遗传算法应用于武汉新冠肺炎传染病发展预测

February 21, 2020

1 摘要

2020年1月，湖北省武汉市爆发了名为“新型冠状病毒肺炎”的疫情，这种疾病由新型冠状病毒（2019-nCoV）引起，本次疫情发展迅速，影响较大。针对这一次疫情，本文提出一种新的武汉肺炎传染病模型：增强型SEIR（E-SEIR），它是基于SEIR基础上，做了如下考虑：考虑病毒在潜伏期也具有相同传染性、考虑确诊患者在医院进行了较好的隔离，这部分患者比率传染性配比了隔离系数、考虑使用带参数的s型函数拟合传染率的下降规律，来模拟日益完善的防疫力量、用遗传算法优化拟合参数，结果显示目前趋势截止至笔者写作当日（2020年2月21日），相疾病的扩散规模当于模型发病前40天左右的趋势。

我们使用遗传算法，通过对武汉市从1月26日起大约20天的疫情数据进行拟合，优化得到模型参数。我们的模型可以与到目前为止的数据较好地拟合，对之后一周武汉市累计确诊患者人数的预测误差率大约为20%，通过模型分析我们得到大概在疾病传播的60天后，传染率会变得很低，发病人数会在模型第50天左右达到最大值，最终武汉市会有累计6到7万人确诊。

2 研究背景

2019年12月，一场名为新型冠状病毒肺炎的瘟疫开始在中国大陆悄悄传播，直到1月24日武汉市封城，这场才受到公众的关注，各省市也立即出台相应的政策尽力抑制本次疫情的发展，国内外社会各界也向疫情区提供大量的支援和帮助。

疫情前期由于未被社会所重视，所以传播较快，直到1月17日，钟南山院士宣布本次的新型冠状病毒可以人传人，才被社会各界关注并提高了警惕。本次新冠病毒肺炎全国的确诊人数从2020年1月26日的2761人，到2月15日已经增加到68584人，可见本次疫情来势猛烈，日前已经被世界卫生组织定为“国际关注的突发公共卫生事件”。随着社会各界加强防范，抑制了疫情的发展，每日新增患者数已经得到控制。

为了对疫情将来的发展做出准确的预测，以方便人们做出正确的判断和决策，需要借助传染病动力学模型来对本次疫情进行分析。由于本次疫情有潜伏期的存在，我们可以采用带有潜伏期的传染病模型SEIR模型，并且使用机器学习的方法确定模型的参数，来帮助来完成模型。

在本文中，我们提出了增强型传染病模型(Enhanced-SEIR)模型，在这个模型中S表示易感者，E表示潜伏期患者，I表示感染者，R表示移出者（治愈或死亡）。由于模型中的各项参数不确定，我们使用机器学习的方法确定该模型的参数。由于在疫情刚开始的时候未被社会重视，而到1月17日本次疫情被指出可以人传染人，开始受到国家的高度重视，社会各界也开始逐步提高警惕。因此本次疫情刚开始传播时的传染率较低，而随着时间的发展，传染率逐步变低，所以本文的模型中的传染率为一个变量，并且由参数控制。最后由遗传算法通过对二十多天的数据分析，得到模型所有的参数。

3 预备知识

3.1 SIR模型与SEIR模型

SIR模型是是一种传播模型，是传染病模型中最经典的模型，其中S表示易感者，I表示感染者，R表示移出者。

SIR传染病模型，是一种传染病领域经典的数学模型，它是由Kermack与McKendrick在1927年利用用动力学的方法建立的。SIR模型将总人口分为以下三类：易感者(susceptibles)，其数量的比例记为S(t)，表示t时刻未染病但有可能被该类疾病传染

