运用 Dropout-LSTM 模型的新冠肺炎趋势预测

王 瑞，闫 方，逯 静*，杨文艺

(河南理工大学计算机科学与技术学院 河南 焦作 454000)

【摘要】为提高新冠肺炎 (COVID-19) 趋势预测精度，该文提出一种运用 Dropout 技术的长短期记忆 (LSTM) 神经网络预测新冠肺炎发展趋势的方法。该方法基于 Python 语言使用网络爬虫技术获取完整的国内新冠肺炎历史数据，提高数据采集效率的同时减少了主观原因导致的数据错误；因为新冠肺炎历史数据为时序性数据，为避免人为添加时间特征及充分挖掘较少时序数据之间的非线性关系，该文构建了层数更多的 LSTM 神经网络预测模型。随后在隐藏层中的非循环部分采用 Dropout 技术，对神经元进行随机概率失活，有效解决了深度学习的过拟合问题。最后用国内累计确诊、现有确诊和累计治愈人数对该方法进行验证，实验证明该方法可较精准预测新冠肺炎传播趋势。

关 键 词 新冠肺炎； Dropout 技术； 长短期记忆神经网络； 网络爬虫

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2020403

COVID-19 Trend Forecasting by Using Dropout - LSTM Model

WANG Rui, YAN Fang, LU Jing*, and YANG Wen-yi

(College of Computer Science and Technology, Henan Polytechnic University Jiaozuo Henan 454000)

Abstract To improve the accuracy of COVID-19 trend forecasting, a method of COVID-19 trend forecasting by using dropout and long short-term memory (LSTM) is proposed. The method uses web crawler based on python to obtain complete domestic historical data of COVID-19, which improves the efficiency of data collection and reduces data errors caused by subjective reasons. To avoid adding time features artificially and explore the nonlinear relationship fully between the less data of COVID-19, the proposed model extends the layers of the deep learning network. Then, the dropout technique is applied to the non-circular part of each hidden layer to randomly deactivate neurons, preventing the neural network from overfitting. The experiment demonstrates that the method can predict the number of cumulative confirmed cases, current confirmed cases and recovered cases.

Key words COVID-19; dropout technique; long short-term memory; web crawler

疫情在全球多点爆发并蔓延，已成为全球性流行病，给各国人民的健康带来巨大威胁。通过各地一系列防控措施，如勤洗手、公众场合戴口罩、外出需出示健康码和通行大数据等，我国疫情得到较好的控制，但是国外疫情仍然比较严峻，因此利用新冠肺炎历史数据预测疫情发展趋势对制定合理的干预防控措施有重要意义。

文献 [1-3] 从多个角度为抗击疫情提供了有力的学术支持。现有方法可分为统计学方法、动力学方法和机器学习方法。统计学方法适合在信息不完整的情况下使用，该方法通过部分样本的情况预测总体趋势，而部分样本与总体传播趋势具有较大差异性[4]，因此该方法的预测误差较大，无法准确体现疫情传播的趋势变化。动力学方法的经典数学模型为 SIR(susceptible infected recovered) 模型[5-6] 和 SEIR(susceptible-exposed-infected-removed) 模型[7-8]。动力学方法对疫情早期传播趋势有较好的预测，但是无法对开放式流动环境下的病毒传播做出准确估计，也无法使假设的疾病传播能力及治愈概率的常数与实际状况相符，因此无法对疫情趋势做长期准确的分析[9]。

随着新冠肺炎数据增多，机器学习展现出了极大的优越性。文献 [10] 在有限的数据下，通过最小二乘准则和梯度下降算法对数据进行非线性

收稿日期：2020－11－15；修回日期：2021－03－15
作者简介：王瑞 (1977－)，男，副教授，主要从事智能信息处理方面的研究.
通信作者：逯静，E-mail：lujing@hpu.edu.cn

回归来预测新冠肺炎确诊人数趋势，但此方法需要人为添加时间特征保证预测准确度。为解决上述问题，文献 [11-13] 运用自回归积分滑动平均模型 (autoregressive integrated moving average model, ARIMA) 对新冠肺炎疫情发展状况进行了预测，此模型对数据的时序性要求高，不需要人为添加时间特征，但是非线性拟合能力不强，随着数据量增加，预测效果下降。为解决上述问题，文献 [14] 建立了基于深度学习的长短期记忆模型 (long short-term memory, LSTM)，通过 Python 实现了模型的拟合和预测，此方法在一定程度上提高了短期新冠肺炎预测的准确度，但是有以下不足：1) 未做到对新冠肺炎发展趋势较为准确的长期预测；2) LSTM 神经网络对数据量需求较大，文中未提供一个简单的数据采集方法；3) 未考虑深层神经网络中由于参数多和模型复杂而带来的过拟合问题。

