新冠病毒疫情预测模型研究方法评述

宁 晴　鲍 泓　徐 成

北京联合大学北京市信息服务工程重点实验室　北京 100101

(191083510904@buu.edu.cn)

摘　要　新型肺炎冠状病毒(世界卫生组织命名为COVID-19,简称新冠病毒)具有很强的传染性,对疫情发展趋势进行准确的预测具有重大的现实意义。对当前国内外具备一定影响力的预测模型研究方法进行了分析和对比,同时对疫情的发展和我国抗疫的实际情况进行了评估,介绍了目前疫情预测的主流算法及其优缺点,以期为制定重大疫情的防控措施提供参考。

关键词:COVID-19;疫情预测;机器学习;SEIR模型;LSTM

中图法分类号　TP391

Survey of Novel Coronavirus Epidemic Prediction Model

NING Qing, BAO Hong and XU Cheng

Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China

Abstract　Novel coronavirus (the world health organization named COVID-19) is highly infectious, it is of great practical significance to accurately predict the development trend of the epidemic situation. This paper analyzes and compares the research methods of prediction models at home and abroad, evaluates the development of epidemic situation and the actual situation of anti epidemic in China, and introduces the mainstream algorithm of epidemic prediction and its advantages and disadvantages, so as to provide reference for the formulation of prevention and control measures of major epidemic situation.

Keywords　COVID-19, Epidemic prediction, Machine learning, SEIR model, LSTM

1　引言

2019年爆发的新型冠状病毒所致的肺炎(现称COVID-19),其临床特征显示此次病毒感染者具有发热症状,表现为呼吸状态受损,且不排除超级传播者的存在[1],同时无症状感染者[2]的存在给疫情防控带来了巨大的挑战。对于疫情防控,有学者分析了消毒[3]、人员流动[4]、温度[5]对病毒传播的影响。准确预测疫情的趋势,将对疫情防控的指导起到重要的作用。

国内外各个团队对疫情趋势预测展开研究,分别采用了统计学模型以及SEIR模型改进的SEIR模型、以及机器学习模型,但预测结果存在较大浮动。文献[6-7]采用SEIR模型对拐点和峰值进行预测。经典的SEIR模型泛化适用于各类疫情,但须考虑人员流动的感染。文献[8]采用修正的SEIR模型

基金项目:国家自然科学基金(61932012);北京市属高校高水平教师队伍建设支持计划项目(IDHT20170511);北京联合大学项目(BPHR2020EZ01,ZK80202001)

This work was supported by the National Natural Science Foundation of China(61932012), Supporting Plan for Cultivating High Level Teachers in Colleges and Universities in Beijing(IDHT20170511) and Beijing Union University Project (BPHR2020EZ01, ZK80202001).

通信作者:鲍泓(baohong@buu.edu.cn)

对 COVID-19 进行拟合分析预测，但由于未考虑防控措施对人员流动的影响，预测结果与国家卫健委报告人数存在较大偏差[9]。文献[4,10-16]在考虑现有疫情防控措施基础上，将人员流动引入 SEIR 模型对疫情进行预测，也得出旅游禁令的有效性的结论。东南大学的科研人员在 medRxiv 上发表论文[17]，采用修正的 SEIR 模型对 2019-nCOV 疫情暴发的流行趋势及风险进行评估。无论是 SEIR 模型，还是修正的 SEIR 模型，均需要分析大量的参数，包括 R0、移除率等，而人员网络存在大量的不确定性因素。对此，文献[4,18-20]采用机器学习的方法进行数据的预测，文献[21]通过随机森林的方法将不同城市分成不同的防控等级，对疫情防控提供了很好的借鉴思路。同时，国外学者对疫情进行了大量的预测和分析，如模拟传染病的统计学家 Sebastian Funk[22]和日本札幌北海道大学的流行病学家 Hiroshi Nishiura[22]团队对高峰进行了预测。