本文使用的关于武汉新冠肺炎数据来源于国家卫健委发布的官方数据，武汉市人口数据来源于最新一次人口普查数据。

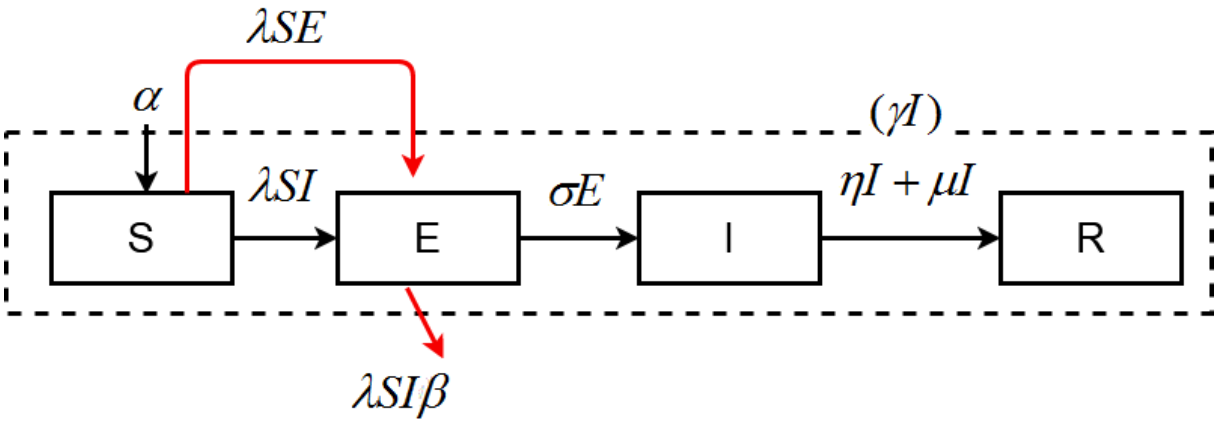

Figure 1: E-SEIR模型示意图。其中红色的线代表对SEIR的主要改进点

的人数占比；染病者(infectives)，其数量的比例记为I(t)，表示t时刻已被感染成为病人而且具有传染力的人数占比；移出者(removed)，其数量比例记为R(t)，表示t时刻已从染病者中移出的人数占比。设所有可能被感染的人口总数比例为N(t)，则有

$$N(t) = S(t) + I(t) + R(t) \tag{1}$$

SIR模型的建立基于以下三个假设：

⑴不考虑人口的出生、死亡、流动等种群动力因素。人口始终保持一个常数，即$N(t) \equiv 1$。也就是实际人口总数保持为K。

⑵一个病人一旦与易感者接触就必然具有一定的传染力。假设 t 时刻单位时间内，假设一个病人能传染给他人的平均概率是λ，则新增感染者数量和当前感染者比例以及易感人群比例有关，从而在t时刻单位时间内被所有病人传染的人数比例为$\lambda S(t)I(t)$。

⑶ t 时刻，单位时间内从染病者中移出的人数与病人数量成正比，比例系数为γ，单位时间内移出者的数量为 $\gamma I(t)$。

但在实际情况中，传染病往往会有潜伏期，在这个期间的病人虽然染病但未表现出症状，而且往往具有传染性，比如本次新冠肺炎就是具有潜伏期的。于是为了将SIR模型适用于具有一定潜伏期的传染病，在传统的SIR模型上添加了潜伏期E，就形成了SEIR模型。它们状态之间的转化大致如 Fifure 1虚线框内所示。

其中，α 表示易感人群的输入或人口增加，μ 表示所有人的死亡率，一旦死亡就从该模型中移除. σ 表示从暴露人群到确诊感染者的比率；η 是感染者的治愈率，感染者一旦被治愈一般不会再次患病。SEIR模型可以由微分方程构建。

$$S = \alpha S \tag{2}$$

$$\frac{dS}{dt} = -\lambda SI \tag{3}$$

$$\frac{dE}{dt} = \beta SI - \sigma E \tag{4}$$

$$\frac{dI}{dt} = \sigma E - (\eta + \mu)I \tag{5}$$

$$\frac{dR}{dt} = \gamma I \tag{6}$$

3.2 遗传算法

遗传算法作为进化算法中最成熟的一类算法，现在已经被广泛应用于科研和实际生活中。该算法模拟的是达尔文生物进化论中的自然选择和遗传学机理的生物进化过程，通过模拟自然进化过程进而搜索最优解的方法。

众所周知，达尔文生物进化论中生物进化的单位是种群，而一个种群则由一定数目的个体组成。每个个体的各种性状都由位于染色体上的基因控制。因此，在一开始需要实现从表现型到基因型的映射即编码工作。由于仿照基因编码的工作很复杂，我们往往进行简化，如二进制编码，初代种群产生之后，按照适者生存和优胜劣汰的原理，逐代演化产生出越来越好的近似解，在每一代，根据问题域中个体的适应度大小选择个体，并借助于自然遗传学的遗传算子进行组合交叉和变异，产生出代表新的解集的种群。这个过程将导致种群像自然进化一样的后生代种群比前代更加适应于环境，末代种群中的最优个体经过解码，即可作为问题近似最

