卫生信息化国际发展动态

北京市卫生健康委信息中心　　徐健 • 2020.11.25

（七）机器学习

1. **标题：** 利用机器学习预测 COVID-19 爆发

来源： medRxiv

时间： 2020 年 4 月 17 日

链接： https://doi.org/10.1101/2020.04.17.20070094

概要：

获得准确疫情预测模型对于深入了解传染病可能传播和后果至关重要。政府和其他立法机构可依靠预测模型提出新政策并评估政策执行的有效性。COVID-19 全球大流行预测模型中，简单的流行病学模型和统计模型越来越受权威机构重视。但是由于疫情的非线性、复杂性，模型周期短、缺乏基础数据，标准流行病学模型面临着提供更可靠结果的新挑战。尽管文献中有一些解决问题的尝试，但现有模型的基本通用性和健壮性仍需改进。于是，利用机器学习(ML)方法建立疫情预测模型近年来得到广泛关注。ML 方法目标是开发具有较高泛化能力的模型，并在更长周期下具有更强预测可靠性。ML 已用于之前流行病(如埃博拉、霍乱、猪瘟、H1N1 流感、登革热、寨卡病毒、牡蛎诺如病毒)预测建模，但针对 COVID-19 的同行评议论文在文献上仍是空白。本文重点探索 ML 在 COVID-19 大流行建模中应用，目的是研究 ML 模型的泛化能力和不同周期的准确性。本文比较分析了进化算法、遗传算法、粒子群优化算法、沃尔夫优化器、机器学习、多层感知器、自适应神经模糊推理系统在预测 COVID-19 疫情上的应用。数据来自于中国、意大利、法国、匈牙利和西班牙五国。本文对 ML 模型和软计算模型在预测 COVID-19 疫情中的应用进行了比较分析。

两种 ML 模型(MLP 和 ANFIS)的结果揭示了长期预测的高度泛化能力。针对本文报道的结果，鉴于 COVID-19 疫情的高度复杂性及各国差异，本研究建议 ML 作为模拟疫情的有效工具。

2. **标题：通过机器学习预测美国 COVID-19 病例和死亡**

来源： medRxiv

时间： 2020 月 8 月

链接： https://doi.org/10.1101/2020.08.13.20174631

概要：

COVID-19 已成为美国等许多国家的重大国家安全问题。随着美国和国际上 COVID-19 病例和死亡总数快速增加，预测冠状病毒传播趋势的模型变得越来越重要。传统流行病学模型更多关注不同类型 COVID-19 预测，但对 COVID-19 预测简单自回归模型的关注很少，而不是优化一个单一模型。本研究的目标是测试一个自回归机器学习模型，与传统流行病学模型相比，现状及预测优化。

这些国家的卫生部门的公共政策和都寄希望于基于未来 COVID-19 死亡和病例预测模型。COVID-19 最常用的模型是流行病学模型和高斯曲线拟合模型，但最近文献表明，这些模型可通过机器学习进行实时验证。但是这些 2019 冠状病毒预测机器学习模型的研究重点是提供一系列不同类型机器学习模型，而不是优化一个单一模型。本研究提出并优化了一个带有梯度优化器的线性机器学习模型，用于预测美国未来 COVID-19 病例和死亡情况。经过比较研究和测试，本研究建议将用于较短范围预测的机器学习模型与用于较长范围预测的高斯曲线拟合或流行病学模型相结合，可以大大提高 COVID-19 预测的准确性。

译文一：

利用机器学习预测COVID-19爆发

Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, Peter M. Atkinson，徐健（译）

摘要

世界各国政府都在用新冠肺炎疫情预测模型进行信息决策和相关控制措施的实施。COVID-19全球大流行预测标准模型中，简单的流行病学和统计模型越来越受权威机构重视，媒体也欢迎。但是由于高度不确定性和缺乏基础数据，标准模型对长期预测的准确性较低。尽管文献中有一些解决问题的尝试，但现有模型的基本通用性和健壮性仍需改进。本文比较分析了预测COVID-19疫情的机器学习和软计算模型。在调研大量机器学习模型基础上，有两个模型显示出良好结果(即多层感知器，MLP和自适应网络模糊推理系统，ANFIS)。基于本文得出的结果，也鉴于COVID-19疫情高度复杂性和各国行为差异性，研究给出的建议是机器学习是模拟疫情的一个有效工具。

关键词

COVID-19;冠状病毒病;SARS-CoV-2;模型;预测;机器学习

介绍

获得准确疫情预测模型对于深入了解传染病可能的传播和后果至关重要。政府和其他立法机构依靠预测模型的洞察力提出新政策，并评估执行政策的有效性。据报告，在全球范围内，新型冠状病毒病(COVID-19)已感染200多万人，确认死亡超过13.2万人。最近全球COVID-19大流行表现出非线性和复杂性。此外，这次暴发与近期其他暴发不同，这使人们对标准模型提供准确结果的能力产生疑问。除了传播过程中涉及众多已知和未知变量之外，不同地缘政治地区的人口行为的复杂性和遏制策略的差异极大地增加了模型的不确定性。因此，标准流行病学模型面临着提供更可靠结果的新挑战。为了克服这一挑战，出现了许多新模型，这些模型引入一些假设(例如，以宵禁、隔离等形式增加社交距离)。

为了详细说明实施这些假设的有效性，理解标准动态流行病学(例如，易感-感染-康复，SIR)模型是必不可少的，模型策略是围绕传染病是通过接触传播的假设，考虑了三种不同类型的良好混合人群：易感人群(S类)、感染人群(I类)和被排除人群(R类用于康复、产生免疫力、被隔离或死亡的人群)。进一步假设类I将感染传播给类S，其中可能传播的数量与接触者总数成正比。作为时间序列的S类个体数量会经常被用于如下基本微分方程计算:

$$\frac{dS}{dt} = -\alpha SI \quad (1)$$

其中“I”为感染人群，“S”为易感人群，均作分子。“α”表示微分方程中每日增长率，调节易感传染病接触者的数量。由微分方程产生的时间序列中‘S’的值会逐渐减小。首先，我们假设在暴发早期’ S≈1 ’，’I’类人数忽略不计。这样，增量$\frac{dI}{dt}$就变成线性，而’I’类最终可按如下计算:

$$\frac{dI}{dt} = \alpha SI - \beta I \quad (2)$$

其中，β通过量化有能力传播的受感染人数对每日新感染率进行调节。此外，R类表示不受感染扩散影响的个体，计算公式如下:

$$\frac{dR}{dt} = \beta I \quad (3)$$

在被排除组无约束条件下(Eq. 3)，暴发指数增长可计算如下:

$$I(t) \approx I_0 exp\{(\alpha - \beta)\} \quad (4)$$

用公式4对各种传染病暴发进行建模。但由于一些国家，比如中国、意大利、法国、匈牙利和西班牙，采取了严格措施，对易感性进行大幅控制，人们自愿隔离并限制社交，SIR模型在COVID-19疫情预测上就无法呈现预期效果，但对于那些控制措施推迟的国家，如美国，该模型还是能显示相对准确性。通过比较实际确诊数和流行病学模型预测发现用于意大利疫情的传统模型的不准确（图1）。通过加入个体感染重要潜伏期，SEIR模型在提高水痘和寨卡疫情模型准确性上取得进展。SEIR模型假设潜伏期是一个随机变量，与SIR模型类似，有一个无病平衡点。有一点需要注意，当参数在时间上不稳定，SEIR模型将不能很好工作。不稳定的一个关键原因是社会混杂因素(它决定了联系网络)会随时间变化。社会混杂决定了易感个体量的再生数R_0。R_0小于1的传染病会消亡。当它大于1时，它会扩散。封锁前新冠肺炎疫情R_0估计为4出现大流行。封锁措施应将R0降至1以下。SEIR模型难适用于COVID-19的关键原因是混杂因素的不稳定性，这是由推动(逐步)干预措施造成的。

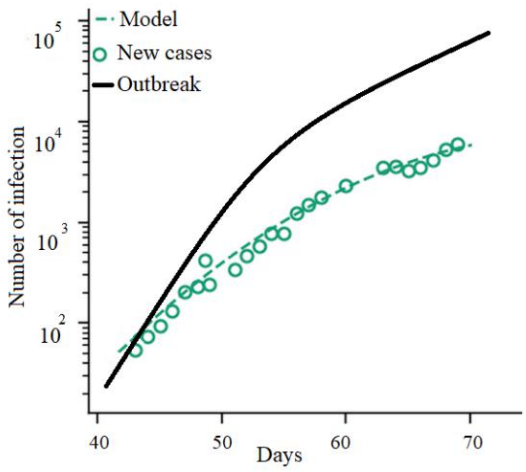

Figure 1. Italy's COVID-19 outbreak: the actual number of confirmed infections vs. epidemiological model.

只有(a)当社会互动在时间上是稳定的(即干预或控制措施没有变化)，或(b)存在大量可以用公式3计算的R类知识，标准的流行病学模型才起效和可靠，我们才可以计算。为了获取R类信息，一些新模型包含了来自社交媒体或呼叫数据记录(CDR)的数据，显示出良好效果。然而，对一些国家的COVID-19行为观察显示出高度不确定性和复杂性。因此，为使流行病学模型能够提供可靠结果，它们必须适应当地情况，并了解对感染的易感性。这极大地限制了常规模型的泛化能力和健壮性。因此，必须提出具有高度泛化能力的精确模型，以便可扩展地对区域和全球流行病进行建模。

人们可以计算出，标准流行病学模型只有在(a)社会互动在时间上是平稳的(即，干预或控制措施没有变化)，或(b)存在大量关于R类的知识，可用来计算Eq. 3。为了获取类R的信息，一些新的模型包含了来自社交媒体或呼叫数据记录(CDR)的数据，显示了良好的效果。但一些国家COVID-19行为显示出高度不确定性和复杂性。因此，为使流行病学模型能提供可靠结果，人们必须采用能洞悉易感的当地数据。这极大地限制了常规模型的普及能力和健壮性。因此，必须提出具有高度普及能力的精确模型，以便可扩展到区域和全球流行中建模。

传统流行病学模型的另一个不足是周期短。为评估模型性能，暴发预测成功中位数呈现出有用信息。中位预测因子的计算方法如下:

$$f = \frac{Prediction}{True\ value}$$

随着周期变长，模型准确性会下降。如意大利COVID-19疫情，后5天以上模型的准确性从前5天的f=1降至第6天的f=0.86。

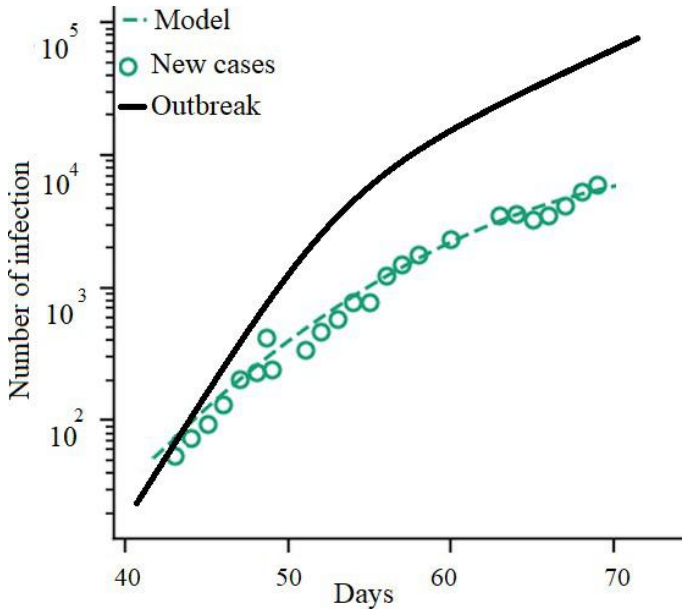

图1　意大利COVID-19疫情:确诊感染的实际数量与流行病学模型

因流行病学模型开发的复杂性和规模性，机器学习(ML)最近在建立疫情预测模型上得到关注。ML方法的目标是开发具有较高泛化能力的模型，并在更长周期下具有更强的预测可靠性。

