生物医学工程研究
Journal of Biomedical Engineering Research 2020,39(1):1~5

网络首发时间:2020-02-13 14:49:34
网络首发地址:http://kns.cnki.net/kcms/detail/37.1413.R.20200213.0956.002.html

基于机器学习的新型冠状病毒(COVID-19)疫情分析及预测*

王志心,刘治△,刘兆军△
(山东大学信息科学与工程学院,青岛 266237)

摘要: 本研究采用数学建模的方式,在有限的数据下,通过机器学习对近期爆发的新型冠状病毒(COVID-19)肺炎确诊人数趋势进行了预测,根据有关部门发布的信息,预测了疫情拐点出现的时间,并对比了各省预计最终确诊人数所占的比例,以此为依据,大致划分了疫情的严重程度,对各省市人民防护工作有指导意义。

关键词: 新型冠状病毒肺炎;传播模型;疫情拐点;最小二乘准则;梯度下降;确诊人数预测
中图分类号: R318 **文献标识码:** A **文章编号:** 1672-6278(2020)01-0001-05

COVID-19 analysis and forecast based on machine learning

WANG Zhixin, LIU Zhi, LIU Zhaojun
(*School of Information Science and Engineering, Shandong University, Tsingtao* 266237, *China*)

Abstract: We used mathematical modeling to predict the trend of the number of newly diagnosed pneumonia outbreaks caused by COVID-19 with limited data through machine learning, and compared the proportion of estimated final diagnoses in each province. Based on that, the epidemic situation was roughly divided. The degree of severity could also be a guiding significance for people's self-protection work in various provinces and cities.

Key words: COVID-19 pneumonia; Propagation model; Inflection point; Least square error principle; Gradient descent; Prediction of diagnosed number

1 引 言

2019年12月起,湖北省武汉市开始出现原因不明的肺炎病例,2020年1月7日,首次检测出一种新型冠状病毒(COVID-19)[1-2]。该病毒主要通过飞沫和接触传播。随着春运的到来,新型冠状病毒肺炎(简称"新冠肺炎")很快波及全国。

在1个多月的时间内,新冠肺炎确诊患者和疑似患者的数量不断创出新高,说明该病毒的传染性较强。

2 传播模型[3]

模型把新冠肺炎传播分为两个阶段,第一阶段是对疫情不够重视的自由传播阶段,等价于疾病传播的SIR过程[4-7],在此阶段,新感染的患者数量以再生数R_0呈现出指数型增长的趋势。第二阶段是政府介入后,媒体对新冠肺炎的报道使人群采取自我保护行为,如待在家中或佩戴口罩出行等,阻断病

DOI 10.19529/j.cnki.1672-6278.2020.01.01
* 山东省自然科学基金重大基础研究资助项目(ZR2019ZD05)。
△通信作者 Email: liuzhi@sdu.edu.cn; zhaojunliu@sdu.edu.cn

毒传播渠道。在此阶段,疾病传播再生数下降至小于1,呈现出新增感染患者数量下降的趋势。

2.1 自然传播阶段的模型

在无外界干预的情况下,假设第一天的感染患者数量为 n,基本再生数为 k,那么第二天新增的感染患者数量为 nk,第三天为 nk^2,以此类推,可以得到在第 t 天的感染患者数量为:

$$\text{total}(t) = \sum_{t=1}^{t} n k^{t-1} = n \times \frac{1-k^t}{1-k} \tag{1}$$

其中,k 为基本再生数,n 为初始感染人数,t 为天数。

利用式(1)对2020年1月13日-2月3日公布的确诊感染数据进行拟合,结果见图1。由图1可知,曲线可以较好地拟合实际数据。

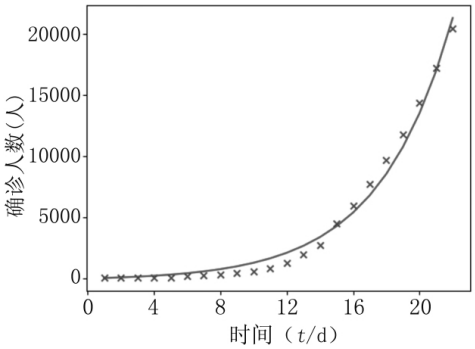


图1 1月13日至2月3日确诊患者数量与拟合曲线

Fig. 1 The fitting curve based on number of confirmed people from 13th, Jan. to 3rd, Feb.