2 理论与方法

国内外各个团队对疫情趋势预测展开研究，分别采用 SEIR 模型、改进的 SEIR 模型以及机器学习模型对疫情趋势进行预测。

2.1 SEIR 模型

SEIR 模型中，S 指易感状态，E 指潜伏状态，I 指感染状态，R 指康复或者死亡。文献[6]建立 SEIR 模型，并利用其对武汉地区感染人数峰值和拐点时间进行预测。$S(t)+E(t)+I(t)+R(t)=N$，其中 $S(t)$，$E(t)$，$I(t)$和 $R(t)$分别为时刻 t 的易感人群数、潜伏人群数、感染人群数、移出人群数，N 为种群的个体数。假设一个易感状态在单位时间里与感染个体接触并被传染的概率为 β，整体以单位时间概率转化为感染个体 γ_1，感染个体数目由潜伏群体转化而来，同时以单位时间概率 γ_2 转化为移除状态。

$$\begin{cases}\dfrac{\mathrm{d}S}{\mathrm{d}t}=-\dfrac{\beta SI}{N}\\ \dfrac{\mathrm{d}E}{\mathrm{d}t}=\dfrac{\beta SI}{N}-\gamma_1 E\\ \dfrac{\mathrm{d}I}{\mathrm{d}t}=\gamma_1 E-\gamma_2 I\\ \dfrac{\mathrm{d}R}{\mathrm{d}t}=\gamma_2 I\end{cases}\tag{1}$$

2.2 基于改进的 SEIR 模型预测

Wu 等[8]通过国际航空数据以及国外确诊数据得到患病比例，预测武汉确诊数据，并通过武汉确诊比例以及从腾讯数据库中获取的客流量预估全国确诊病例数量。通过 SEIR 种群模型，主要对湖北、重庆等城市的确诊数量进行了估计。$z(t)$为人畜共患病的传染力，L_{IW}为从国际进入到武汉的人数，L_{CW}为从国内进入武汉的人数，L_{WI}为武汉旅游到国际的人数，L_{WC}为武汉旅游到国内的人数，D_{E}和 D_{I}是潜伏时间的均值和感染时间的均值。

$$\begin{cases}\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-\dfrac{S(t)}{N}\left(\dfrac{R_0}{D_{\mathrm{I}}}I(t)+z(t)\right)+L_{\mathrm{IW}}+\\ \qquad L_{\mathrm{CW}}(t)-\left(\dfrac{L_{\mathrm{WI}}}{N}+\dfrac{L_{\mathrm{WC}}(t)}{N}\right)S(t)\\ \dfrac{\mathrm{d}E}{\mathrm{d}t}=\dfrac{S(t)}{N}\left(\dfrac{R_0}{D_{\mathrm{I}}}I(t)+z(t)\right)-\dfrac{E(t)}{D_E}\left(\dfrac{L_{\mathrm{WI}}}{N}+\right.\\ \qquad \left.\dfrac{L_{\mathrm{WC}}(t)}{N}\right)E(t)\\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=\dfrac{E(t)}{D_{\mathrm{E}}}-\dfrac{I(t)}{D_{\mathrm{I}}}-\left(\dfrac{L_{\mathrm{WI}}}{N}+\dfrac{L_{\mathrm{WC}}(t)}{N}\right)I(t)\end{cases}\tag{2}$$

2.3 机器学习方法

文献[4]使用 LSTM 模型——一种用于处理和预测各种时间序列问题的递归神经网络来预测随着时间的推移新感染的人数。该研究使用了 2003 年 SARS 疫情统计数据，纳入了 COVID-19 流行病学参数，如传播概率、潜伏期、恢复率和接触次数。由于数据集相对较小，作者开发了一个更简单的网络结构来防止过度拟合；同时使用 Adam 优化器，并运行了 500 次迭代。

$$\begin{cases}f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f)\\ i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i)\\ \tilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c)\\ C_t=f_c*C_{t-1}+i_t*\tilde{C}_t\\ o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o)\\ h_t=o_t*\tanh(C_t)\end{cases}\tag{3}$$

3 扩展讨论与个案分析

从疫情的影响因素和预测的角度对上述研究进行分析比较。疫情的影响因素包括年龄、潜伏期、无症状感染者，从预测的拐点和峰值进行对比，如表 1 所列。

表 1　已有预测结果的对比

Table 1　Comparison of prediction results

文献	方法	预测
Sebastian Funk 等[22-23]	改进 SEIR 模型	峰值 100 万
Hiroshi Nishiura 等[24]	贝叶斯定理	峰值 3—5 月，峰值感染人数 5～6 亿
Joseph T Wu 等[8]	改进 SEIR 模型	到 1 月 25 日，武汉市将达到 7.5 万人，4 月将达到疫情高峰
东南大学[18]	改进 SEIR 模型	湖北以外于 2 月中旬达到拐点，3 月初达到峰值 13806
Read 等[25]	改进 SEIR 模型	到 2 月 4 日将出现 19 万例病例的高峰
Yang 等[4]	改进 SEIR 模型和 LSTM 模型	2 月 4—7 日每日新增确诊人数达到高峰，有接近 4 000 例，4 月底累计确诊人数达到 95811 人