尽管ML已用于之前流行病(如埃博拉、霍乱、猪瘟、H1N1流感、登革热、寨卡病毒、牡蛎诺如病毒)预测建模，但针对COVID-19的同行评议论文在文献上仍是空白。表1显示了有名的用于暴发预测的ML方法。这些ML方法仅限于随机森林、神经网络、贝叶斯网络、朴素贝叶斯、遗传规划和分类回归树(CART)的基本方法。虽然ML作为自然灾害和天气预报建模的标准工具已建立很久，但它在流行病建模中的应用仍处于早期阶段。更复杂的ML方法(如混合、综合)还有待探索。因此，本文的贡献是探索ML在COVID-19大流行建模中的应用。本文目的是研究ML模型的泛化能力和不同周期的准确性。

表1 有名的暴发预测ML方法

作者	杂志	暴发疫情	机器学习
[39]	跨界和新出现的疾病	猪瘟	随机森林
[35]	地理空间的健康	登革热	神经网络
[42]	BMC研究笔记	流感	随机森林
[41]	公共卫生医学杂志	登革热/伊蚊	贝叶斯网络
[38]	Informatica	登革热	对数递增
[8]	全球生态学和生物地理学	甲型H1N1流感	神经网络
[34]	当前的科学	登革热	采用多元回归和朴素贝叶斯
[36]	国际环境	牡蛎诺瓦克病毒	神经网络
[37]	水的研究	牡蛎诺瓦克病毒	遗传规划
[43]	传染病模型	登革热	分类回归树(CART)

文章其余部分组织如下：第二部分介绍了方法和材料，第三节给出结果，第四节和第五节分别是讨论和结论。

材料和方法

数据收集自https://www.worldometers.info/coronavirus/country五个国家，包括意大利、德国、伊朗、美国和中国，共计30天病例。图2显示了所考虑国家的病例总数(累计统计数字)。目前，为控制疫情，各国政府采取了各种措施，通过限制人们流动和社会活动来减少传播。虽然基于社会距离变化信息的流行病学模型推进非常必要，但是不需要任何假设的机器学习建模也是需要的。从图2中可以看出，在疾病发生的最初几周，中国增长率要高于意大利、伊朗、德国和美国。

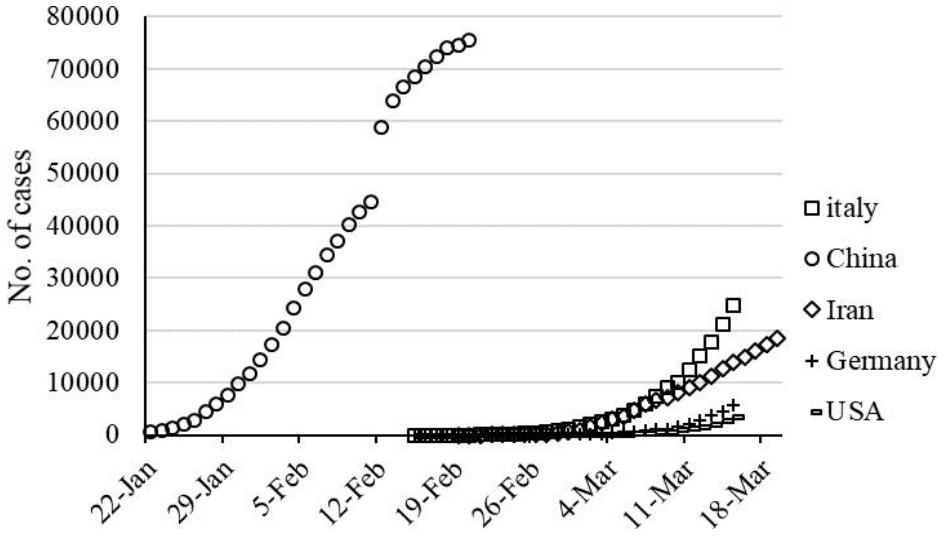

图2　5个国家30天内累计病例数

(https://www.worldometers.info/coronavirus/country)

下一步是建立估计时间序列数据的最佳模型。使用Logistic、线性、对数、二次、三次、复合、幂和指数方程(表2)来建立所需模型。

表2　数学预报模型

Model description	Model name	Equation number
R=A/(1+exp(((4*μ)*(L-x)/A)+2))	Logistic	(6)
R=Ax-B	Linear	(7)
R=A+ Blog(x)	Logarithmic	(8)
$R=A+Bx+Cx^2$	Quadratic	(9)
$R=A+Bx+Cx^2+Dx^3$	Cubic	(10)
$R=AB^x$	Compound	(11)
$R=Ax^B$	Power	(12)
R=AEXP(Bx)	Exponential	(13)

A、B、C、μ和L是描述上述函数的参数(常量)。这些常数需要估值，以形成一个准确的估值模型。这项研究的一个目标是建立基于logistic微生物成长模型的时间序列数据模型。为此，使用逻辑回归的修正方程估算和预测疾病流行率(即在给定时间点的I/人数)作为时间的函数。参数的估算使用进化算法，如GA、粒子群优化器和

沃尔夫优化器。下面将讨论这些算法。

进化算法

进化算法(EA)是通过智能方法解决优化问题的强大工具。这些算法往往受自然过程启发，以寻找作为优化问题的所有可能答案。本文利用常用的遗传算法(GA)、粒子群优化算法(PSO)和沃尔夫优化算法(GWO)通过求解成本函数来估计参数。

遗传算法(GA)

GA是“计算模型”一个子集，灵感来自演化概念。这些算法对“类染色体”数据结构中的特定问题使用“潜在解决方案”或“候选解决方案”或“可能的假设”。GA通过对类染色体数据结构应用“重组操作符”来维护存储在这些染色体数据结构中的重要信息。在很多情况下，GA被用作“函数优化器”算法，即用于优化“目标函数”的算法。当然，使用遗传算法解决问题的应用范围非常广。遗传算法的实施通常要首先产生随机染色体群，并由问题的变量上下绑定。接着对生成的数据结构(染色体)进行评估，能更好显示问题的最佳解决方案的染色体更有可能被用于生成新的染色体。答案的“好”程度通常是由当前候选人的答案数量来衡量。一个遗传算法的主要算法如图3示。

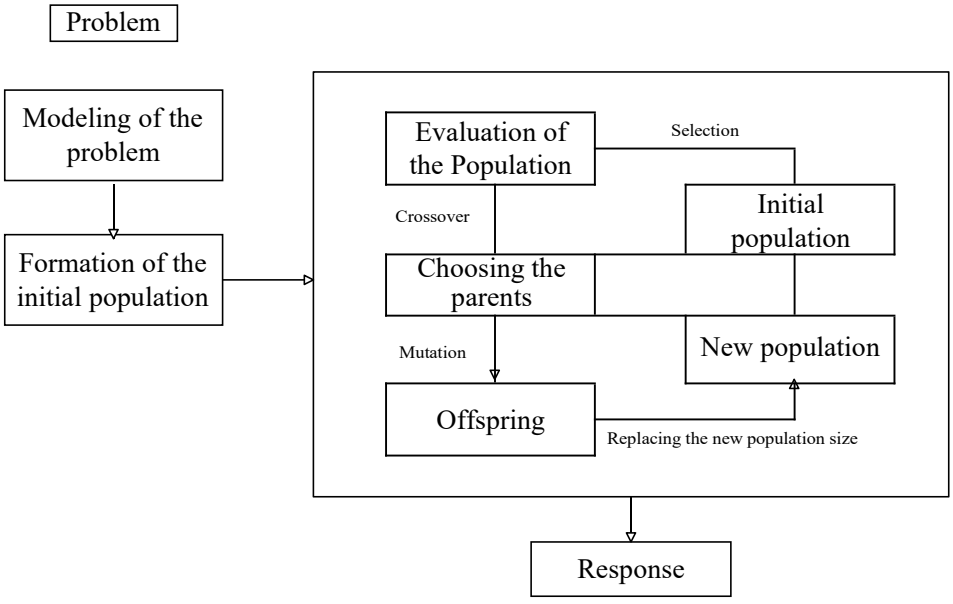

图3. GA 算法

在本研究中，采用遗传算法对等式 6到13的参数进行估值。通过不同的试错过程，选择种群数量为300，并确定最大生成(作为迭代次数)为500，以降低成本函数值。成本函数用公式14定义为目标值与估值间均方误差:

$$MSE = \sqrt{\overline{(Es-T)^2}} \tag{14}$$

式中，Es为估计值，T为目标值，N为数据个数。

粒子群优化算法(PSO)

1995年，Kennedy和Eberhart将粒子群优化算法作为一种不确定搜索方法引入优化。该算法的灵感来自觅食鸟群规模运动。一群鸟在一个空间里临时找食物，在搜寻空间里只有一块食物。粒子群算法中的每个解称为一个粒子，相当于鸟群规模运动算法中的一只鸟。每个粒子都有一个由能力函数计算的值，能力函数随着搜索空间中的粒子接近目标(鸟运动模型中食物)而增加。每个粒子也有一个速度来引导粒子的运动。通过跟踪当前状态下的最优粒子，每个粒子继续在问题空间中移动。PSO方法源于Reynolds的工作，是对鸟类社会行为的早期模拟。自然界中粒子的质量代表着群体智慧。想想鱼在水中的群体运动或鸟在迁徙过程中的群体运动。所有的动物都和睦相处，如果它们要被捕猎，它们就会一起去捕猎；如果它们要被捕食，它们就会移到另一个猎物上，以逃脱捕食者的追捕。该算法中的粒子属性包括:

- 每个粒子独立寻找最佳点
- 每个粒子每一步以同样的速度移动
- 每个粒子记得它在空间中的最佳位置
- 粒子一起工作互相通知他们正在寻找的地方
- 每个粒子接触其相邻的粒子
- 每个粒子都是意识到在附近的粒子
- 每个粒子被称为最好的粒子在其附近

PSO的实施步骤可以总结为:第一步建立初级种群并对其进行评估。第二步确定最好个体记忆和最好集体记忆。第三步更新速度和位置。如果不满足停止的条件，循环回到第二步。

粒子群算法是一种基于种群的算法。这个特性使它不太可能陷入局部最小值。该算法是根据可能的规则而不是确定的规则来操作。因此，粒子群优化算法是一种能够搜索非指定复杂区域的随机优化算法。这使得粒子群优化算法比传统方法更加灵活和持久。由于粒子群算法利用信息结果(性能指标或目标函数)来指导问题区域的搜索，因此粒子群算法可处理非微分目标函数。提出的路由响应的质量不依赖于初始种群。算法从搜索空间的任意位置开始，最终收敛于最优解。粒子群算法在控制局部搜索空间和全局搜索空间的平衡方面具有很大的灵活性。这种独特的粒子群算法克服了算法的非收敛性，提高了算法的搜索容量。所有这些特征使得粒子群算法不同于遗传算法和其他创新算法。

在本研究中，PSO被用来估算公式6到13的参数。通过不同的试错过程，选择种群数量为1000，迭代次数为500，根据公式14降低成本函数值。成本数定义为目标值与估值间的均方误差。

沃尔夫优化器(GWO)

沃尔夫算法是最近开发的一种智能优化算法，引起许多研究者关注。像大多数其他智能算法一样，GWO受大自然启发。沃尔夫算法的主要思想是基于狼群领导等级及它们如何狩猎。一般来说，在灰狼群中有四种狼:阿尔法、贝塔、德尔塔和欧米伽。阿尔法狼位于牧群领导金字塔的顶端，其余的狼听从阿尔法的命令并跟随他们(通常每个牧群中只有一只阿尔法狼)。贝塔狼在较低的等级，但是他们相对于德尔塔狼和欧米伽狼的优势让他们可以为阿尔法狼提供建议和帮助。贝塔狼负责根据阿尔法的移动来调节和定向兽群。德尔塔狼，是狼群中权力金字塔线上下一个，通常由守卫者、老年人、受伤狼看护者等组成。欧米茄狼也是权力阶层中最弱的。公式15-18用于建模狩猎工具:

$$\vec{D} = |C, \vec{X_p}(t) - \vec{X}(t)| \tag{15}$$

$$X(t+1) = \vec{X_p}(t) - A, \vec{D} \tag{16}$$

$$A = 2a, \vec{r_1} - a \tag{17}$$

$$C = 2\vec{r}_2 \tag{18}$$

其中t为算法循环。A和C是猎物点的向量，X表示灰狼位置。在循环过程中，a从2线性降至0。r1和r2是随机向量，其中每个元素都可在范围[0.1]内实现。GWO算法流程图如图4所示。