2.2 干预后的传播模型

2.2.1 控制手段干预后的模型 对传染病的控制在于控制传染源,切断传播渠道,保护易感人群,这些都可归结为人为地降低基本再生数 k。

由SIR传播理论可知,只有当再生数小于1时,传染病才可被控制。当采取的控制手段力度大时,干预后再生数下降大,当采取的控制手段力度小时,干预后再生数下降小。在此,我们讨论干预后再生数小于1的情况。

假设在第 t_0 天,再生数小于1,那么在 t_0 天之前,传染病感染患者数量以指数型增长,此后总确诊患者数量呈现下降的趋势,由此可以得出:

$$\widetilde{\text{total}}(t) = \begin{cases} \text{total}(t), t \leqslant t_0 \\ \text{total}(t_0) + \widetilde{\text{total}}(t-1)(\tilde{k}+1), t > t_0 \end{cases} \tag{2}$$

其中,$\tilde{k}$ 为干预后再生数,是关于时间和政府干预强度的变量。

我们通过对国家卫生健康委员会公布的每日确诊患者数量来估算人际间传染率,对 $\tilde{k}$ 进行估计。

2.2.2 模型拐点的确定 现在需要确定 t_0 点,即传播人数的拐点。只有当大部分人都重视防护时,疫情才可能得到控制,查阅资料可发现在1月23日浙江省首先宣布启动重大突发公共卫生事件一级响应,1月25日晚30个省市均启动了重大突发公共卫生事件一级响应。已知该新型冠状病毒的潜伏期最长可达14天[8],由于人群反应会有延迟,因此,在经过约16~18天后,各种措施开始见效,疾病传播力度也逐渐减弱,即在2月13日左右将出现拐点。因此,我们用 $\widetilde{\text{total}}(t)$ 对数据进行拟合,以此得到疫情确诊人数的预测趋势图。

2.3 机器学习算法

回归算法是机器学习中最常见也是使用最广的一种算法,是一种有监督学习的算法。在这里我们使用最小二乘准则(least square error,LSE)和梯度下降算法对数据进行非线性回归,寻找天数与确诊患者数量的非线性关系。

2.3.1 最小二乘准则 最小二乘准则提供了一种损失函数的表达方法,基本思路是使得所有样本点到曲线或一面的距离最小。通过最小二乘准则可以很容易地写出损失函数,即:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(\widetilde{\text{total}}_\theta(x_i) - y_i\right)^2 \tag{3}$$

其中,$J(\theta)$ 为损失函数,y_i 为样本观测值。

2.3.2 梯度下降法 梯度下降算法在机器学习中的应用十分广泛,主要通过迭代找到目标函数的极小值,但多数情况下,其较难找到全局最优解,一般只能找到局部最优解,因此,对模型预测参数的准确性可能会产生一定的影响。

首先,我们对 θ 进行随机初始化,然后沿着负梯度的方向进行迭代,使得更新后的 θ 令 $J(\theta)$ 更小,公式如下:

$$\theta = \theta - \eta\frac{\partial J}{\partial \theta} \tag{4}$$

其中 θ 为参数,J 为损失函数,η 为学习率。

当 θ 下降到某个无法下降的点或者某个定义的极小值时,停止下降,并将得到的 θ 代入损失函数中,得到极小值,完成对参数的估计,见图2。

由式(5)求损失函数 $J(\theta)$ 对 θ_i 的偏导数:

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\frac{1}{2}\left(\widetilde{\text{total}}_\theta(x) - y\right)^2 = \left(\widetilde{\text{total}}_\theta(x) - y\right)x_i \tag{5}$$