根据国家卫健委官方通报数据（如图 1 所示），中国新增确诊人数于 2020 年 2 月 12 日达到最大，为 15152 人，4 月 25 日的累计确诊人数为 82827 人，新增确诊人数为 11 人，与文献[8]预测结果高度吻合。使用 LSTM 将美国截止到 6 月 30 日的确诊人数输入模型中，预测结果为总确诊人数于 9 月 10 日达到峰值，为 1 135 737，接近 390 万，如图 2 所示。

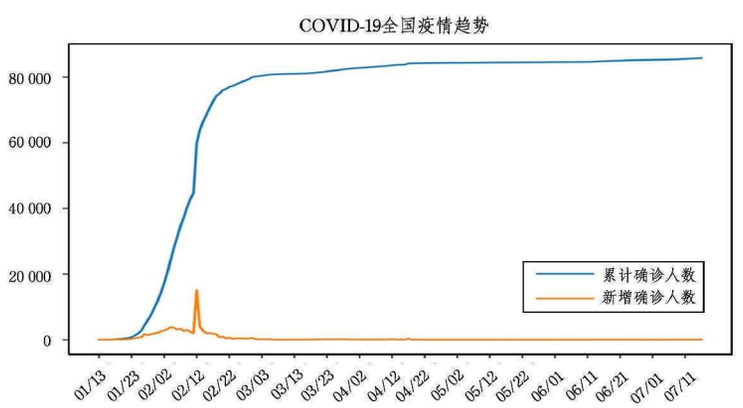

图 1　中国新增确诊人数、累计确诊人数数据

Fig. 1　The number of new confirmed cases and the cumulative number of confirmed cases in China

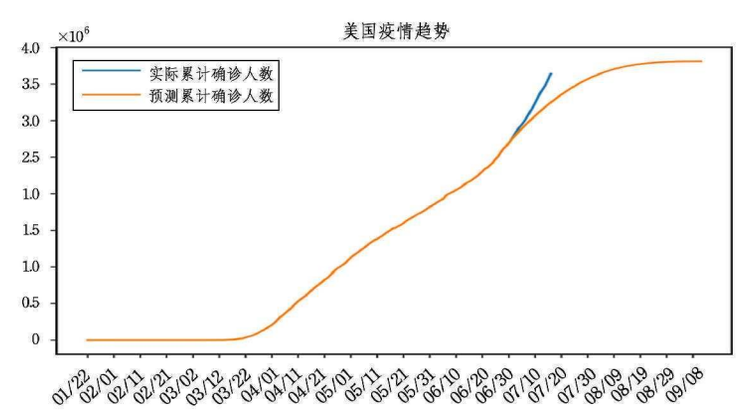

图 2　美国疫情趋势

Fig. 2　Epidemic trend in the US

对 7 月 1 日－16 日的预测数据进行跟进验证对比，如图 2 所示。由于模型参数调整为用前三天（$n-2$，$n-1$，n）的数据预测未来一天 $n+1$ 的数据，为了预测未来整体趋势，采取将预测值加入已知数据集中的策略，这种方法存在误差累计被放大的现象。可从两方面对如上预测建议进行改进：1）更改训练集，缩小误差；2）组合深度模型方法，如 LSTM、强化学习、CNN、GAN

等，以达到缩小误差的目的。

根据论文预测方法和结果对比，传统方法从底向上的处理思路需要分析人员流动、人口密度和具体防控措施等参数之间的关系，对不同参数影响的分析对预测结果有较大的影响，须准确引入相关参数并完成预测，模型的泛化能力较差。而机器学习方法通过现有数据建立模型，仅考虑数据相关性，数据的相关性也直接决定着准确性；对未来几日数据进行预测，可得到较小的误差，但随着时间的推移，误差累积，会出现误差增大的现象。

结束语 新冠病毒疫情的预测，是根据现有数据，包括已确诊人数、死亡人数、康复人数等数据，来预测疫情的拐点和规模，疫情的准确预测能够为疫情防控争取大量的时间；同时，分析疫情影响关键因素为制定及时有效的防控策略提供了数据基础。文中从传统模型和机器学习两个方向分析了不同预测模型，比较了模型的预测结果，并与现有数据进行了分析对比。

更多地考虑现有防控措施，尤其是人员流动，对疫情蔓延的预测表现出了更高的准确性；但疫情预测模型在不同地区、不同国家的泛化能力，以及疫情预测的实时性，仍是当前的研究重点。传统模型需要考虑更多的参数，机器学习方法仅考虑数据相关性，未来可将两者进行结合；同时，如果能更早地预测疫情的发展，将对疫情防控具有更重要的现实意义。

参 考 文 献

[1] GUAN W. Clinical Characteristics of Coronavirus Disease 2019 in China[J]. New England Journal of Medicine, 2020, 382(18): 1708-1720.