LSTM 神经网络在趋势预测中有许多改进机制，大致分为两类：1) 针对 LSTM 神经网络自身进行改进；2) 引入其他方法改进 LSTM 算法。文献 [15] 针对 LSTM 神经网络自身改进，网络训练时，将输出反馈回输入端，使其二次训练达到提高泛化能力的目的。文献 [16] 利用卷积神经网络 (convolutional neural network, CNN) 提取挖掘相邻路口交通流量的空间关联性，通过 LSTM 模型挖掘交通流量的时序特征，将提取的时空特征进行特征融合，实现短期流量预测。文献 [17] 引入集合经验模态分解 (ensemble empirical mode decomposition, EEMD)，构建多层级 LSTM 预测模型提升模型预测准确率。本文提出的 Dropout-LSTM 模型属于第二种改进方法。新冠肺炎历史数据为时序性数据，而 LSTM 神经网络擅长处理时序性数据，可根据实验效果调整具体预测的天数，达到对新冠肺炎发展趋势预测的目的。

由于新冠肺炎历史数据较多，传统的人工搜集方法不再适用，因此本文使用网络爬虫技术从腾讯新闻网站中获取相关数据供本次研究使用。考虑到新冠肺炎历史数据之间有很强的时序性，故本文使用擅长处理时序性数据的 LSTM 神经网络作为基本模型对疫情趋势进行预测。该模型与 ARIMA 模型相比有更强的非线性拟合能力，且不需要人为添加时间特征，可最大限度地挖掘时序数据之间的非线性关系。针对 LSTM 神经网络，新冠肺炎数据量较小，为避免在多层网络训练时出现过拟合问题，本文构建多层 LSTM 神经网络，并引入 Dropout 技术按随机概率让神经元失活。最后使用国内累计确诊、现有确诊和累计治愈人数验证此方法的准确性。

1　新冠肺炎历史数据获取

针对 2020 年 1 月 13 日−2020 年 9 月 12 日的 244 条数据，传统的人工搜集方法效率低且容易由于主观原因导致收集的数据与真实值不符。为避免以上问题，本文采用基于 Python 语言的网络爬虫技术[18] 获取新冠肺炎历史数据，并将其保存至 CSV 文件，以供后期实验使用。具体流程如图 1 所示。

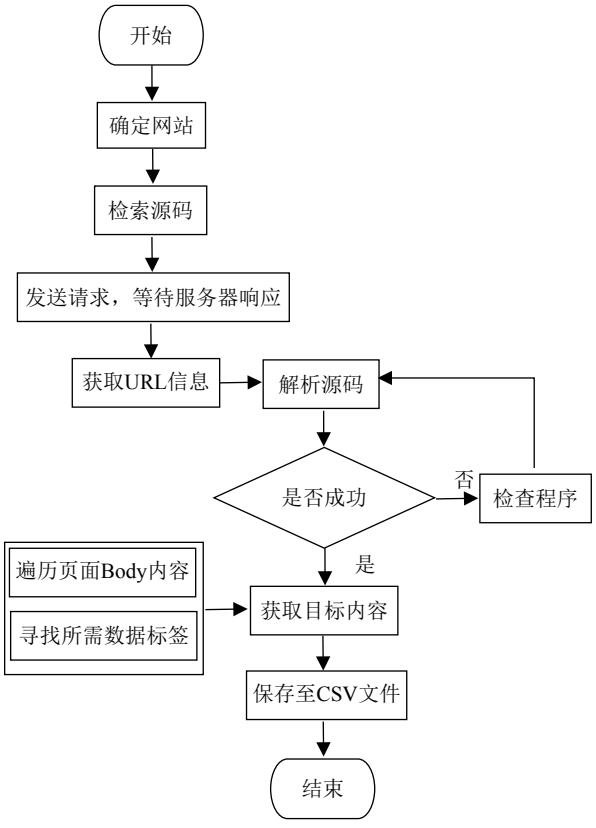


图 1　获取新冠肺炎历史数据流程图

具体步骤为：

1) 确定需要爬取数据的网址：https://view.inews.qq.com/g2/getOnsInfo?name=disease_other，在网页空白处，点击 F12 查看目标网站源码，找到数据接口；

2) 向服务器发送请求，等待响应。若响应成功，可获得 URL 的信息，并保存响应结果，以便后续对数据处理；若响应失败，需检查程序，重新

发送请求；

3) 对源码进行解析，若解析成功，在遍历源码 Body 的基础上寻找目标内容的标签，如：日期、累计确诊、新增确诊和累计治愈等，对目标内容进行获取并保存至 CSV 文件；若解析失败，则需检查程序，重新解析源码。

2 LSTM 模型

每日更新的新冠肺炎历史数据属于时序性数据，根据神经网络的特征，本文选择擅长处理时序性数据的 LSTM 神经网络作为基本模型对新冠肺炎发展趋势进行预测，可最大限度地挖掘数据时序性与非线性之间的关系。

2.1 LSTM 模型简介

LSTM 神经网络[19] 属于循环神经网络 (recurrent neural networks, RNN)。RNN 在处理序列信息中有良好的性能，但是当历史数据和预测数据的位置间隔不断增大时，梯度越传越弱，上一层的网络权重无法更新，丧失从历史数据中学习信息的能力，即梯度消失问题。

LSTM 神经网络能够解决 RNN 的梯度消失问题，是因为该网络中放置了遗忘门 f_t、输入门 i_t 和输出门 o_t[20]。可根据式 (1)～(6) 判断该数据是否符合算法认证，符合认证的数据留下，不符合的数据则通过遗忘门遗忘，因此 LSTM 神经网络能在更长的时序数据中有更好的表现[21]。LSTM 神经网络标准结构如图 2 所示。