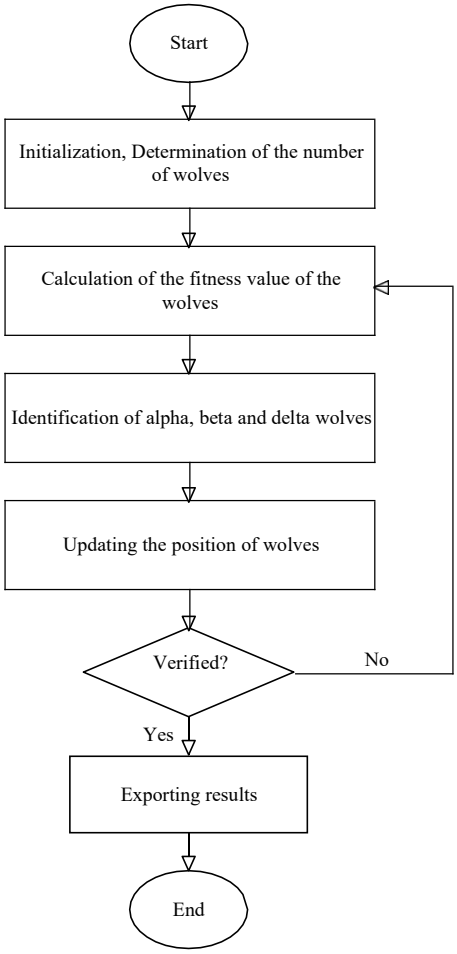

Figure 4. GWO algorithm

在本研究中，我们使用GWO来估计e1 ~ 8参数。通过不同试错，选择种群数量为500，迭代次数为1000，降低成本函数值。成本函数根据公式14定义为目标值与估计值之间的均方误差。

机器学习(ML)

ML被认为是AI的一个子集。使用ML技术，计算机学会使用数据(处理过的信息)中模式或“训练样本”来预测或在没有明显计划的情况下做出明智决定。换句话说，ML是计算机系统使用算法和统计模型的科学研究，这些计算机系统使用模式和推理来执行任务，而不是使用明确的指令。

时间序列是在一段时间内收集的数据序列，可作为ML算法输入。这类数据反映了一种现象在一段时间内发生的变化。设Xt为时间序列向量，其中Xt为时间点t的爆发，t为所有等距时间点的集合。为有效训练ML方法，我们定义了两个场景，如表3所示。

表3 输入和输出变量训练ML方法的时间序列数据

	输入	输入数	输出
Scenario 1	x_{t-1}, x_{t-7}, x_{t-14}, and x_{t-21}	4	x_t (outbreak)
Scenario 2	x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, and x_{t-5}	4	x_t (outbreak)

从表3中可看出，场景1使用三周数据来预测第t天爆发，场景2使用五天爆发数据来预测第t

天的爆发。使用这两种场景来拟合ML方法。本研究采用多层感知器(MLP)和基于自适应网络的模糊推理系统(ANFIS)两种常用的ML方法对五个国家的疫情进行预测。

多层感知器(MLP)

ANN是一个灵感来自生物神经系统的想法，它像大脑一样处理信息。这一思想的关键要素是信息处理系统的新结构。这个系统是由几个高度相互连接的被称为神经元处理元素组成，这些元素通过一起工作来解决问题。ANNs是像人一样通过榜样来学习。神经网络是在学习过程中建立的，用于执行特定任务，如识别模式和分类信息。在生物系统中，学习是由神经间突触连接来调节。这种方法也被用于神经网络。人工神经网络通过对实验数据处理，将数据背后的知识或规律转移到网络结构中，这就是学习。基本上，学习能力是这样一个智能系统最重要的特征。学习系统更灵活、更容易规划，因此可以更好应对过程中新问题和变化。

在ANNs中，编程设计了一个数据结构，可以像神经元一样工作。这种数据结构称为节点。在这个结构中，这些节点间的网络通过具有教育意义的算法进行训练。在这个运行空间或神经网络中，节点有两种活动状态(on或off)和一种不活动状态(off或0)，每条边(突触或节点间连接)都有一个权重。正权值会刺激或激活下一个不活动的节点，负权值会使下一个连接节点(如果是活动的)失活或抑制。在神经网络结构中，对于神经细胞c，输入b_p从前一个细胞p进入细胞。w_{pc}为输入b_p相对于细胞c的权值，ac为各输入与其权值乘积之和:

$$a_c = \sum w_{pc} b_{p\,c} \tag{19}$$

一个非线性函数Θc是应用于a_c。根据b_c可计算出因此，b_c可以用公式20计算出:

$$b_c = \theta_c(a_c) \tag{20}$$

同理，w_{cn}是c到n的输出b_{cn}权重，W是神经网络所有权重集合，对于输入x和输出y，$h_w(x)$ 是神经网络的输出。主要目标是学习这些权值来减少y和$h_w(x)$)之间的误差值。即目标是最小化成本函数Q(W)，公式 21：

$$Q(W) = \frac{1}{2} \sum_{=1} (y_i - o_i)^2 \tag{21}$$

现研究中，一种被称为MLP的再生使用的ANN类被用于预测暴发。MLP用与两种场景相关的数据集进行训练(见表2)。1为网络训练，8、12和16的内神经元要获得最佳反馈。用RMSE及相关系数对结果进行评价，降低成本函数值。图5展示了MLP的体系结构。

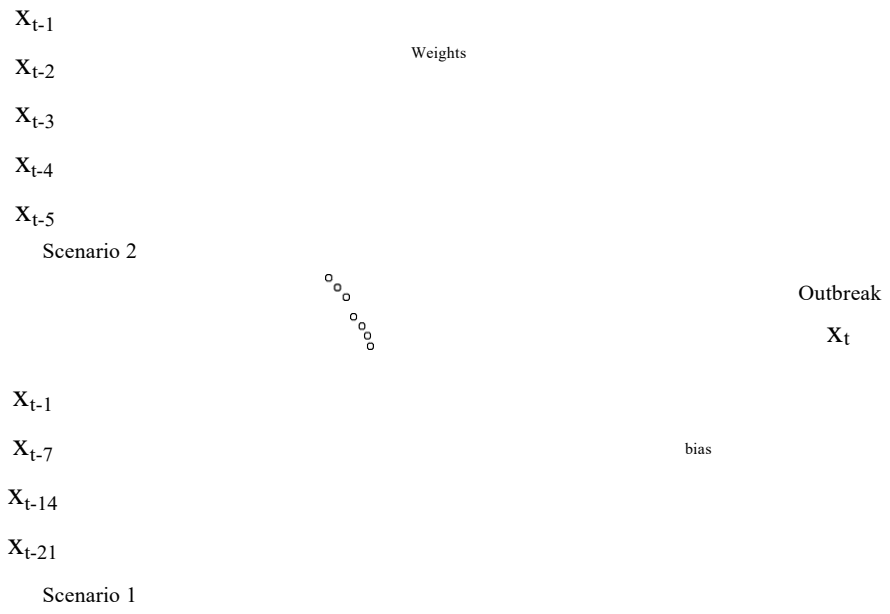

图 5. MLP架构

自适应神经模糊推理系统

自适应神经模糊推理系统是一种基于Takagi-Sugeno模糊系统的神经网络。这种方法是在20世纪90年代早期发展起来的。由于该系统集成了神经网络和模糊逻辑的概念，它可以在一个统一的框架内利用这两种能力。此技术是使用最频繁、最健壮的混合ML技术之一。它与一组模糊if-then规则一致，这些规则可被学习用于近似非线性函数。因此，ANFIS被提出作为一种通用估计量。模糊系统的一个重要元素是输入空间的模糊划分。对于输入k，输入空间中的模糊规则使k面模糊立方体。实现非线性反演的柔性划分是非平凡的。该模型的思想是建立一个神经网络，其输出是属于每个类别的某个程度的输入。该模型的隶属度函数(MFs)可是非线性、多维的，因此不同于传统的模糊系统。在ANFIS中，神经网络被用来提高模糊系统的效率。设计神经网络的方法是采用模糊系统或基于模糊的结构。该模型是一种分而治之的方法。在这个模型中，不是使用一个神经网络来处理所有的输入和输出数据，而是创建了几个网络:

- 模糊分离器在多个类集群输入-输出数据
- 为每个类神经网络
- 训练神经网络输出与输入数据在相应的类中

图6是ANFIS的一个简单架构。

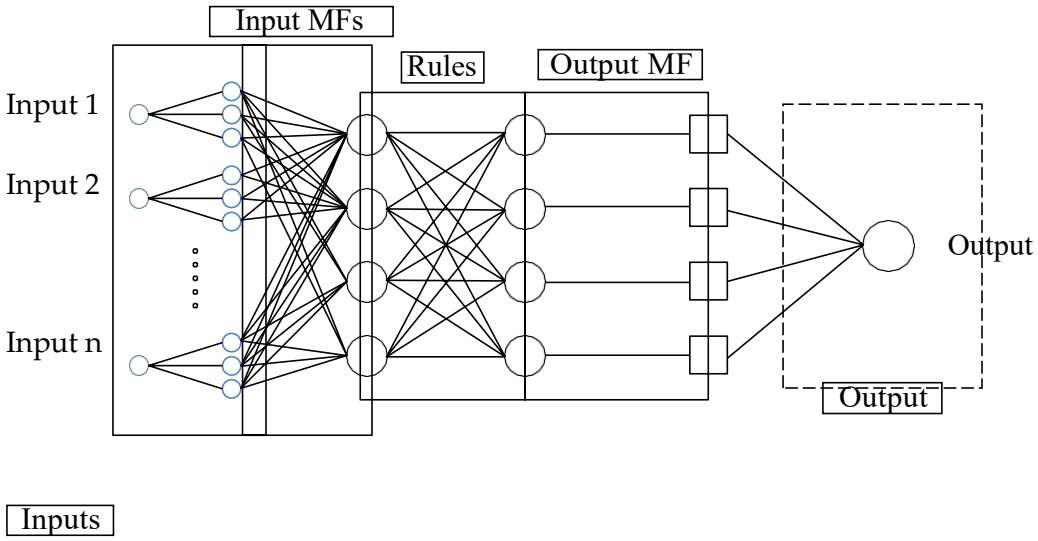

Figure 6. ANFIS架构

在本研究中，开发ANFIS来处理表3中描述的两种情况。每个输入包含Tri模型、Trap模型和Gauss模型的两个MFs。输出MF类型为线性混合优化型。

评估标准

评价采用均方根误差(RMSE)和相关系数。这些统计数据比较目标值和输出值，并计算一个分数，作为衡量所开发方法的性能和准确性的指标。表4给出评价标准方程。

表 4. 模型评价指标

Accuracy and Performance Index	
Correlation coefficient= $\frac{N\sum(AP)-\sum(A)\sum(P)}{\sqrt{[N\sum A^2-(\sum A)^2][N\sum P^2-(\sum AP)^2]}}$	(22)
RMSE= $\sqrt{\frac{1}{}\sum(A-P)^2}$	(23)

其中，N为数据个数，P和A分别为预测(输出)值和期望(目标)值。

结果

表5至表12分别显示了logistic、线性、对数、二次、三次、复合、幂和指数方程的精确统计结果。用三种ML优化器GA、PSO和GWO计算各方程的系数。该表包含国家名称、模型名称、种群大小、迭代次数、处理时间、RMSE和相关系数。

表5. logistic模型的准确性统计

Country	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	82 s	1028.98	0.996
	PSO	1000	500	36 s	3358.1	0.997