2.3.3 计算数据与拟合结果 为求解拟合系数,所

需的数据如下：

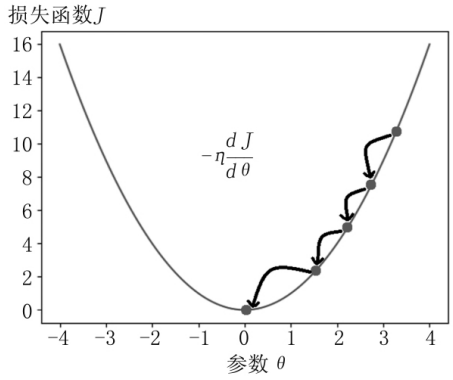


图 2　梯度下降示意图

Fig. 2　Diagram of gradient descent

$$X = [x_1, x_2, \cdots, x_n] \tag{6}$$

$$Y = [y_1, y_2, \cdots, y_n] \tag{7}$$

$$\widetilde{\text{total}}(t) = \begin{cases} \text{total}(t), t \leqslant t_0 \\ \text{total}(t_0) + \widetilde{\text{total}}(t-1)(\hat{k}+1), t > t_0 \end{cases} \tag{8}$$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(\widetilde{\text{total}}_\theta(x_i) - y_i\right)^2 \tag{9}$$

其中，X，Y 为样本数据，$\widetilde{\text{total}}(t)$ 为模型函数，$J(\theta)$ 为损失函数。

根据上述步骤计算出拟合值后，便可对确诊患者数量进行预测，结果见图 3。由图 3 可知，在拐点到来后 20 天左右，由于干预强度逐渐增大，疫情开始趋于稳定，最终确诊患者数量在 75 000 人左右。

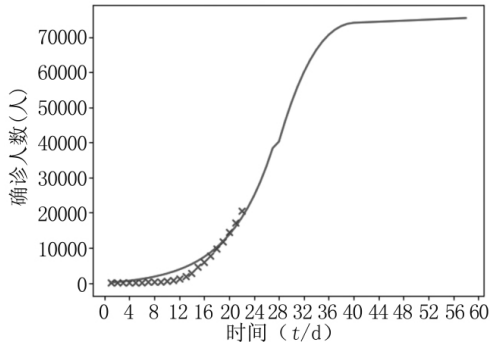


图 3　基于 1 月 13 日至 2 月 3 日确诊人数拟合数据的预测曲线

Fig. 3　Prediction based on the fitting curve from 13th, Jan. to 3rd, Feb.

3　主要省份的感染规模预测

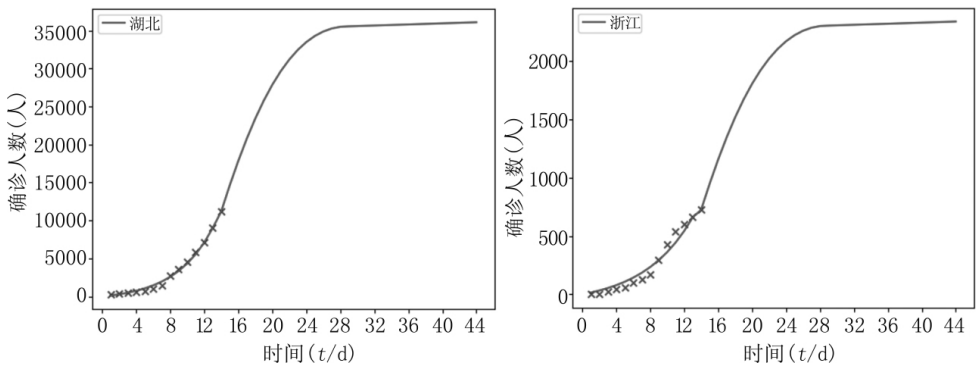
根据上述模型对部分省市的感染规模做简单推算，表 1 为 2020 年 1 月 20 日至 2 月 2 日部分省市卫生健康委员会发布的确诊患者数量。