Table 1: 武汉市2月13日到2月19日实际确诊人数与预测确诊人数对比

日期	2.13	2.14	2.15	2.16	2.17	2.18	2.19
实际现存感染者确诊人数	17360	30035	32952	34256	35304	36336	37118
预测现存感染者人数	20541	22578	24587	26541	28418	30196	31861

优解。

遗传算法的具体流程为：

1、初始化，将进化代数置为0，并且设置最大代数，并且随机地生成初始的种群。

2、个体评价，计算每个个体的适应度。

3、选择，基于适应度优胜劣汰地从种群中选择若干个体。

4、交叉变异，被选择的个体交换基因，基因随机地突变，进而生成新的个体。

5、重复步骤2到5，直到到达最大迭代的代数。

4 模型建立

4.1 E-SEIR传染病模型的建立

为了对疫情的发展做出测，我们对本次疫情建立模型来做出分析，由于新型冠状病毒肺炎有潜伏期，并且潜伏期内具有传染性，且过去的动力学模型较少能把变化的传染率这一要素考虑，因此我们提出了一种增强型SEIR模型。我们仍然以S表示易感者，E表示潜伏期患者，I表示发病期的感染者，R表示移出者。瘟疫开始流行的前几天，社会并没有对本次疫情很好地重视起来，导致初始几天的传染率较高，随着时间的发展，传染率逐步降低。因此，我们用公式定义了变化的传染率λ。

$$\lambda = 1 - \frac{1}{1 + e^{-\omega(t-d)}} \tag{7}$$

在这个公式中，如图所示，随着时间t的增长传染率λ会越来越低，传染率由两个参数 ω 和 d 控制，这两个参数都由接下来的遗传算法拟合数据得到。患者一旦确诊，那么有很大概率会被隔离，我们设隔离率为 β，在本文中我们认为已知为0.8，假设潜伏期和发病期间的传染率是一样的，那么每天能感染其他人的数量为 $I(1-\beta)+E$ ，进而每天减少的的易感人群的数量，应该等于每天新增的被感染而进入潜伏期的人数，为感染率与易感人群数量以及有感染能力的人数之积。

$$\frac{dS}{dt} = \lambda S(I(1-\beta)+E) \tag{8}$$

那么每天新增的潜伏期患者的人数，就应该是每天新增的进入潜伏期的人数减去每天发病的人数。本次非冠病毒肺炎的平均潜伏期是一周，那么每天大概有1/7的潜伏期患者进入发病期，设为σ平均潜伏期的倒数，那么我们得到每天新增的潜伏期患者的人数的计算公式。

$$\frac{dE}{dt} = \lambda S(I(1-\beta)+E) - \sigma E \tag{9}$$

每天新增的发病期的人数，应该为由潜伏期到发病期的人数减去移除（治愈或死亡）的人数，设移除率为γ，那么每天被移除的人数为γI，进而我们得到每天新增的发病期的人数的计算公式。

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{10}$$

每天移除的患者人数则只和当天现存患者以及移除率有关系。

$$\frac{dR}{dt} = \gamma I \tag{11}$$

Table 2: 遗传算法拟合的模型的参数值

模型参数	-
移除率γ	2.38 %
传染率曲线参数ω	0.1004
传染率曲线参数d	10.55
时间参数	13

4.2 遗传算法参数优化

我们的模型已经建立完成，我们需要用遗传算法估算的参数有：控制传染率的两个参数ω和d、移除率γ、以及初始的天数。初始天数未知，是因为我们得到了从1月26起的数据，但是之前疾病就已经存在了，我们并不知道我们获得的1月26日数据是模型的第几天。我们的训练集为武汉市1月26日