	GWO	500	1000	14 s	187.15	0.999
China	GA	300	500	79 s	42160.4	0.982
	PSO	1000	500	35 s	2524.44	0.994
	GWO	500	1000	13 s	2270.58	0.995
Iran	GA	300	500	81 s	1267.04	0.992
	PSO	1000	500	36 s	628.62	0.997
	GWO	500	1000	13 s	392.88	0.996
USA	GA	300	500	82 s	1028.98	0.999
	PSO	1000	500	38 s	350.33	0.999
	GWO	500	1000	15 s	22.35	0.999
Germany	GA	300	500	86 s	5339.5	0.983
	PSO	1000	500	39 s	555.32	0.997
	GWO	500	1000	16 s	55.54	0.999

表6. 准确性统计为线性模型

Country	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	92 s	3774.06	0.845
	PSO	1000	500	42 s	3645.76	0.844
	GWO	500	1000	16 s	3642.44	0.844
China	GA	300	500	91 s	7188.95	0.981
	PSO	1000	500	39s	6644.16	0.982
	GWO	500	1000	14 s	5039.48	0.982
Iran	GA	300	500	96 s	3330.45	0.943
	PSO	1000	500	45 s	2072.71	0.944
	GWO	500	1000	18 s	1981.97	0.944
USA	GA	300	500	88 s	850.22	0.745
	PSO	1000	500	40 s	596.69	0.746
	GWO	500	1000	17 s	592.48	0.746
Germany	GA	300	500	93 s	1118.77	0.758
	PSO	1000	500	47 s	964.46	0.759
	GWO	500	1000	20 s	951.63	0.759

表7. 对数模型的准确性统计

	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	98 s	8325.33	0.634
	PSO	1000	500	51 s	8818.2	0.634
	GWO	500	1000	20 s	9296.59	0.634
China	GA	300	500	96 s	40828.2	0.847
	PSO	1000	500	42 s	43835.37	0.847
	GWO	500	1000	17 s	42714.93	0.847
Iran	GA	300	500	102 s	4929.97	0.757
	PSO	1000	500	59 s	8775.56	0.757
	GWO	500	1000	22 s	8995.52	0.756
USA	GA	300	500	94 s	889.15	0.538
	PSO	1000	500	37 s	1130.33	0.538
	GWO	500	1000	15 s	1135.12	0.538
	GA	300	500	95 s	1552.22	0.548

Germa ny	PSO	1000	500	45 s	1966.81	0.548
	GWO	500	1000	21 s	1878.67	0.548

表8. 对数模型的准确性统计

	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	102 s	6710.01	0.976
	PSO	1000	500	54 s	5102.4	0.953
	GWO	500	1000	26 s	1272.1	0.982
China	GA	300	500	100 s	7921.33	0.992
	PSO	1000	500	46 s	4328.71	0.993
	GWO	500	1000	20 s	3710.16	0.993
Iran	GA	300	500	105 s	6771.74	0.995
	PSO	1000	500	62 s	822.09	0.998
	GWO	500	1000	24 s	310.02	0.998
USA	GA	300	500	98 s	754.6	0.931
	PSO	1000	500	38 s	791.92	0.853
	GWO	500	1000	19 s	307.58	0.938
Germany	GA	300	500	101 s	7577	0.904
	PSO	1000	500	49 s	752.95	0.923
	GWO	500	1000	26 s	472.62	0.946

表9. 对数模型的准确性统计

	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	112 s	7973.11	0.993
	PSO	1000	500	61 s	4827.08	0.996
	GWO	500	1000	34 s	324.33	0.998
China	GA	300	500	113 s	15697.84	0.971
	PSO	1000	500	59 s	3611.15	0.995
	GWO	500	1000	34 s	2429.45	0.995
Iran	GA	300	500	120 s	5852.66	0.995
	PSO	1000	500	88 s	3809.76	0.997
	GWO	500	1000	39 s	250.2	0.999
USA	GA	300	500	110 s	37766.56	0.875
	PSO	1000	500	49 s	678.36	0.979
	GWO	500	1000	25 s	118.24	0.991
Germany	GA	300	500	116 s	1709.06	0.744
	PSO	1000	500	59 s	1812.78	0.967
	GWO	500	1000	29 s	196.8	0.99

表10. 复合模型的准确性统计

	Model	Pop. size	iteration	Processing time	RMSE	Correlation coefficient
Italy	GA	300	500	92 s	8347.51	0.912
	PSO	1000	500	53 s	195705.52	0.918
	GWO	500	1000	22 s	12585.79	0.951

China	GA	300	500	90 s	41544.05	0.986
	PSO	1000	500	48 s	40195.9	0.988
	GWO	500	1000	23 s	24987.34	0.895
Iran	GA	300	500	99 s	1487501.93	0.782
	PSO	1000	500	81 s	8216.81	0.986
	GWO	500	1000	26 s	13635.01	0.864
USA	GA	300	500	96 s	655.62	0.994
	PSO	1000	500	32 s	1026.03	0.827
	GWO	500	1000	16 s	364.87	0.988
Germany	GA	300	500	98 s	15333537.7	0.93
	PSO	1000	500	72 s	1557.23	0.976
	GWO	500	1000	20 s	431.97	0.998

表11. 强模型的准确性统计

	Model	**Pop. size**	**iteration**	**Processing time**	**RMSE**	**Correlation coefficient**
Italy	GA	300	500	72 s	7063.4	0.983
	PSO	1000	500	40 s	6150.52	0.982
	GWO	500	1000	13 s	3450.96	0.991
China	GA	300	500	65 s	39669.92	0.976
	PSO	1000	500	39 s	19365.58	0.987
	GWO	500	1000	12 s	4078.99	0.989
Iran	GA	300	500	83 s	2343032.5	0.951
	PSO	1000	500	65 s	92755.53	0.975
	GWO	500	1000	15 s	1031.6	0.991
USA	GA	300	500	79 s	1030.01	0.779
	PSO	1000	500	24 s	1005.27	0.751
	GWO	500	1000	11 s	790.16	0.837
Germany	GA	300	500	85 s	1475.39	0.871
	PSO	1000	500	69 s	1387.94	0.916
	GWO	500	1000	14 s	1341.91	0.875

表12. 准确性统计指数模型

	Model	**Pop. size**	**iteration**	**Processing time**	**RMSE**	**Correlation coefficient**
Italy	GA	300	500	79 s	8163.1	0.995
	PSO	1000	500	48 s	52075925.37	0.839
	GWO	500	1000	18 s	12585.79	0.951
China	GA	300	500	71 s	68991.73	0.866
	PSO	1000	500	45 s	80104.27	0.865
	GWO	500	1000	17 s	24987.34	0.895
Iran	GA	300	500	89 s	1436025.84	0.767
	PSO	1000	500	70 s	3745673.26	0.744
	GWO	500	1000	21 s	13635.01	0.864
USA	GA	300	500	84 s	457051.4	0.974
	PSO	1000	500	30 s	982.37	0.932

	GWO	500	1000	15 s	364.87	0.988
Germany	GA	300	500	87 s	8176.54	0.981
	PSO	1000	500	74 s	3278.55	0.998
	GWO	500	1000	19 s	431.97	0.998

根据表5-表12，与PSO和GA相比，GWO在拟合所有5个国家的logistic、线性、对数、二次、三次、幂、复合和指数方程时提供了最高的精度(最小的RMSE和最大的相关系数)和最小处理时间。与粒子群算法和遗传算法相比，基于可接受的处理时间，GWO是一种可持续的优化器。因此，与粒子群算法和遗传算法相比，GWO具有最高的优化精度，被选为最佳优化器。总的来说，GWO通过为表2所示的函数提供最佳参数值，可以认为与PSO和GA相比，GWO提高了COVID-19疫情预测的准确性。因此，本研究选择GWO衍生的函数作为最佳预测因子。

表13-表17给出GWO估计的线性、对数、二次、三次、复合、幂、指数和logistic方程的描述和系数。表13-表17还分别给出了中国、意大利、伊朗、德国和美国的数据所拟合的每个方程的RMSE和r-square值。

表13. GWO适合中国的模型描述

Model name	Description	RMSE	r-square
Linear	$R = 3036{,}4 \times x - 13509{,}84$	5039.48	0.964
Logarithmic	$R = -33948{,}15 + 27124{,}70 \times \log(x)$	42714.93	0.718
Quadratic	$R = -5080{,}88 + 1455{,}98 \times x + 50{,}98 \times x^2$	3710.16	0.98
Cubic	$R = 3984{,}73 - 1790{,}2 \times x + 308{,}52 \times x^2 - 5{,}53 \times x^3$	2429.45	0.99
Compound	$R = 1601{,}03 \times 1.16^x$	24987.34	0.801
Power	$R = 262{,}27 \times x^{1{,}69}$	4078.99	0.98
Exponential	$R = 1601{,}03 \times EXP(0{,}15 \times x)$	24987.34	0.801
Logistic	$R = 85011{,}297/(1 + EXP(((4 \times 4483{,}304) * (9{,}423 - x)/85011{,}297) + 2))$	2270.58	0.992

表14. 适合意大利的GWO模型描述

Model name	Description	RMSE	r-square
Linear	$R = 663{,}71 \times x - 5437{,}25$	3642.44	0.713
Logarithmic	$R = -7997{,}93 + 5162{,}83 \times \log(x)$	9296.59	0.402
Quadratic	$R = 2998{,}21 - 917{,}93 \times x + 51{,}02 \times x^2$	1272.1	0.965

Cubic	$R = -978{,}55 + 506{,}05 \times B2 - 61{,}95 \times x^2 + 2{,}42 \times x^3$	324.33	0.997
Compound	$R = 2{,}78 \times 1{,}406^x$	12585.79	0.904
Power	$R = 0{,}096 \times x^{3{,}476}$	3450.96	0.984
Exponential	$R = 2{,}786 \times EXP(0{,}341 \times x)$	12585.79	0.904
Logistic	$R = 70731{,}084/(1 + EXP(((4 \times 3962{,}88) \times (23{,}88 - x)/70731{,}08) + 2))$	187.15	0.999

表15. GWO拟合的伊朗模型描述

Model name	Description	RMSE	r-square
Linear	$R = 656{,}068 \times x - 4527{,}69$	1981.97	0.891
Logarithmic	$R = -7921{,}009 + 5449{,}784 \times \log(x)$	8995.52	0.574
Quadratic	$R = 310{,}48 - 251{,}09 \times x + 29{,}26 \times x^2$	310.027	0.997
Cubic	$R = 902{,}33 - 463{,}02 \times x + 46{,}07 \times x^2 - 0{,}36 \times x^3$	250.204	0.998
Compound	$R = 13{,}26 \times 1{,}33^x$	13635.014	0.748
Power	$R = 0{,}51 \times x^{3{,}09}$	1031.607	0.982
Exponential	$R = 13{,}26 \times EXP(0{,}28 \times x)$	13635.014	0.748
Logistic	$R = 21936{,}052/(1 + EXP(((4 * 1255{,}36) \times (14{,}66 - x)/21936{,}052) + 2))$	392.88	0.996

表16. GWO适合德国的模型描述

Model name	Description	RMSE	r-square
Linear	$R = 128{,}421 \times x - 1130{,}294$	951.635	0.577
Logarithmic	$R = -1528{,}684 + 959{,}941 \times \log(x)$	1878.672	0.3
Quadratic	$R = 911{,}113 - 254{,}342 \times x + 12{,}347 \times x^2$	472.624	0.895
Cubic	$R = -478{,}087 + 243{,}097 \times x - 27{,}118 \times x^2 + 0{,}848 \times x^3$	196.809	0.981
Compound	$R = 3{,}821 \times 1{,}263^x$	431.975	0.996

Power	$R = 0{,}937x^{2{,}021}$	1341.911	0.766
Exponential	$R = 3{,}821 \times EXP(0{,}233 \times x)$	431.975	0.996
Logistic	$R = 55179{,}669/(1 + EXP(((4 \times 3740{,}457) \times (30{,}49 - x)/55179{,}669) + 2))$	55.546	0.998