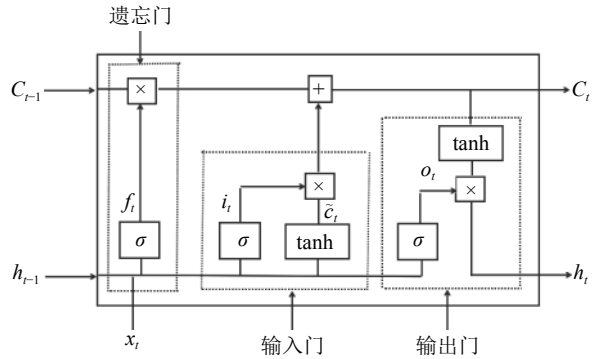


图 2 LSTM 网络结构

2.2 LSTM 记忆单元更新

1) 遗忘门

遗忘门负责对t时刻的输入x_t选择性忘记，通过计算f_t控制上一单元候选状态C_{t-1}中的数据保留或忘记。若f_t为 1，表示保留上一单元的数据；若f_t为 0，表示忘记上一单元的数据：

$$f_t = \sigma(\boldsymbol{W}_f[h_{t-1}, x_t] + b_f) \tag{1}$$

式中，σ为 sigmoid 激活函数；$\boldsymbol{W}_f$为上一单元隐藏层的输出h_{t-1}和当前输入数据x_t相乘的权重矩阵；b_f为遗忘门的偏置。

2) 输入门

输入门i_t负责对t时刻的输入有选择性地记忆。

当前输入内容由式 (2) 得到，σ用来控制需要更新的输入值，而输入的门控信号由式 (3) 控制，tanh 用来控制当前记忆单元候选状态：

$$i_t = \sigma(\boldsymbol{W}_{it}[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(\boldsymbol{W}_c[h_{t-1}, x_t] + b_c) \tag{3}$$

式中，$\boldsymbol{W}_{it}$为输入门i_t在t时刻的输入x_t的权重矩阵；$\boldsymbol{W}_c$为新生成信息在t时刻的权重矩阵；b_i、b_c为输入数据和新生成信息在当前单元的偏置。

传递给$t+1$时刻的状态C_t由两部分组成：式 (1) 得到的遗忘门f_t的输出与上一时刻$t-1$候选状态C_{t-1}之积；式 (2) 得到的输入门i_t的输出与t时刻候选状态$\tilde{C}_t$之积。此过程旨在抛弃当前单元中无用信息，保留有用信息：

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{4}$$

3) 输出门

输出门o_t通过 sigmoid 激活函数控制得到初始输出，接着使用 tanh 层将传递给下一时刻$t+1$的状态C_t缩放至 (−1,1) 上，最后与初始输出相乘得到最终输出结果：

$$o_t = \sigma(\boldsymbol{W}_{ot}[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \tanh(C_t) \tag{6}$$

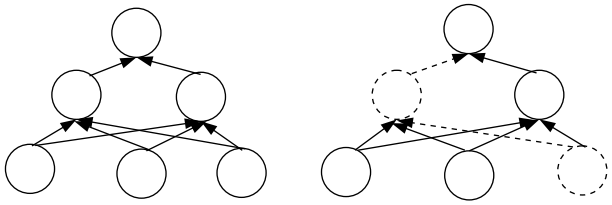
式中，$\boldsymbol{W}_{ot}$为输出门o_t的权重矩阵；b_o为偏置。

3 Dropout 技术

新冠肺炎历史数据量的大小是相对的，即对传统的人工搜集方法来说，新冠肺炎历史数据量较大，搜集较为麻烦；但是对于网络规模较大的 LSTM 神经网络来说，新冠肺炎历史数据量较小，较小的数据量输入较复杂的 LSTM 神经网络容易出现过拟合现象，即在训练集上精确率很高，在测试集中精确率却较低。在数据没有过拟合的前提下，调参很容易实现更高程度的拟合，但是会出现过拟合问题，无论如何调参，测试集的准确度依然不高。过拟合问题严重影响了模型的预测精度，为解决模型在训练时出现的过拟合问

题，提高模型的训练精确度，本文引入 Dropout 技术。

Dropout 是一种防止过拟合技术，分为权重 Dropout 和神经元 Dropout。权重 Dropout 选择神经层权重矩阵中的部分权重使之失活，而神经元 Dropout 则是选择神经层中部分神经元使之失活[22]。图 3 为神经网络运用 Dropout 技术前后对比，图 3a 为标准神经网络，图 3b 为运用 Dropout 技术后的神经网络。



a. 标准神经网络　　b. 运用Dropout技术后

图 3　运用 Dropout 技术前后对比

本文选择神经元 Dropout 技术按照概率把部分神经元的激活值设置为 0，随机使其失活，虚线部分表示失活后的神经元无法参与网络训练，这样可减弱神经元节点间的联结，提高模型泛化能力，减轻过拟合问题。

Dropout 技术在测试集上不需要使用。因为在测试阶段并不期望输出结果是随机的，若在测试阶段使用 Dropout 技术可能会导致预测值产生随机变化 (因为 Dropout 使节点随机失活)，预测值会受到干扰。