表 1　各省市卫健委发布的累计确诊患者数量

Table 1　Number of confirmed cases issued by provincial health committees

日期	1/20	1/21	1/22	1/23	1/24	1/25	1/26	1/27	1/28	1/29	1/30	1/31	2/1	2/2
湖北	270	375	444	549	729	1 052	1 423	2 714	3 554	4 586	5 806	7 153	9 074	11 177
浙江	0	4	25	42	61	103	127	172	295	428	537	599	661	724
广东	14	22	28	49	74	94	142	184	237	311	393	520	604	683
河南	0	0	0	4	27	74	119	159	197	278	352	422	493	566
湖南	0	1	4	19	38	64	95	138	216	277	332	389	463	521
安徽	0	0	5	15	39	60	70	106	152	200	237	297	340	408
重庆	0	4	10	26	56	74	109	131	146	165	206	238	262	300
江西	0	0	5	11	16	34	46	70	107	162	240	286	333	391
四川	3	3	5	8	26	42	67	88	106	142	177	207	231	254
山东	0	2	6	9	21	39	63	87	121	136	178	202	225	246

对表 1 数据用模型进行拟合，然后对各省市最终感染患者数量进行预测，结果见图 4。



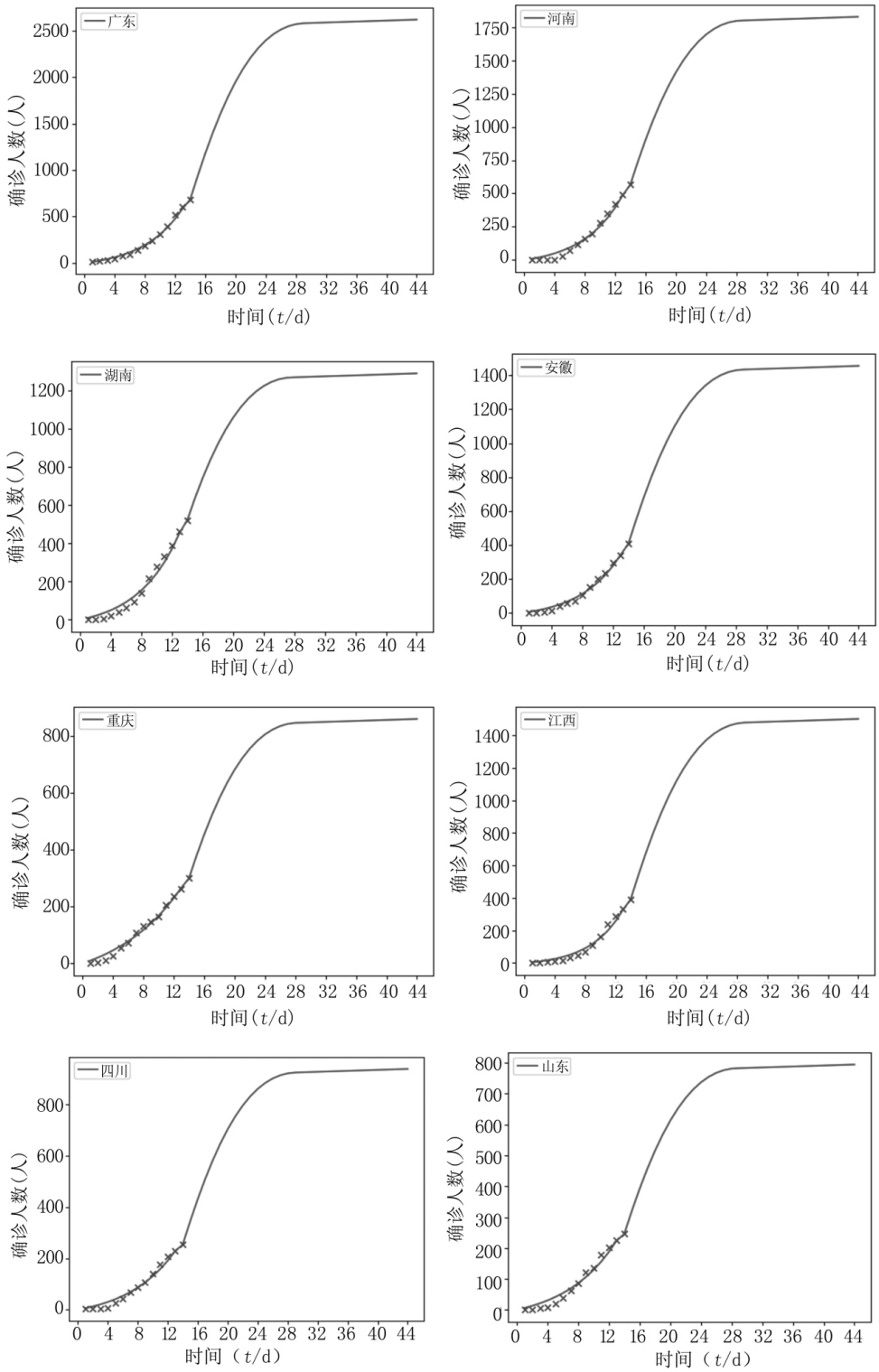


图 4 各省市确诊患者数量预测曲线

Fig. 4 Prediction curve of each province

由图 4 可以得出各省市的预计感染患者数量，将其与本省市人口进行对比，确定确诊患者数量在本省市的占比，以此评估各省市新冠肺炎的严重程度，见表 2。由表 2 可知，湖北预估确诊人数最多，其次是浙江和广东，这与当前疫情严重程度相吻合。

4 结论

本研究通过对新型冠状病毒肺炎的传播模型进行建模，并根据时间节点等数据预测了拐点出现的时间。结果表明，疫情在 2020 年 1 月 25 日后 16 ~

表 2　各省市预估确诊患者数量在本省市人口的占比

Table 2　Proportion of the estimated numbers of confirmed cases in the population ofeach province