表17. 适用于美国的GWO模型描述

Model name	Description	RMSE	r-square
Linear	$R = 76{,}833 \times x - 666{,}79$	592.486	0.557
Logarithmic	$R = -902{,}637 + 573{,}32 \times \log(x)$	1135.124	0.289
Quadratic	$R = 584{,}76 - 157{,}831 \times x + 7{,}569 \times x^2$	307.585	0.88
Cubic	$R = -333{,}235 + 170{,}881 \times x - 18{,}509 \times x^2 + 0{,}56 \times x^3$	118.247	0.982
Compound	$R = 6{,}296 \times 1{,}214^x$	364.875	0.977
Power	$R = 1{,}707 \times x^{1{,}735}$	790.163	0.702
Exponential	$R = 6{,}296 \times EXP(0{,}194 \times x)$	364.875	0.977
Logistic	$R = 32604{,}552/(1 + EXP(((4 \times 2288{,}932) \times (30{,}303 - x)/32604{,}552) + 2))$	22.354	0.999

从表13-表17可以看出，一般情况下，logistic方程和二次方程和三次方程为预测COVID-19疫情提供了最小的RMSE和最大的r平方值。从图7 - 表11中也可看出这种主张，图7 - 11展示了利用GWO推导出的各模型对中国、意大利、伊朗、德国和美国COVID-19病例预测的能力和趋势。

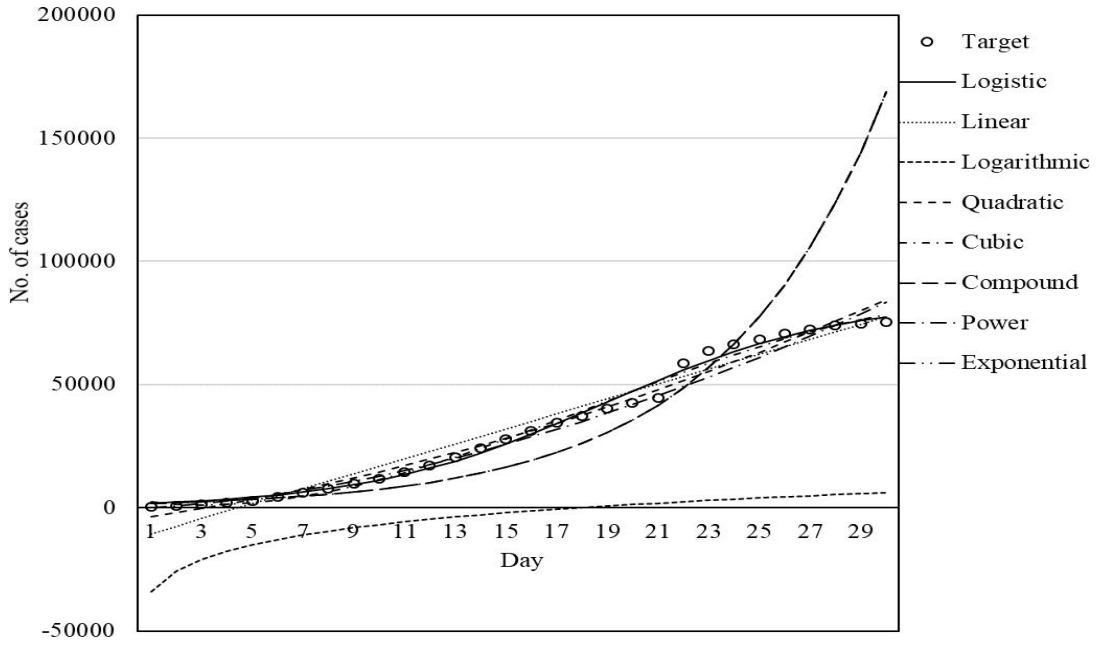

图7. GWO拟合的中国适应度图

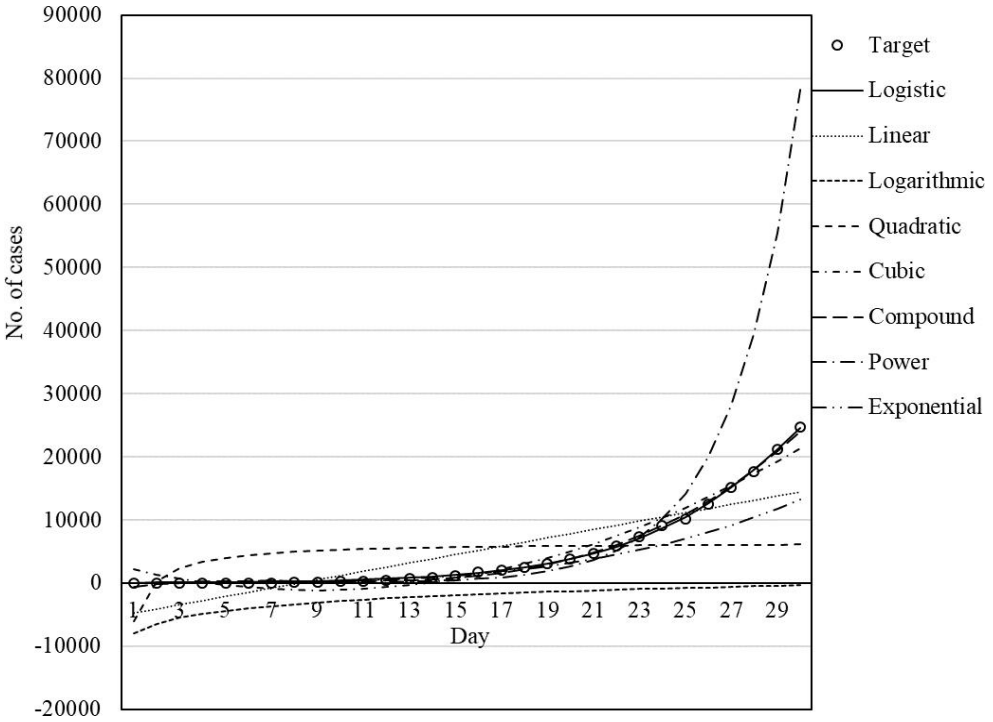

图8. 适合意大利的GWO模型

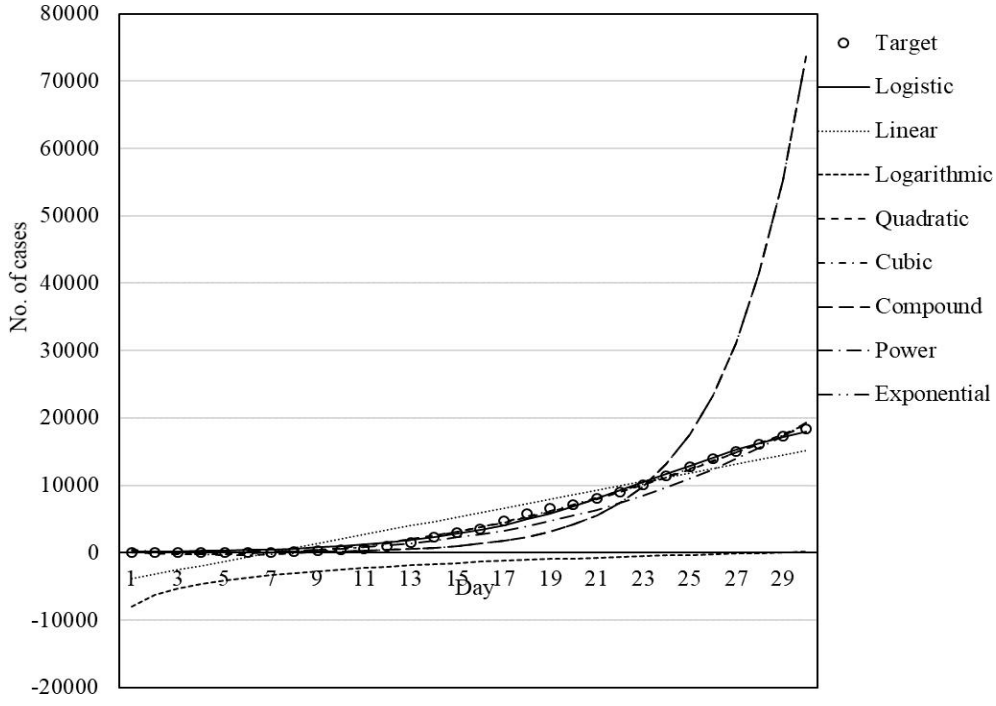

图9. 适合伊朗的GWO模型

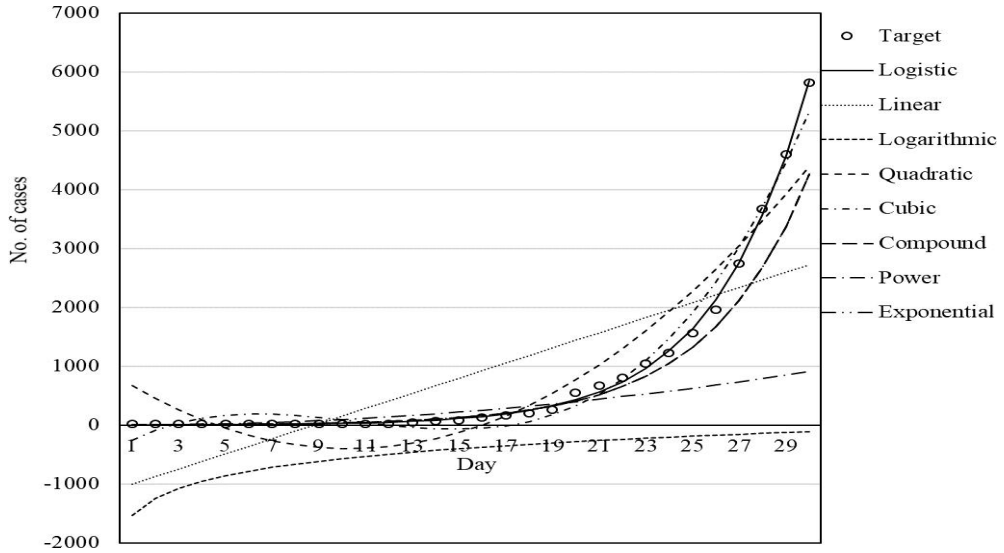

图10. 适合德国的GWO模型

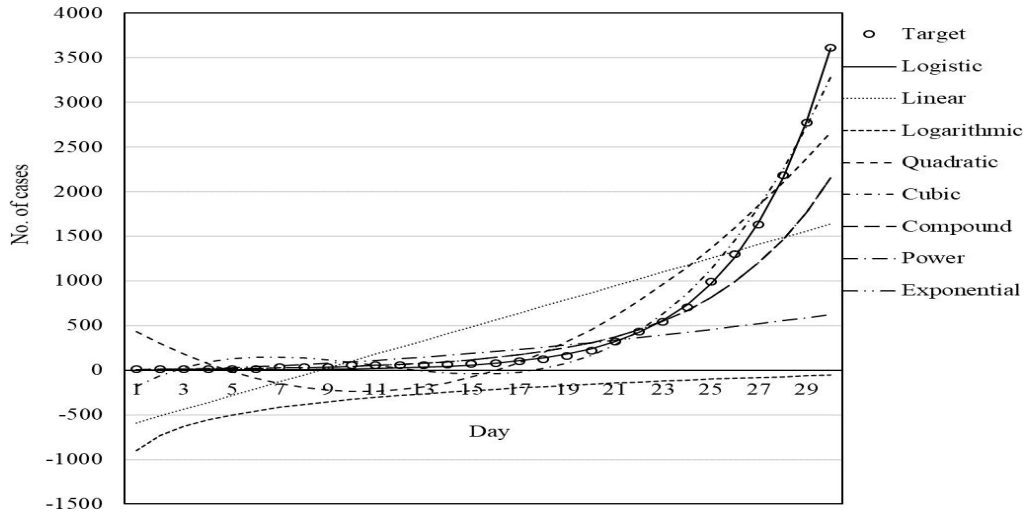

图11. 适合美国的GWO模型

图7-11说明本文所研究模型的拟合情况。最适合COVID-19情况下的预测是中国(图7)的立方和二次建模后的逻辑模型 ，意大利(图8)卡方模型后逻辑模型，伊朗(图9)逻辑和二次模型后卡方模型，德国逻辑模型(图10)及美国逻辑模型(图11)。