4　Dropout-LSTM 模型

4.1　Dropout-LSTM 模型微观层面

如图 4 所示，x为 LSTM 神经网络输入信息，y为输出信息，每个矩形代表一个 LSTM 神经元。Dropout 技术应用于 LSTM 神经网络，必须置放于网络的非循环部分，否则信息会随着循环丢失。这是因为：若把 Dropout 技术设置在隐藏状态上，即图中实线部分，每经一次循环，剩余信息便会以概率 P 丢失，也就是对式 (6)h_t随机进行置 0 操作。若序列较长，循环次数较多，到最后信息早已丢失。把 Dropout 技术设置在输入信息x上，那么 Dropout 技术造成的信息丢失与循环次数无关，只与网络层数相关。

同一层神经元不同时刻之间的信息传递不使用 Dropout 技术，而在同一时刻层与层之间的神经元传递信息时使用，将其神经元按照一定概率 P 随机置 0，使其失活。如图中粗线部分所示：$t-2$时刻的输入x_{t-2}先传入第一层 LSTM 神经网络，此过程使用 Dropout 技术，信息从第一层的$t-2$时刻传到t时刻不进行 Dropout 技术；接着从第一层的t时刻向第二层神经网络传递信息时使用 Dropout 技术。

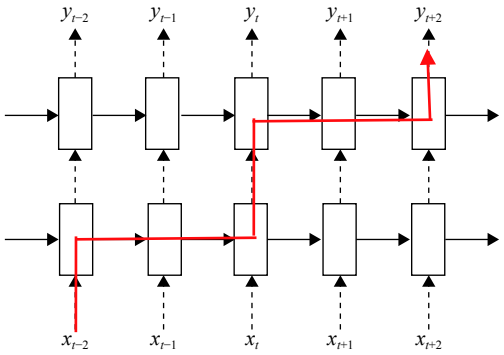


图 4　Dropout-LSTM 模型微观图

4.2　Dropout-LSTM 模型宏观层面

LSTM 神经网络实施 Dropout 技术宏观结构如图 5 所示，$x_0, x_1, \cdots, x_{243}$为输入数据。虚线圆表示采用 Dropout 技术后随机失活的神经元，虚线箭头表示失活后的神经元不传递信息。

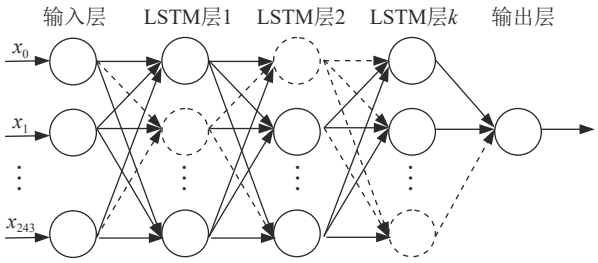


图 5　Dropout-LSTM 模型宏观图

加入 Dropout 技术的 LSTM 神经网络具体工作步骤为：

1) 设定随机概率 P，使网络中的每一层神经元按照随机概率 P 失活，但是不删除这些失活的神经元，停止工作的神经元不参与网络训练中的正向传播，输入神经元和输出神经元保持不变；

2) 将信息x输入 Dropout-LSTM 模型进行从输入层到输出层的正向传播，计算并存储神经网络的中间变量；沿着从输出层到输入层的顺序根据损失函数进行反向传播，计算并存储神经网络的中间变量和梯度参数；在未失活的神经元上更新相关参数；

3) 恢复失活神经元，此时刚恢复活性的神经元保持原样，而未失活的神经元参数经过上一轮过程已更新，重复步骤 1) ～2)。

通过随机忽略隐藏层神经元，可避免 LSTM 神经网络过度依赖某些局部特征，一定程度上降低了迭代过程中的过拟合现象。

5 实验过程

5.1 实验环境

在 Windows8.1 系统中使用 Pycharm，Python3.6 为实验平台，运用 Tensorflow 深度学习框架所提供的 LSTM 神经网络用于仿真实验。

5.2 实验数据及其预处理

5.2.1 实验数据说明

实验数据均由网络爬虫技术在腾讯网站获取。日期范围为 2020 年 1 月 13 日−2020 年 9 月 12 日，共计 244 条数据，将其按照训练集 214 条数据、测试集 30 条数据划分，针对国内累计确诊、现有确诊、累计治愈人数进行预测。

5.2.2 实验数据预处理

1) 数据提取、划分：在爬取到的新冠肺炎历史数据 CSV 文件中提取出来日期和累计确诊两列，将累计确诊列以矩阵形式表示：

$$\boldsymbol{X} = [x_0, x_1, \cdots, x_{243}]^{\mathrm{T}} \tag{7}$$

训练集、测试集为：

$$X_{\mathrm{train}} = [x_0, x_1, \cdots, x_{213}]^{\mathrm{T}} \tag{8}$$

$$X_{\mathrm{test}} = [x_{214}, x_{215}, \cdots, x_{243}]^{\mathrm{T}} \tag{9}$$

2) 数据归一化处理：为缩短模型训练时间，加速 loss 下降，对训练数据和测试数据分别进行归一化处理，这样可以最大程度保留数据特征，更好拟合新冠肺炎数据之间的非线性关系。本文采用最大最小归一化方法对数据进行处理：

$$x_{\mathrm{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

式中，x代表样本数据；$x_{\max}$、$x_{\min}$分别代表训练数据或测试数据中的最大值和最小值。设归一化处理后的训练集和测试集为X_{ntrain}和X_{ntest}，对现有确诊和累计治愈数据做同样处理。

