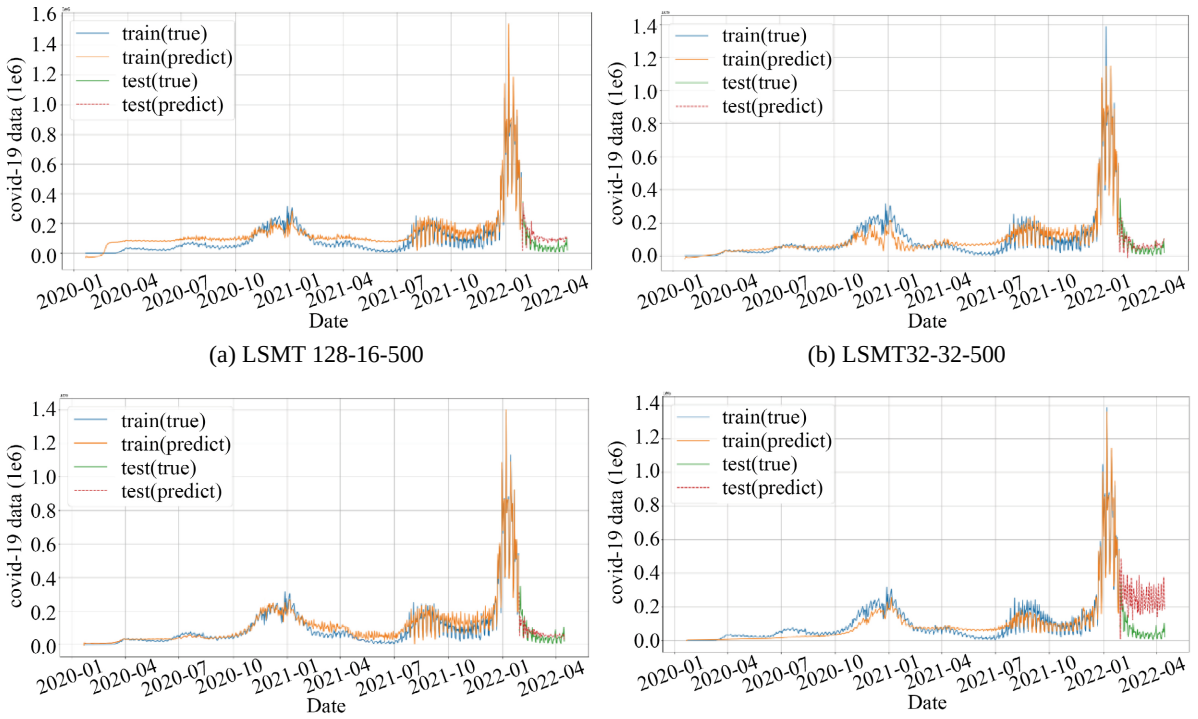


Figure 7. Model loss function
图 7. 模型损失函数

Table 2. Prediction accuracy of LSTM in each round
表 2. 每轮 LSTM 预测精度

timestep	epoch	LSTM(1)	LSTM(2)	MSE	RMSE	MAE
3	500	16	16	2695541.53	1641.81	2704.12
3	500	32	16	2277305.95	1509.07	2587.78
3	500	64	16	2088418.17	1445.14	2412.38
5	500	128	16	916710.50	957.45	1786.65
5	500	16	32	802135.18	895.62	1697.32
5	500	32	32	576506.12	759.28	1508.39
5	500	64	32	394534.73	628.12	1219.65
5	500	128	32	2532267.51	1591.31	2685.57



(a) LSMT 128-16-500

(b) LSMT32-32-500

(c) LSMT 64-32-500

(d) LSMT 128-32-500

Figure 8. Comparison between LSTM actual data and predicted data
图 8. LSTM 实际数据与预测数据对比

3.3. 实验结论

针对美国 COVID-19 数据，根据上述实验，与其他模型 SIR、RNN、单向 LSTM、双向 RNN 模型进行对比，本文选择的双向 LSTM 模型得到的预测数据与实际数据最为接近，且均方误差最少，模型选择较优的的预测精度如表 3。

Table 3. Comparison of prediction accuracy under different models
表 3. 不同模型下预测精度对比

model	MSE	RMSE	MAE
SIR	1217359.16	1103.34	2101.87
RNN	1155947.52	1075.15	1891.63
LSTM (单向)	660042.50	812.43	1252.53
RNN (双向)	531849.32	729.28	1002.45
ARIMA	1357596.1943	1165.1593	1783.4039
LSTM (双向)	394534.73	628.12	1219.65