省市	预估确诊患者数量	本省市人口（亿）	预估确诊患者数量占比（%）
湖北	37 000	0.5917	0.06
浙江	2500	0.5737	0.004
广东	2500	1.13	0.002
河南	1700	0.9605	0.002
湖南	1400	0.6899	0.002
安徽	100	0.6324	0.002
重庆	600	0.3102	0.002
江西	1400	0.4648	0.003
四川	800	0.8341	0.0009
山东	800	1	0.0008

18 天左右将会出现拐点，在一个月左右确诊患者数量将会趋向平稳，新增确诊患者数量将很少。从预估确诊患者数量在各省市中占比来看，湖北的严重程度为第一梯队，浙江、广东、河南、湖南、安徽、重庆、江西为第二梯队，如果生活在以上几个省市应尽量减少外出，外出时应避免前往人群聚集的地方，并采取自我防护措施，注意佩戴口罩。

参考文献：

[1] Chen N S，Zhou M，Dong X，et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan，China：a descriptive study [J]. The Lancet，2020（Pre－publis）.

[2] Lu R J，Zhao X，Li J，et al. Genomic characterization and epidemiology of 2019 novel coronavirus：implications for virus origins and receptor binding [J]. The Lancet，2020（Pre－publis）.

[3] 马知恩，周义仓，王稳地，等. 传染病动力学的数学建模与研究 [M]. 北京：科学出版社，2004.

[4] 霍阔，李世霖. 甲型 H1N1 流感传播的 SIR 模型研究 [J]. 湖南工业大学学报，2010，4：40－42.

[5] 刘来福，曾文艺. 数学模型与数学建模 [M]. 北京：北京师范大学出版社，1997.

[6] 王汝发. SARS 传播的数学模型分析 [J]. 数理医药学杂志，2004，17（2）：99－100.

[7] 宇永仁，杨颖，明珠，等. SARS 传播模型及其对经济的影响 [J]. 辽宁大学学报（自然科学版），2005，32（1）：48－49.

[8] 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防控方案（第二版）[Z]. 2020－01－22.

（收稿日期：2020－02－11）