5.3 LSTM 模型参数

本文提出的模型由 1 个输入层、多个隐藏层和 1 个输出层组成。上一个隐藏层的输出作为下一个隐藏层的输入，输入层与隐藏层共同实现输入数据特征的提取，最后一个隐藏层的输出为一维列向量，经线性回归即得到处理后的预测值[23]。经反复实验，最终超参数设置如下：激活函数为 Linear、迭代次数为 15 000，隐藏层为 100，Dropout 为 0.05，学习率为 0.000 1，优化器为 Adam。

5.4 实验流程

目前缺乏利用深度学习对新冠肺炎发展趋势预测的研究，这是因为：1) 深度学习对数据量要求较高，利用传统手工搜集方法不易获取相关数据；2) 训练过程中易出现过拟合问题，严重影响预测精度。在已有的 LSTM 神经网络用于新冠肺炎预测的研究中，有以下不足：1) 未提供简单的数据获取方法；2) LSTM 层过少，不能充分挖掘时序数据之间的非线性关系；3) 未考虑过拟合问题给实验结果带来的影响。针对以上问题，本文利用网络爬虫技术获取新冠肺炎历史数据组成实验数据集；在网络层上构建层次更深的 LSTM 新冠肺炎发展趋势预测模型；在每个 LSTM 单元构成的隐含层中的非循环部分采用 Dropout 技术对神经元进行随机概率失活，有效避免 LSTM 神经网络中的过拟合问题。本文实验流程图如图 6 所示。

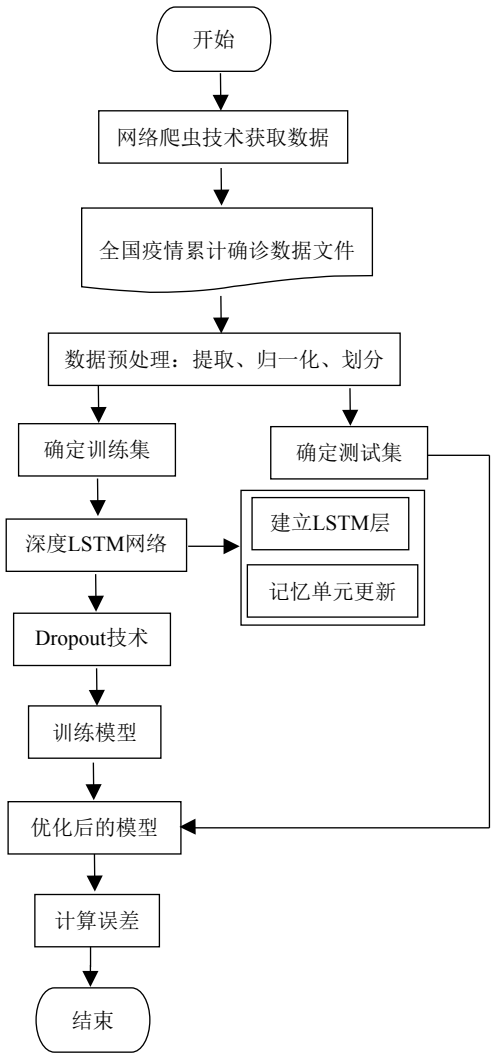


图 6 运用 LSTM 的全国累计确诊预测流程

6　实验结果分析

6.1　实验评价指标

本文预测评价指标采用平均绝对百分误差 (mean absolute percentage error, MAPE)、平均绝对误差 (mean absolute error, MAE) 和均方根误差 (root mean square error, RMSE) 衡量：

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=0}^{n}\left|\frac{y_i-y_p}{y_i}\right|\times 100\% \tag{11}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=0}^{n}|y_i-y_p| \tag{12}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i-y_p)^2} \tag{13}$$

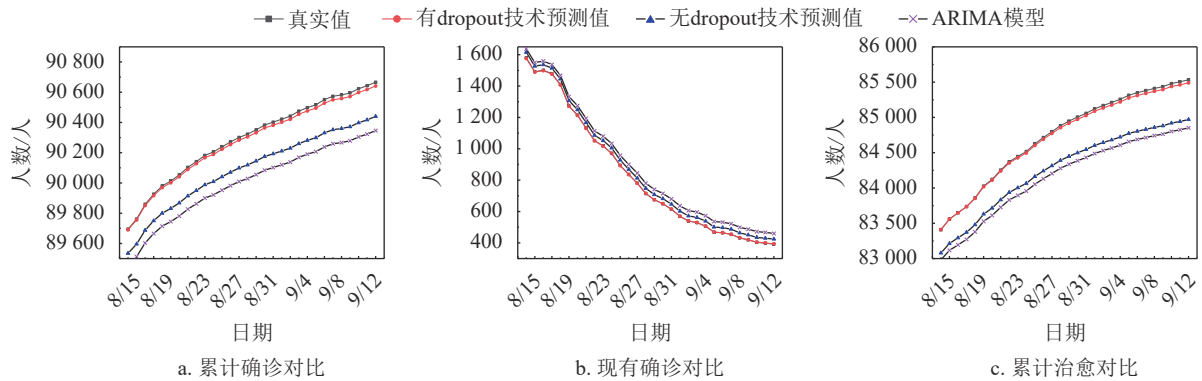
式中，y_i表示真实值；y_p表示预测值；n 表示待预测的新冠肺炎天数，即 30。