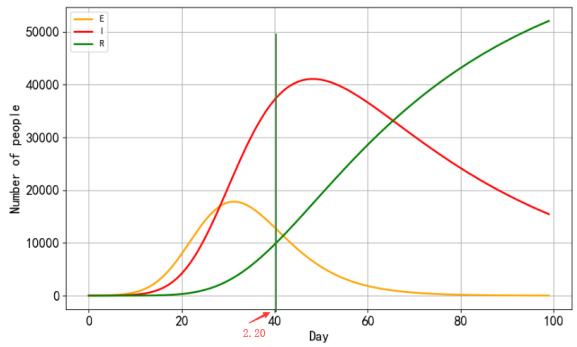

Figure 2: 模型的预情总体发展趋势。I代表武汉市现存感染者数量，E代表潜伏期患者人数，R代表治愈和死亡人数之和

到2月12日的确诊、治愈和死亡人数等各项数据，测试集为2月13日到2月19日的数据。

1、初始化，我们采用二进制编码的方式。我们定义了种群个体数目为400，最大遗传代数为100，生产种群染色体矩阵Chrom，对第一代种群进行解码，计算个体的目标函数值，计算当代最优个体序号。

2、个体评价，我们的遗传算法，先计算多天内预测的感染人数和移除人数与实际的感染人数和移除人数差，再将各天之和相加，越小的值说明拟合的越好，有更大概率遗传下去。

3、个体选择，我们采用“轮盘赌”选择方法。这是一种回放式随机采样方法。每个个体进入下一代的概率等于它的适应度值与整个种群中个体适应度值和的比例

4、交叉变异，采用两点交叉方式，交叉概率为0.7，即每一代有0.3的父代计入下次选择。

5、重复上述步骤，计算每一代的目标函数值，并且选择，直到最大代数

目标函数 我们由国家卫健委获取的数据包含武汉市实时更新的累计感染者人数、治愈人数和死亡人数。我们需要让模型同时拟合现存感染者人数（I）和移除者人数（R）。因此我们设定遗传算法的函数如下：

$$L = L_2(R' - R) + L_2(I' - I) \tag{12}$$

其中R'和I'分别为移除者和感染者的预测值，R和I分别为移除者和感染者的官方数据值。我们把它们的残差二范数之和作为我们的目标函数。

5 实验

从表Table 2可以看出，根据我们的模型，武汉市每天大概会有2.38%的确诊患者治愈或死亡。

Figure 3是根据模型拟合的和两个参数的值，绘制出传染率随时间的发展曲线，我们可以看出前几天传染率非常高，随着时间增长，传染率会逐渐降低，并且在到60天左右时降低到一个非常低的值。

通过对图 Figure 2的分析，我们可以看出前30天武汉市潜伏期患者的数量在不断增长，之后会缓慢下降，发病期的患者会在第50天左右时到达最大，之后会缓慢下降，预计当时全武汉市会有4万多人发病，累计大概有6万人确诊。

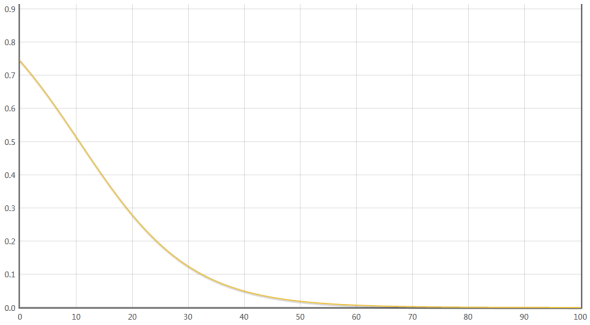

Figure 3: 传染率随时间曲线图

模型评价 时间参数为13天，表示在这个模型中1月26日为疫情开始的第13天，这个结果与实际不太符合，因为我们知道疫情最早在12月就出现了，这是当时对确诊的要求较高导致的确诊人数较少，而直到1月17日才改变了确诊的标准，使得1月17日以来确诊人数增长较快，于是模型出现了一定的误差。

通过表Table 1，我们可以得到实际累计确诊人数与预测的累计确诊人数发展趋势一致，我们取2月13日到2月19日这7天的数据作为我们的测试集，计算一周内的感染者和移除者平均误差都在20%以内。

讨论分析 现在还处在疫情的关键期，现实中很多因素也没有办法估计。本模型是一种数学上简化的模型，在实际中，潜伏期患者和发病期患者的传染率可能相差很大。鉴于面对新病毒时检测手段有限等情况，官方通报的数据也未必能反映实际的发病人数。随着抗疫的推进，治愈率预期也会有提升变化。因此本模型面对现实情况还有很多的改进空间。但是通过分析模型，的确能够量化各决策对于病情控制的效果，也是本文的意义所在。