[2] MIZUMOTO K, KAGAYA K, ZAREBSKI A, et al. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020[J]. Eurosurveillance, 2020, 25(10): 2000180.

[3] KAMPF G. Potential role of inanimate surfaces for the spread of coronaviruses and their inactivation with disinfectant agents[J]. Infection Prevention in Practice, 2020, 2(2): 100044.

[4] YANG Z, ZENG Z, WANG K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. Journal of Thoracic Disease, 2020, 12(2): 165-174.

[5] WANG M. Temperature significant change COVID-19 Transmission in 429 cities[J]. medRxiv, 2020: 2020.02.22.20025791.

[6] 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.

[7] 蔡洁，贾浩源，王珂. 基于 SEIR 模型对武汉市新型冠状病毒肺炎疫情发展趋势预测[J]. 山东医药，2020，60(6)：1-4.

[8] WU J T, LEUNG K, LEUNG G M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study[J]. The Lancet, 2020, 395(10225): 689-697.

[9] 中华人民共和国国家卫生健康委员会. 新型冠状病毒感染的肺炎疫情最新情况[EB/OL]. [2020-01-25]. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.

[10] WAN K, CHEN J, LU C, et al. When will the battle against novel coronavirus end in Wuhan: A SEIR modeling analysis[J]. J Glob Health, 2020, 10(1): 011002.

[11] 曹盛力，冯沛华，时朋朋. 修正 SEIR 传染病动力学模型应用于湖北省 2019 冠状病毒病(COVID-19)疫情预测和评估[J]. 浙江大学学报(医学版)，2020，49(2)：178-184.

[12] CHINAZZI M. The effect of travel restrictions on the spread of the 2019 novel corona virus (COVID-19) outbreak[J]. Science, 2020: eaba9757.

[13] 颜铭江，董一鸿，贾香恩，等. 新型冠状病毒肺炎的疫情趋势预测[J]. 病毒学报，2020，36(4)：560-569.

[14] 林俊锋. 基于引入隐形传播者的 SEIR 模型的 COVID-19 疫情分析和预测[J]. 电子科技大学学报，2020，49(3)：375-382.

[15] TANG Z, LI X, LI H. Prediction of New Coronavirus Infection Based on a Modified SEIR Model[J]. medRxiv, 2020: p. 2020.03.03.20030858.

[16] ZHANG Y. The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study[J]. medRxiv, 2020: 2020.03.04.20031187.

[17] LIU Q. Assessing the Tendency of 2019-nCoV (COVID-19) Outbreak in China[J]. medRxiv, 2020: 2020.02.09.20021444.

[18] 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.

[19] BANDYOPADHYAY S K，DUTTA S. Machine Learning Approach for Confirmation of COVID-19 Cases：Positive，Negative，Death and Release[J]. medRxiv，2020：2020. 03. 25. 20043505.

[20] ZHENG N N，DU S Y，WANG J J，et al. Predicting COVID-19 in China Using Hybrid AI Model[J]. IEEE Transactions on Cybernetics，2020，50(7)：2891-2904.

[21] LI X. Risk map of the novel coronavirus (2019-nCoV) in China：proportionate control is needed[J]. medRxiv，2020：2020. 02. 16. 20023838.

[22] CYRANOSKI D. When will the coronavirus outbreak peak? [EB/OL]. (2020-02-18)[2020-05-26]. https：//www. nature. com/articles/d41586-020-00361-5.

[23] KUCHARSKI A J. Early dynamics of transmission and control of COVID-19：a mathematical modelling study [J]. The Lancet Infectious Diseases，2020，20(5)：553-558.

[24] NISHIURA H. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)[J]. medRxiv，2020：020. 02. 03. 20020248.

[25] READ J M，BRIDGEN J R E，CUMMINGS DAT，et al. Novel coronavirus 2019-nCoV：early estimation of epidemiological parameters and epidemic predictions[J]. medRxiv，2020. 01. 23. 20018549.

NING Qing，a master student in software engineering at Beijing Union University. Her current research interests include intelligent vehicles and digital image processing.

BAO Hong，received his Ph.D degree from the School of Computer and Information Technology，Beijing Jiaotong University，Beijing，China. He is a professor at the Beijing Union University. His current research interests include intelligent control and intelligent vehicles.