6.2　模型预测结果

为证明本模型 (Dropout-LSTM 模型) 的普遍适用性与优越性，运用 ARIMA 模型、LSTM 模型分别对 2020 年 8 月 14 日−2020 年 9 月 12 日的国内累计确诊、现有确诊和累计治愈数据进行预测，并与 Dropout-LSTM 模型预测结果进行对比。Dropout-LSTM 模型预测值和误差如表 1 所示，3 种模型的可视化对比如图 7 所示。

表 1　全国新冠肺炎发展趋势预测

日期	全国累计确诊/人			全国现有确诊/人			全国累计治愈/人			日期	全国累计确诊/人			全国现有确诊/人			全国累计治愈/人		
	真实值	预测值	误差	真实值	预测值	误差	真实值	预测值	误差		真实值	预测值	误差	真实值	预测值	误差	真实值	预测值	误差
8.14	89695	89690	−5	1580	1576	−4	83407	83412	5	8.29	90351	90332	−19	675	675	0	84948	84918	−30
8.15	89761	89755	−6	1492	1489	−3	83559	83561	2	8.30	90383	90364	−19	649	650	1	85005	84974	−31
8.16	89859	89851	−8	1501	1498	−3	83648	83648	0	8.31	90402	90382	−20	614	615	1	85058	85026	−32
8.17	89926	89916	−10	1479	1476	−3	83737	83735	−2	9.1	90422	90402	−20	569	570	1	85122	85088	−34
8.18	89980	89969	−11	1410	1407	−3	83858	83853	−5	9.2	90442	90421	−21	539	540	1	85169	85134	−35
8.19	90013	90001	−12	1273	1271	−2	84027	84018	−9	9.3	90475	90454	−21	529	530	1	85211	85175	−36
8.20	90053	90041	−12	1215	1213	−2	84122	84111	−11	9.4	90498	90476	−22	506	507	1	85257	85220	−37
8.21	90103	90089	−14	1133	1131	−2	84254	84240	−14	9.5	90517	90495	−22	468	469	1	85314	85276	−38
8.22	90141	90127	−14	1052	1051	−1	84372	84355	−17	9.6	90551	90528	−23	464	465	1	85350	85311	−39
8.23	90182	90167	−15	1018	1017	−1	84446	84428	−18	9.7	90573	90550	−23	454	455	1	85380	85340	−40
8.24	90205	90189	−16	971	970	−1	84516	84496	−20	9.8	90582	90558	−24	431	433	2	85411	85370	−41
8.25	90239	90223	−16	894	893	−1	84626	84603	−23	9.9	90595	90571	−24	419	420	1	85436	85395	−41
8.26	90271	90254	−17	836	836	0	84715	84690	−25	9.10	90623	90598	−25	403	405	2	85480	85438	−42
8.27	90301	90283	−18	780	780	0	84799	84773	−26	9.11	90643	90618	−25	398	400	2	85505	85462	−43
8.28	90323	90305	−18	715	715	0	84883	84855	−28	9.12	90666	90641	−25	392	394	2	85533	85490	−43



a. 累计确诊对比　　b. 现有确诊对比　　c. 累计治愈对比

图 7　新冠肺炎预测效果

误差为预测值与真实值之差。误差为负，代表预测值小于真实值；误差为正，代表预测值大于真实值。本文针对 30 天的数据进行预测，当预测日期后移时，将预测日期之前的数据加进来。若仅用当前数据去预测未来两个月甚至更久，从表 3 可看出误差有恶化趋势，选择适当的预测天数和日渐增多的数据量可解决误差趋势恶化问题。

若仅预测 30 天的数据，恶化程度是可控的，

误差会固定在一定区间，对基数较大的累计确诊数和累计治愈，数据误差区间为 [−25, −5] 和 [−43, 8]，基数较小的现有确诊误差仅为 [−4, 2]。随着新冠肺炎数据量增多，神经网络预测也会更加准确。

由图 7 可明显看出，累计确诊和累计治愈真实值总体呈上升趋势，现有确诊真实值呈下降趋势。

对 3 组数据分别进行预测时，ARIMA 模型预测曲线与真实值曲线差别均较大，拟合效果较差。这是因为 ARIMA 模型网络结构较简单，非线性拟合能力较弱。LSTM 模型拟合效果次之，这是因为虽然 LSTM 神经网络对时序数据有较好的拟合能力，但是出现了过拟合问题，无论如何调参，测试集的准确度依然不高。Dropout-LSTM 模型预测曲线与真实值曲线最接近，尤其是在现有确诊预测中，几乎与真实值曲线重合，这是因为添加 Dropout 技术后解决了 LSTM 神经网络中的过拟合问题，从而提高了准确度。

6.3 评价指标对比

为进一步说明 Dropout-LSTM 模型具有较高的拟合能力，采用 ARIMA 模型、LSTM 模型和 Dropout-LSTM 模型对 3 组数据分别进行预测，表 2 对其 MAPE、MAE 和 RMSE 进行了对比。