机器学习结果

本节介绍了ML方法训练阶段结果。MLP和ANFIS分别作为单一和杂合ML方法。使用场景1和场景2相关的两个数据集训练ML方法。表14展示了培训阶段的结果。

表18. 结果:ML法训练阶段

	Scenario 1						Scenario2					
	MLP			ANFIS			MLP			ANFIS		
	No. of neurons	r	RMSE	MF type	r	RMSE	No. of neurons	r	RMSE	MF type	r	RMSE

	8	0.999	190.81	Tri.	0.999	189.76	8	0.999	199.52	**Tri.**	**0.999**	**188.55**
Italy	12	0.999	194.84	Trap.	0.841	3743.63	12	0.999	195.79	Trap.	0.876	3276
	16	**0.999**	**188.18**	Gauss	0.998	320.93	16	0.999	195.2	Gauss	0.999	206.66
Average		**0.999**	**191.27**		**0.946**	**1418.1**	**Average**	**0.999**	**196.83**		**0.958**	**1223.73**
China	8	0.995	2287.55	Tri.	0.996	2293.09	8	0.996	2265.95	Tri.	0.996	2272.13
	12	**0.996**	**2259.95**	Trap.	0.987	4231.05	12	0.996	2285.73	Trap.	0.989	3835.34
	16	0.995	2407.16	Gauss	0.996	2358.3	**16**	**0.996**	**2260.05**	Gauss	0.996	2272.58
Average		**0.995**	**2318.22**		**0.993**	**2960.81**	**Average**	**0.996**	**2270.57**		**0.993**	**2793.35**
	8	0.998	392.17	Tri.	0.998	395.33	8	0.998	404.21	Tri.	0.998	394.04
Iran	**12**	**0.998**	**391.04**	Trap.	0.977	1282.33	12	0.998	392.77	Trap.	0.986	994
	16	0.998	392.19	Gauss	0.998	396.51	16	0.998	395.43	**Gauss**	**0.998**	**391.96**
Average		**0.998**	**391.8**		**0.991**	**391.39**	**Average**	**0.998**	**397.47**		**0.994**	**593.33**
	8	0.999	55.6	Tri.	0.999	56.25	8	0.999	55.58	Tri.	0.999	55.63
Germany	**12**	**0.999**	**55.38**	Trap.	0.12	1658.7	**12**	**0.999**	**55.56**	Trap.	0.13	1537.26
	16	0.999	55.58	Gauss	0.998	154.99	16	0.999	55.56	Gauss	0.999	62.91
Average		**0.999**	**55.52**		**0.705**	**623.31**	**Average**	**0.999**	**55.56**		**0.709**	**551.93**
	8	**0.999**	**21.65**	Tri.	0.999	21.75	8	0.999	22.31	Tri.	0.999	22.52
USA	12	0.999	22.36	Trap.	0.22	861.08	**12**	**0.999**	**22.3**	Trap.	0.2	935.41
	16	0.999	22.31	Gauss	0.998	86.32	16	0.999	22.4	Gauss	0.999	25.03
Average		**0.999**	**22.1**		**0.739**	**323.05**	**Average**	**0.999**	**22.33**		**0.739**	**327.65**

根据表18，与场景1和场景2相关的数据集具有不同性能值。因此，对于意大利，具有16个神经元的MLP对scenario 1和具有tri的ANFIS提供了最高的准确性。MF为场景2提供了最高的精度。通过考虑RMSE和相关系数的平均值，可以得出结论，场景1更适合建模意大利的疫情，因为它提供了比场景2更高的准确性(最小的RMSE和最大的相关系数)。

对于与中国相关的数据集，在两个场景中，MLP在场景1和场景2中分别有12和16个神经元，与ANFIS模型相比，MLP提供了最高的准确性。考虑RMSE和相关系数的平均值，可以得出平均相关系数比情景1大，平均RMSE更小的情景2更适合中国疫情建模。

对于伊朗数据集，场景1隐含层有12个神经元的MLP和场景2高斯MF类型的ANFIS对爆发的预测性能最好。考虑RMSE的平均值和相关系数，可以得出情景1比情景2的性能更好的结论。同时，与ANFIS方法相比，MLP方法具有更高的预测精度。

在德国，隐含层有12个神经元的MLP具有最高的准确性(RMSE最小，相关系数最大)。考虑RMSE平均值和相关系数，可以得出情景1比情景2更适合预测德国疫情的结论。

在美国，在情景1和情景2中，分别有8个和12个神经元的MLP比ANFIS模型提供了更高的准确性(RMSE最小，相关系数最大)。考虑RMSE和相关系数的平均值，可以得出情景1比情景2更适合疫情预测，MLP比ANFIS更适合疫情预测。