4. 结论

本研究对 COVID-19 的传播数据进行建模，以预测疫情传播走势，疫情数据是使用了美国每日增加患者的数据，经过多个模型的对比，在短期预测中，对于这种增幅较大、波动较大的数据集，本文提出的将双向 LSTM 加入模型中，得到了较优的预测结果。

因此，将双向 LSTM 模型应用到疫情数据中，预测每日新增确诊病例数，可以为疫情的发展提供有价值的参考建议，以便采取更多措施来控制确诊人数的增长。而且当前许多实验表明，在 COVID-19 病毒感染后要多注意远离人群，及时隔离，而随着许多无症状感染者出现，本文研究的模型还需要进一步的优化。

参考文献

[1] Liu, X.X.., Fong, S.J., Dey, N., *et al*. (2021) A New SEAIRD Pandemic Prediction Model with Clinical and Epidemiological Data Analysis on COVID-19 Outbreak. *Applied Intelligence*, **51**, 4162-4198. https://doi.org/10.1007/s10489-020-01938-3

[2] Alizargar, J. (2020) Risk of Reactivation or Reinfection of Novelcoronavirus (COVID-19). *Journal of the Formosan Medical Association*, **119**, 1123. https://doi.org/10.1016/j.jfma.2020.04.013

[3] Hernandez-Matamoros, A., Fujita, H., Hayashi, T., *et al*. (2020) Forecasting of COVID19 per Regions Using ARIMA Models and Polynomial Functions. *Applied Soft Computing*, **96**, Article ID: 106610. https://doi.org/10.1016/j.asoc.2020.106610

[4] 喻国明, 靳一, 张洪忠, 等. 信息透明化处理的传播效果——SARS 事件中的民意调查及分析[J]. 新闻记者, 2003(7): 28-32.

[5] Bayyurt, L. and Bayyurt, B. (2020) Forecasting of COVID-19 Cases and Deaths Using ARIMA Models. *MedRxiv*. https://doi.org/10.1101/2020.04.17.20069237

[6] Rahman, M.R., Islam, A.H.M.H. and Islam, M.N. (2020) Geospatial Modelling on the Spread and Dynamics of 154 Day Outbreak of the Novel Coronavirus (COVID-19) Pandemic in Bangladesh towards Vulnerability Zoning and Management Approaches. *Modeling Earth Systems and Environment*, **7**, 2059-2087. https://doi.org/10.1007/s40808-020-00962-z

[7] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. and Lachhab, A. (2018) Forecasting of Demand Using ARIMA Model. *International Journal of Engineering Business Management*, **10**. https://doi.org/10.1177/1847979018808673

[8] Tuli, S., Tuli, S., Tuli, R. and Gill, S.S. (2020) Predicting the Growth and Trend of COVID-19 Pandemic Using Machine Learning and Cloud Computing. *Internet Things*, **11**, Article ID: 100222. https://doi.org/10.1016/j.iot.2020.100222

[9] Ku Kucharski, A.J., *et al*. (2020) Early Dynamics of Transmission and Control of COVID-19: A Mathematical Modelling Study. *Lancet Infectious Diseases*, **20**, 553-558. https://doi.org/10.1016/S1473-3099(20)30144-4

[10] Dey, S.K., Rahman, M.M., Siddiqi, U.R. and Howlader, A. (2020) Analyzing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis Approach. *Journal of Medical Virology*, **92**, 632-638. https://doi.org/10.1002/jmv.25743

[11] Tiwari, S., Kumar, S. and Guleria, K. (2020) Outbreak Trends of Coronavirus Disease-2019 in India: A Prediction. *Disaster Medicine and Public Health Preparedness*, **14**, e33-e38. https://doi.org/10.1017/dmp.2020.115

[12] Gatto, M., *et al*. (2020) Spread and Dynamics of the COVID-19 Epidemic in Italy: Effects of Emergency Containment Measures. *Proceedings of the National Academy of Sciences of the USA*, **117**, 10484-10491. https://doi.org/10.1073/pnas.2004978117

[13] Liu, F., Lu, Y. and Cai, M. (2020) A Hybrid Method with Adaptive Sub-Series Clustering and Attention-Based Stacked Residual LSTMs for Multivariate Time Series Forecasting. *IEEE Access*, **8**, 62423-62438. https://doi.org/10.1109/ACCESS.2020.2981506