表 2 评价指标对比

类别	模型	MAPE/%	MAE	RMSE	时间/s
累计确诊	ARIMA	0.320	288.458	289.403	290
	LSTM	0.221	199.453	200.437	223
	Dropout-lSTM	0.019	16.793	17.849	229
现有确诊	ARIMA	9.330	63.816	63.945	267
	LSTM	3.588	25.551	25.566	252
	Dropout-lSTM	0.203	17.034	17.162	266
累计治愈	ARIMA	0.681	577.583	583.498	302
	LSTM	0.252	227.256	228.230	248
	Dropout-lSTM	0.029	24.449	28.077	323

从模型角度看，Dropout-LSTM 模型的 MAPE、MAE 和 RMSE 在 3 组数据中均有明显下降，说明误差降低，准确率提高，排除了只对一组数据有较好预测效果的可能。这是因为 Dropout 技术使神经元按照一定概率失活，神经网络不会偏向于某一个节点，从而使每一个节点的权重不会过大，减轻了 LSTM 神经网络的过拟合现象，提高了预测准确度。从时间角度看，LSTM 模型训练时间最短，这是因为 LSTM 神经网络共享参数，节省了训练时间；而 Dropout-LSTM 模型比 LSTM 模型训练时间长，这是因为神经元按照随机概率失活后，神经网络结构变得更“粗糙”，为使曲线更平滑地到达理想值，收敛到全局最优，增加了训练时间。

虽然向 LSTM 模型中接入 Dropout 技术增加了网络训练时间，但是随着数据量增多，需要的网络层数、训练轮数会相应减少，过拟合问题也会减轻，使网络训练在可接受的时间内完成。

7 结 束 语

本文提出的运用 Dropout-LSTM 模型预测新冠肺炎发展趋势的方法，通过网络爬虫技术获取新冠肺炎数据用于实验，解决了人工搜集方法的不足，提高了数据收集速度；构建层数更多的 LSTM 神经网络用于国内累计确诊、现有确诊和累计治愈的预测，在此基础上引入 Dropout 技术对神经元进行随机概率失活，有效避免了神经网络过拟合问题，充分挖掘了数据的时序性与非线性关系。通过全国累计确诊、现有确诊和累计治愈人数验证了运用 Dropout-LSTM 模型对新冠肺炎趋势预测是完全可行的，并且准确度较高。

参 考 文 献

[1] CHAN J F, YUAN Shuo-feng, KOK K, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster[J]. The Lancet, 2020, 395(10223): 514-523.

[2] WEI Yan-qiu, LU Yan-jun, XIA Li-ming, et al. Analysis of 2019 novel coronavirus infection and clinical characteristics of outpatients: An epidemiological study from a fever clinic in Wuhan, China[J]. Journal of Medical Virology, 2020: 2758-2767.

[3] ZHANG Qi, YU Yu. Epidemiological features of the 2019 novel coronavirus outbreak in China[J]. Current Topics in Medicinal Chemistry, 2020, 20(13): 1137-1140.

[4] 杨政, 原子霞, 贾祖瑶. 基于迁徙数据估计武汉感染新型冠状病毒的人员数量[J]. 电子科技大学学报, 2020, 49(3): 330-338.
YANG Zheng, YUAN Zi-xia, JIA Zu-yao. Estimating the number of people infected with COVID-19 in Wuhan based on migration data[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(3): 330-338.

[5] CHANG H J. Estimation of basic reproduction number of the Middle East respiratory syndrome coronavirus (MERS-CoV) during the outbreak in South Korea, 2015[J]. Biomedical Engineering Online, 2017, 16(1): 79.

[6] ILSU C, HO L D, KUK K Y. Effects of Timely control intervention on the spread of middle east respiratory syndrome coronavirus infection[J]. Osong Public Health and Research Perspectives, 2017, 8(6): 373-376.

[7] 武文韬, 柏如海, 李达宁, 等. 广东省新型冠状病毒肺炎疫情流行趋势的初步预测[J]. 暨南大学学报 (自然科学与

医学版), 2020, 41(2): 181-185.
WU Wen-tao, BAI Ru-hai, LI Da-ning, et al. Preliminary prediction of the epidemic trend of corona virus disease 2019 in Guangdong province[J]. Journal of Jinan University (Natural Science & Medicine Edition), 2020, 41(2): 181-185.

[8] 范如国, 王奕博, 罗明, 等. 基于 SEIR 的新冠肺炎传播模型及拐点预测分析[J]. 电子科技大学学报, 2020, 49(3): 369-374.
FAN Ru-guo, WANG Yi-bo, LUO Ming, et al. SEIR-based COVID-19 transmission model and inflection point prediction analysis[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(3): 369-374.

[9] 梅文娟, 刘震, 朱静怡, 等. 新冠肺炎疫情极限 IR 实时预测模型[J]. 电子科技大学学报, 2020, 49(3): 362-368.
MEI Wen-juan, LIU Zhen, ZHU Jing-yi, et al. Extreme IR model for COVID-19 real-time forecasting[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(3): 362-368.