图12 – 16所示的模型分别适用于意大利、中国、伊朗、德国和美国。通过对比图12、16

和图7、11可知，GWO拟合的MLP和logistic模型拟合效果优于其他模型。此外，与其他模型相比，ML方法具有更好的性能。

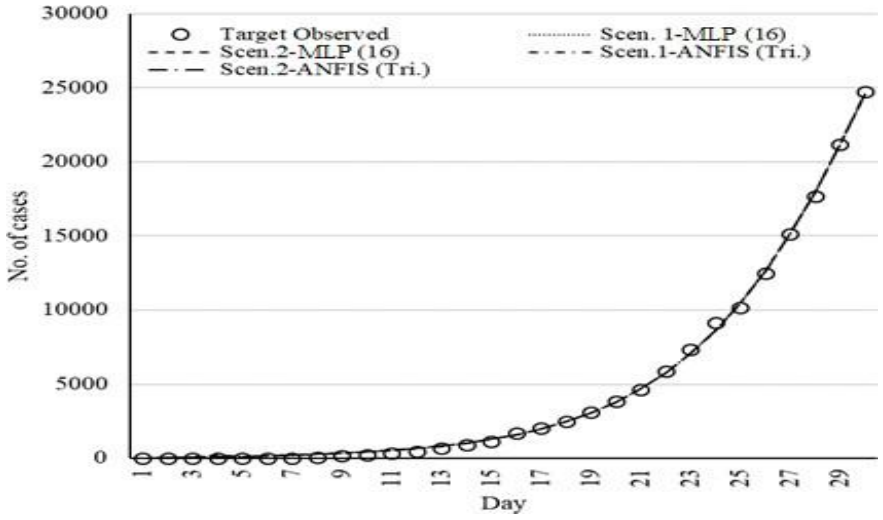

图12. 适合意大利的ML方法模型

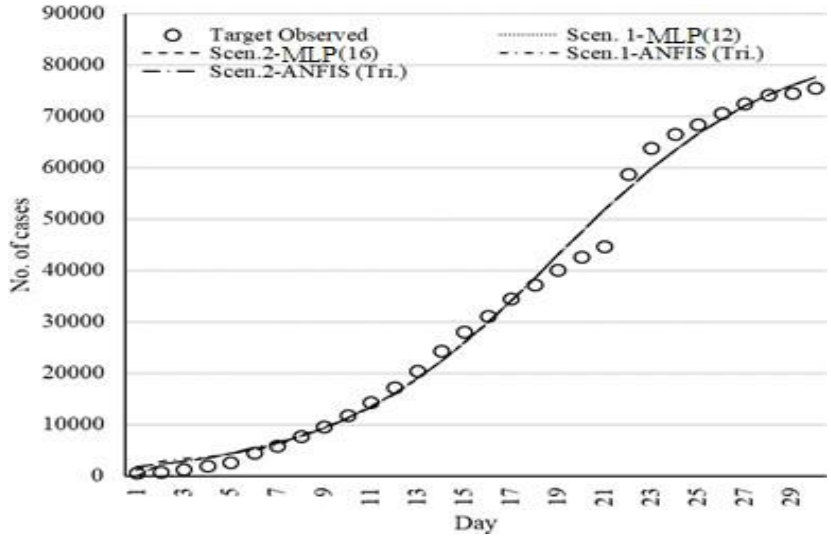

图13.适合中国的ML方法模型

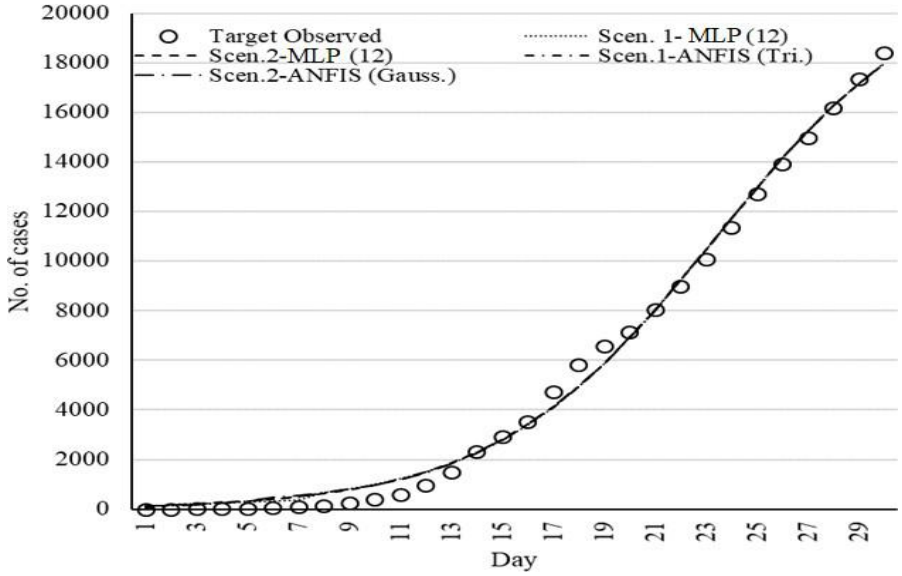

图14. 拟合伊朗的ML方法模型

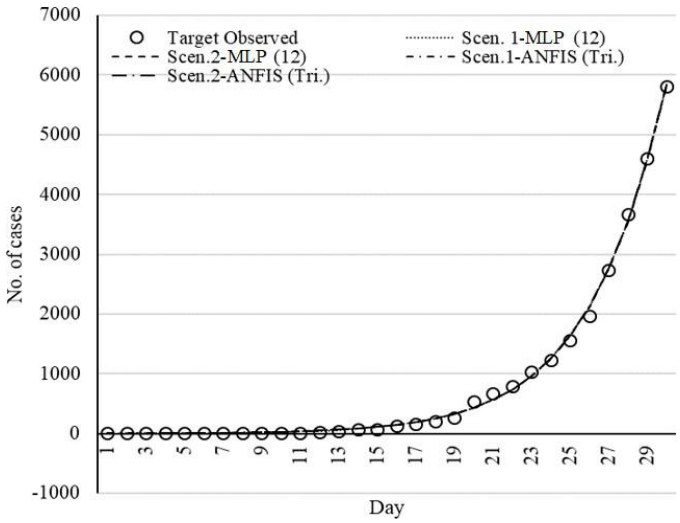

图15. 适合德国的ML方法模型

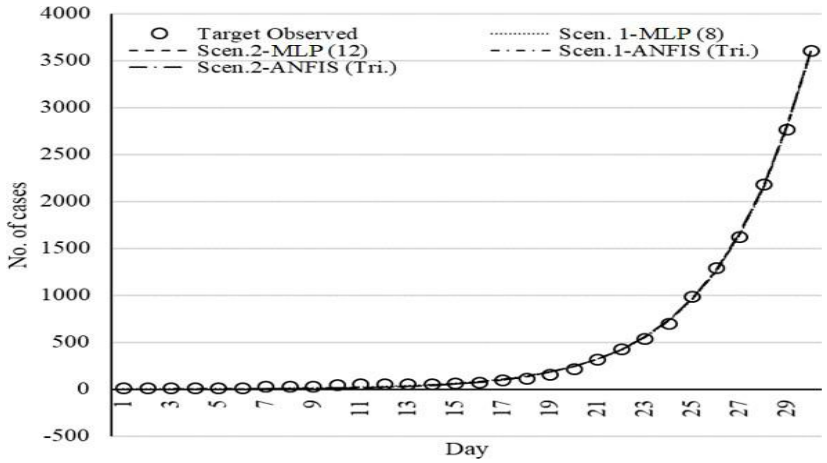

图16. 拟合美国ML方法模型

拟合模型比较

本节将比较所选模型对30天疫情预测的准确性和性能。图17到图21显示了与所选模型的目标值的偏差。

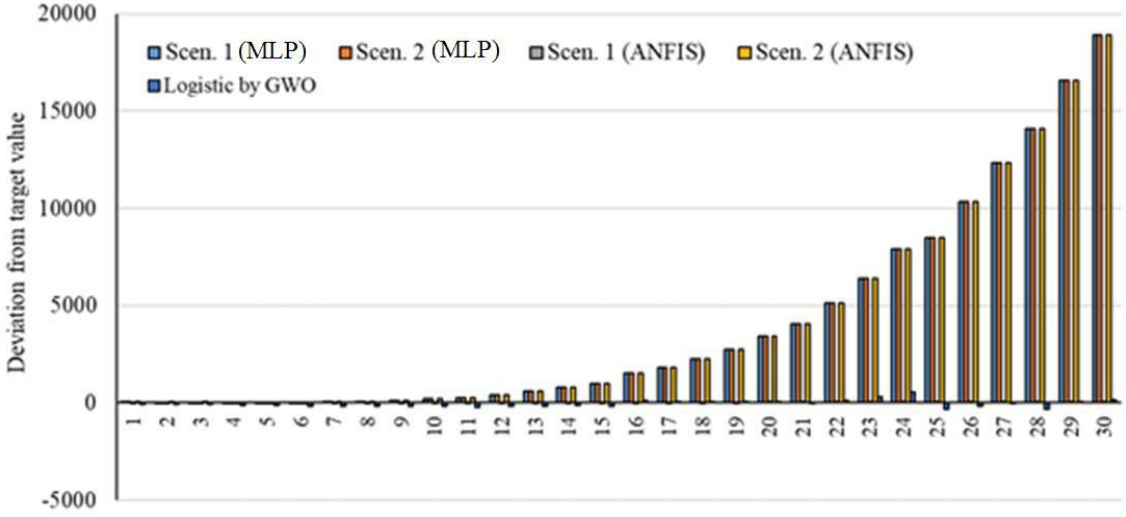

图17. 意大利相关模型的目标偏离值

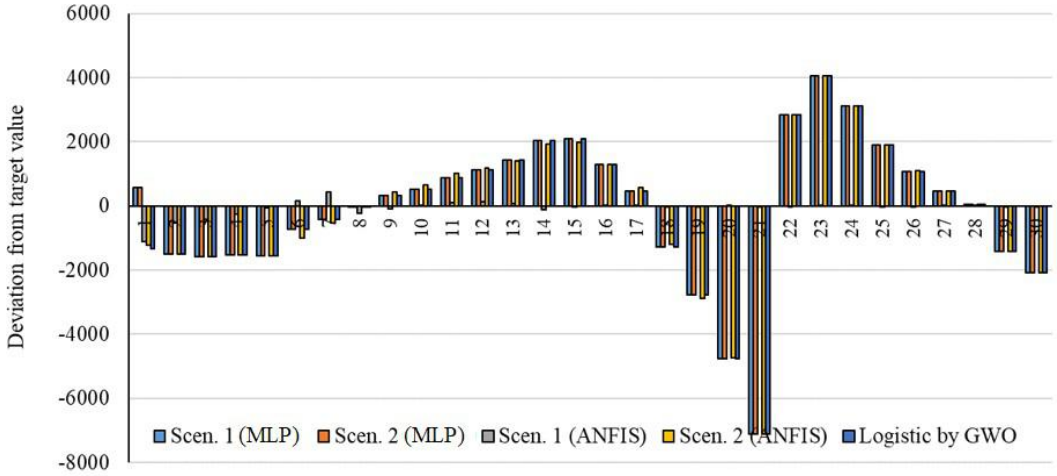

图18. 与中国相关模型的目标值偏离

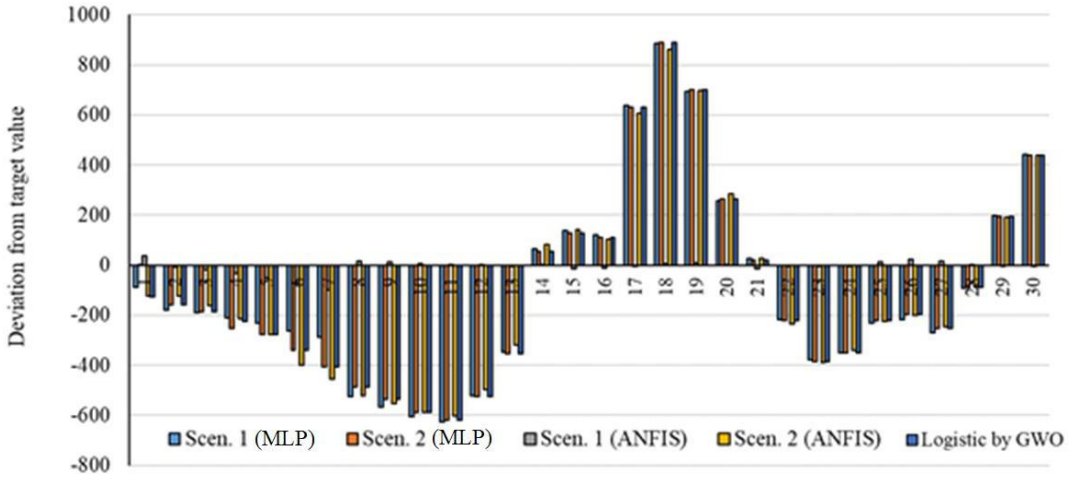

图19.伊朗相关模型目标偏离值

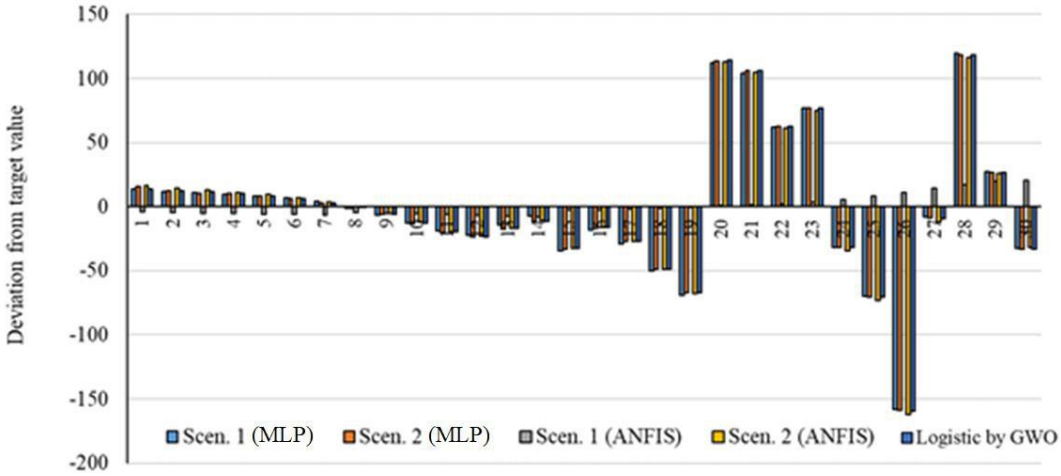

图20. 德国相关模型的目标偏离值

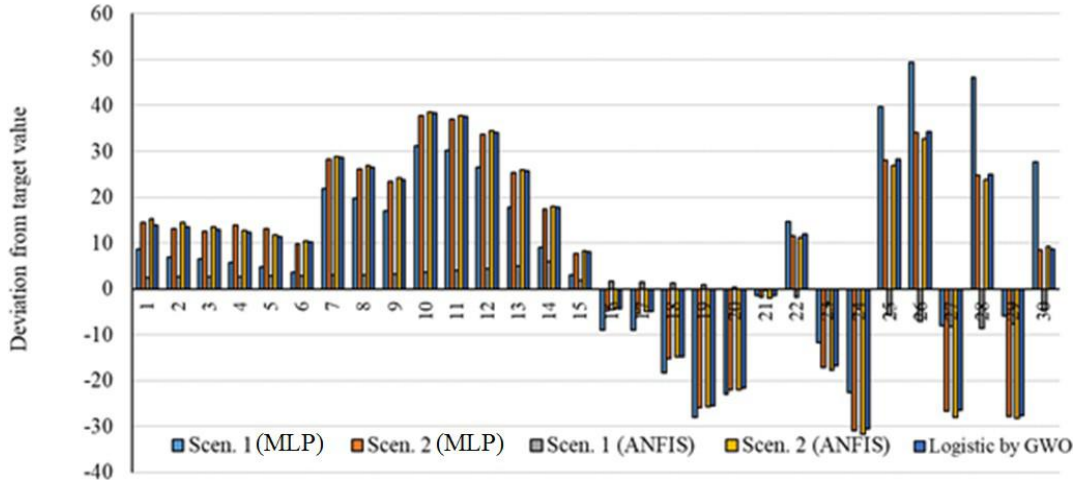

图21. 与美国相关模型目标值偏差

从图17到图21中可清楚看出，与目标偏差最小与场景2的MLP之后的场景1的MLP有关。这表明MLP方法在预测爆发方面的最高性能。

图22至图26显示了75天的爆发预测，表19至23显示了150天的爆发预测。

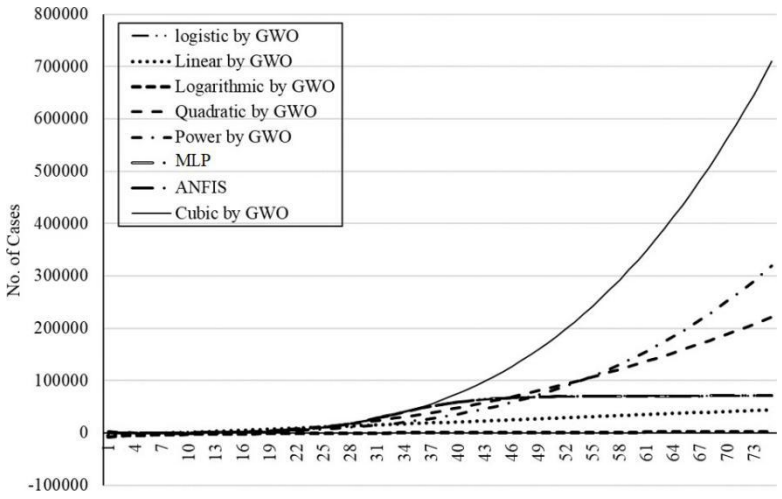

图22. 对意大利75天的疫情预测

表19. 意大利150天内的疫情预测

	Logistic by GWO	Linear by GWO	Logarithmic by GWO	Quadratic by GWO	Power by GWO	MLP	ANFIS
Day 20th	3794.045	7837.054	-1280.93	5047.906	3225.523	3792.734	3796.738
Day 40th	58966.55	21111.37	273.235	47914.4	35898.08	58966.74	58964.96
Day 60th	70571.86	34385.68	1182.365	131597.7	146966.2	70571.66	70572.12
Day 80th	70729.28	47659.99	1827.402	256097.8	399523.4	70729.27	70729.15
Day 100th	70731.06	60934.31	2327.733	421414.7	867822	70731.09	70730.93
Day 120th	70731.08	74208.62	2736.532	627548.4	1635643	70731.14	70730.87

Day 140th	70731.08	87482.94	3082.167	874498.9	2795218	70731.19	70730.79
Day 150th	70731.08	94120.09	3236.862	1013280	3552851	70731.21	70730.75