[10] 王志心, 刘治, 刘兆军. 基于机器学习的新型冠状病毒(COVID-19)疫情分析及预测[J]. 生物医学工程研究, 2020, 39(1): 1-5.
WANG Zhi-xin, LIU Zhi-xin, LIU Zhao-jun. COVID-19 analysis and forecast based on machine learning[J]. Journal of Biomedical Engineering Research, 2020, 39(1): 1-5.

[11] 林德双, 金秀玲, 刘文鑫, 等. 新冠肺炎疫情预测分析[J]. 黑龙江工业学院学报 (综合版), 2020, 20(9): 114-119.
LIN De-shaung, JIN Xiu-ling, LIU Wen-xin, et al. Prediction and analysis of new coronavirus epidemic situation[J]. Journal of Heilongjiang University of Technology, 2020, 20(9): 114-119.

[12] 纪安之, 杨雪梅. 基于 ARIMA 模型的新冠肺炎序列分析预测[J]. 价值工程, 2020, 39(18): 107-109.
JI An-zhi, YANG Xue-mei. Analysis and prediction of time series of 2019-nCoV based on ARIMA model[J]. Journal of Value Engineering, 2020, 39(18): 107-109.

[13] 白璐, 郭佩汶, 范晋蓉. 湖北省新冠肺炎确诊人数的建模与预测分析[J]. 检验检疫学刊, 2020, 30(2): 10-12.
BAI Lu, GUO Pei-wen, FAN Jin-rong. Modeling and prediction of the number of confirmed cases of new coronavirus pneumonia in Hubei Province[J]. Journal of Inspection and Quarantine, 2020, 30(2): 10-12.

[14] 赵行健. 基于深度学习的新型冠状病毒肺炎疫情的动态监测研究[J]. 现代商贸工业, 2020, 41(20): 156-157.
ZHAO Xing-jian. Research on dynamic monitoring of COVID-19 based on deep learning[J]. Journal of Modern Trade Industry, 2020, 41(20): 156-157.

[15] 曾鹏飞, 刘辉. 基于二次相似性度量的即时学习转炉炼钢终点碳温软测量方法[EB/OL]. [2020-10-11]. http://kns.cnki.net/kcms/detail/11.5946.tp.20201217.1558.002.html.
ZENG Peng-fei, LIU Hui. A soft-sensing method for carbon temperature at the end of converter steelmaking based on quadratic similarity measurement[EB/OL]. [2020-10-11]. http://kns.cnki.net/kcms/detail/11.5946.tp.20201217.1558.002.html.

[16] 晏臻, 于重重, 韩璐, 等. 基于 CNN+LSTM 的短时交通流量预测方法[J]. 计算机工程与设计, 2019, 40(9): 2620-2624, 2659.
YAN Zhen, YU Chong-chong, HAN Lu, et al. Shot-term traffic flow forecasting method based on CNN+LSTM[J]. Journal of Computer Engineering and Design, 2019, 40(9): 2620-2624, 2659.

[17] 李瑞津, 刘斌, 张学敏, 等. 基于改进 LSTM 的变电站铅酸电池寿命预测[EB/OL]. [2020-10-11]. http://kns.cnki.net/kcms/detail/43.1129.TM.20201202.1340.030.html.
LI Rui-jin, LIU Bin, ZHANG Xue-min, et al. Life prediction of lead-acid battery in substation based on improved LSTM[EB/OL]. [2020-10-11]. http://kns.cnki.net/kcms/detail/43.1129.TM.20201202.1340.030.html.

[18] CASTILLO C. Effective web crawling[J]. ACM SIGIR Forum, 2005, 39(1): 55-56.

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

[20] 曹有为, 闫双红, 刘海涛. 基于降噪时序深度学习网络的风电功率短期预测方法[J]. 电力系统及其自动化学报, 2020, 32(1): 145-150.
CAO You-wei, YAN Shuang-hong, LIU Hai-tao. Short-term wind power forecasting method based on noise-reduction time-series deep learning network[J]. Journal of Proceedings of the CSU-EPSA, 2020, 32(1): 145-150.

[21] 王瑞, 闫方, 逯静, 等. 运用相似日和 LSTM 的短期负荷双向组合预测[J]. 电力系统及其自动化学报, 2020, DOI: 10.19635/j.cnki.csu-epsa.000671.
WANG Rui, YAN Fang, LU Jing, et al. Bidirectional combined short-term load forecasting by using similar days and LSTM[J]. Journal of Proceedings of the CSU-EPSA, 2020, DOI: 10.19635/j.cnki.csu-epsa.000671.

[22] 张柏翰, 凌捷. 改进的基于 DNN 的恶意软件检测方法 [J]. 计算机工程与应用, 2020, DOI: 11.2127.TP.20200819.1505.020.
ZHANG Bo-han, LING Jie. An improved malware detection method based on DNN[J]. Journal of Computer Engineering and Applications, 2020, DOI: 11.2127.TP.20200819.1505.020.

[23] 张宇帆, 艾芊, 林琳. 基于深度长短时记忆网络的区域级超短期负荷预测方法[J]. 电网技术, 2019, 43(6): 1884-1892.
ZHANG Yu-fan, AI Qian, LIN Lin. A very short-term load forecasting method based on deep LSTM RNN at zone level[J]. Journal of Power System Technology, 2019, 43(6): 1884-1892.

编　辑　叶　芳