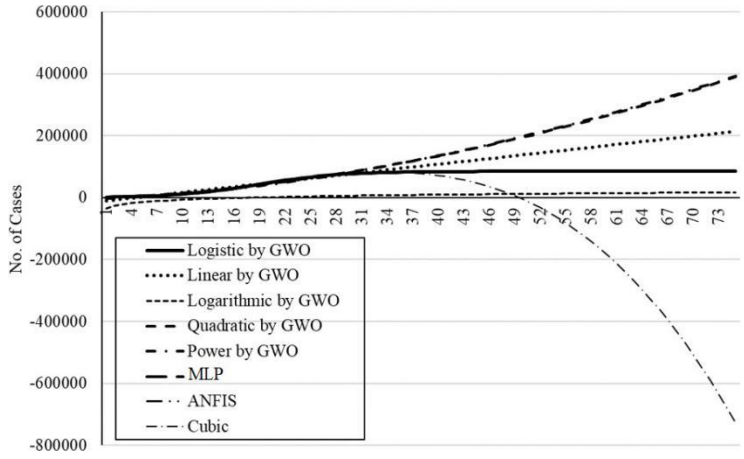

表23. 对中国疫情暴发的预测为150天

表20. 对中国疫情暴发的预测为150天

	Logistic by GWO	Linear by GWO	Logarithmic by GWO	Quadratic by GWO	Power by GWO	MLP	ANFIS
Day 20th	47397.6	47218.47	1341.899	44431.48	41916.55	47397.6	47360.98
Day 40th	84030.16	107946.8	9507.249	134729.1	135599.1	84030.17	84030.39
Day 60th	84996.7	168675.1	14283.67	265812	269471.3	84996.7	84996.67
Day 80th	85011.08	229403.4	17672.6	437680.2	438660.2	85011.08	85011.05
Day 100th	85011.29	290131.7	20301.26	650333.6	640132.8	85011.3	85011.22
Day 120th	85011.3	350860	22449.02	903772.3	871733.6	85011.34	85011.13
Day 140th	85011.3	411588.3	24264.94	1197996	1131815	85011.38	85011.05
Day 150th	85011.3	441952.5	25077.68	1360403	1272113	85011.41	85011.01

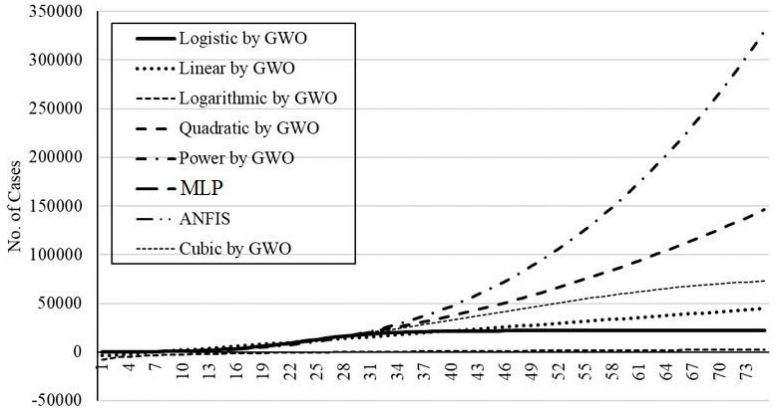

图24. 对伊朗爆发75天的预测

表21. 美国75天内的疫情预测

	Logistic by GWO	Linear by GWO	Logarithmic by GWO	Quadratic by GWO	Power by GWO	MLP	ANFIS
Day 20th	6898.344	8593.676	-830.677	6993.955	5494.377	6902.315	6875.585
Day 40th	21455.58	21715.05	809.8719	37087.98	47060.48	21457.4	21456.65
Day 60th	21931.01	34836.43	1769.531	90592.56	165300.1	21932.24	21930.68
Day 80th	21936	47957.8	2450.42	167507.7	403082.8	21935.1	21935.54
Day 100th	21936.05	61079.18	2978.559	267833.4	804764.4	21935.11	21935.6
Day 120th	21936.05	74200.55	3410.08	391569.6	1415829	21935.12	21935.63
Day 140th	21936.05	87321.93	3774.925	538716.4	2282679	21935.13	21935.65
Day 150th	21936.05	93882.61	3938.219	621068.7	2826737	21935.13	21935.67

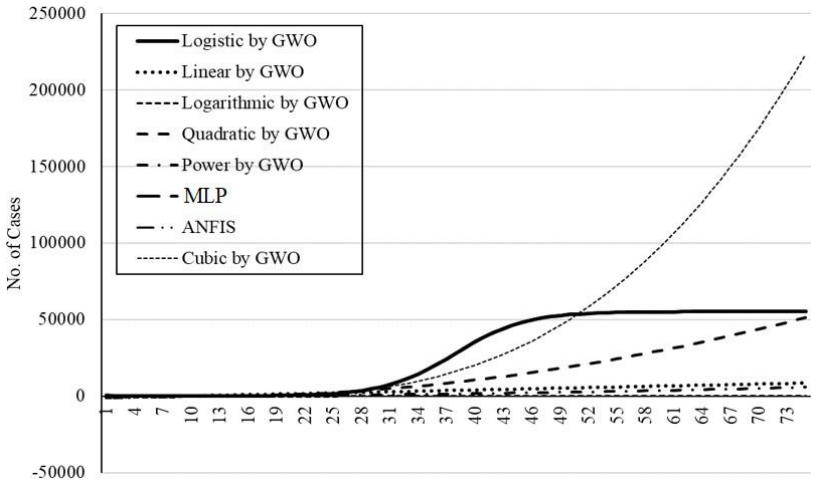

图25. 美国75天内的疫情预测

表 22. 美国75天内疫情预测

	Logistic by GWO	Linear by GWO	Logarithmic by GWO	Quadratic by GWO	Power by GWO	MLP	ANFIS
Day 20th	431.027	1438.128	-279.772	763.1467	400.0548	432.8991	431.8119
Day 40th	35356.27	4006.551	9.199328	10492.96	1624.405	35355.14	35355.72
Day 60th	55043.44	6574.974	178.2366	30100.56	3687.126	55036.14	55044.03
Day 80th	55179.07	9143.397	298.1705	59585.93	6595.829	55179.05	55178.88
Day 100th	55179.67	11711.82	391.1984	98949.09	10355.87	55179.9	55179.47
Day 120th	55179.67	14280.24	467.2078	148190	14971.42	55179.92	55179.42

Day 140th	55179.67	16848.66	531.4728	207308.7	20445.86	55179.94	55179.37
Day 150th	55179.67	18132.88	560.2357	240572.3	23506.09	55179.96	55179.35

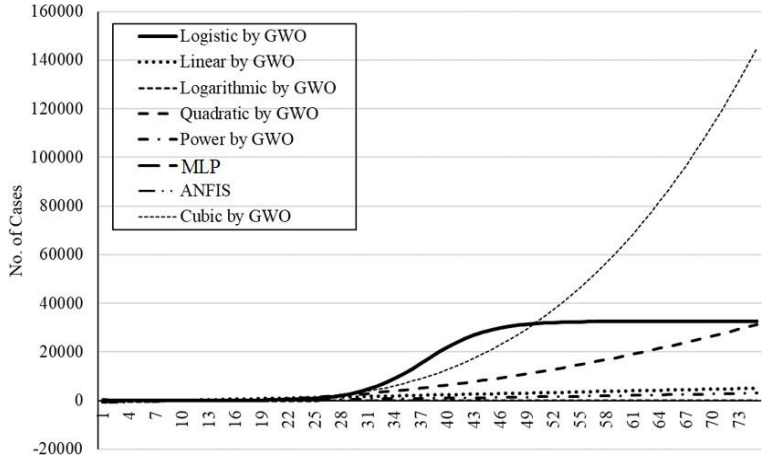

图 26. 美国75天内疫情预测

表23. 对美国爆发150天的预测

	Logistic by GWO	Linear by GWO	Logarithmic by GWO	Quadratic by GWO	Power by GWO	MLP	ANFIS
Day 20th	242.6091	869.8855	-156.73	456.0663	309.616	244.0038	243.6504
Day 40th	21951.15	2406.562	15.85698	6383.264	1031.324	21942.25	21948.25
Day 60th	32547.08	3943.238	116.8138	18366.35	2084.876	32552.6	32548.47
Day 80th	32604.34	5479.914	188.4437	36405.33	3435.319	32606.19	32604.47
Day 100th	32604.55	7016.591	244.0043	60500.21	5060.548	32606.63	32604.72
Day 120th	32604.55	8553.267	289.4005	90650.97	6944.676	32606.7	32604.76
Day 140th	32604.55	10089.94	327.7825	126857.6	9075.446	32606.78	32604.8
Day 150th	32604.55	10858.28	344.9611	147231.9	10230.16	32606.81	32604.82

讨论

利用遗传算法(GA)、粒子群算法(PSO)和GWO拟合了logistic、线性、对数、二次、三次、复合、幂和指数等几个简单数学模型的参数。

logistic模型优于其他方法，并在30天的训练中显示出良好的结果。

考虑到新的统计数据，超出最初30天观测范围的外推预测不应被认为是现实的。拟合模型对五个国家的精度普遍较低，泛化能力较弱。尽管对中国的预测很有希望，但正如预期的那样，该模型不足以进行推断。相应地，逻辑GWO的性能优于粒子群算法和遗传算法，其计算成本也令人满意。因此，为了进一步评价ML模型，我们采用符合GWO的逻辑模型进行比较分析。

接下来，为了引入机器学习方法进行时间序列预测，提出了两种情景。场景1考虑来自前些天疫情的四个数据样本，如表3所示。对于场景1，数据处理的采样每周进行一次。但是，场景2用于对所有连续的前几天进行每日抽样。提供这两种情况扩大了本研究的范围。在这两种情况下考虑了两种机器学习模型(MLP和ANFIS)的训练和测试结果。详细的调查也进行了探索最合适的神经元数目。对于MLP，使用8、12和16个神经元的性能在整个研究中被分析。对于ANFIS，在整个研究中，我们分析了Tri、Trap和Gauss的隶属函数(MF)类型。包括意大利、中国、伊朗、德国和美国这五个国家。这两种ML模型在这些国家的性能在两种不同的场景中有所不同。鉴于观察到的结果，不可能选择最合适的场景。因此，每天和每周的采样都可以用于机器学习建模。使用偏离目标值的分析模型和机器学习模型之间的比较(图17到21)表明，在这两个场景中，MLP提供了最准确的结果。使用ML模型对长达150天的长期预测进行了外推检验。报告了5个国家的MLP和ANFIS的实际预测，显示了疫情的进展。

结论

严重急性呼吸系统综合征2型冠状病毒(SARS-CoV-2)全球大流行已成为许多国家的首要国家安全问题。改进疫情的准确预测模型对于深入了解这一传染病的传播和后果至关重要。由于高度不确定性和缺乏关键数据，标准流行病学模型对长期预测的准确性较低。本文对ML模型和软计算模型在预测COVID-19疫情中的应用进行了比较分析。两种ML模型(MLP和ANFIS)的结果揭示了长期预测的高度泛化能力。针对本文报道的结果，鉴于COVID-19疫情的高度复杂性及各国差异，本研究建议ML作为模拟疫情的有效工具。

为了提高长期预测的性能模型，未来研究应致力于对各国各种ML模型进行比较研究。由于疫情在各国间存在根本差异，推进具有推广能力的全球模型是不可行的。正如许多研究中观察到的和报告的那样，个别疫情不太可能在其他地方重演。

虽然最困难的预测是估计感染患者的最大数量，但估计n(死亡)/ n(感染)也很重要。死亡率对于准确估计重症监护病房的病人数量和所需床位尤其重要。对于未来研究来说，建立死亡率模型对于各国规划新设施至关重要。

术语

Multi-layered perceptron	MLP	Grey wolf optimization	GWO
Adaptive network-based fuzzy inference system	ANFIS	Mean square error	MSE
Susceptible-infected-recovered	SIR	Root mean square error	RMSE
Call data record	CDR	Artificial intelligence	AI
Classification and regression tree	CART	Artificial neural network	ANN
Evolutionary algorithms	EA	Triangular	Tri.
Genetic algorithm	GA	Gaussian	Gauss.
Particle swarm optimization	PSO	Trapezoidal	Trap.

Membership function	MF	Machine learning	ML

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**

译文二：

通过机器学习预测美国COVID-19病例和死亡

Anaiy Somalwar，徐健（译）

摘要

COVID-19已成为美国等许多国家的重大国家安全问题，这些国家的公共政策和卫生保健部门都是基于未来COVID-19死亡和病例预测模型。COVID-19最常用的模型是流行病学模型和高斯曲线拟合模型，但最近文献表明，这些模型可通过机器学习进行实时验证。但是这些2019冠状病毒预测机器学习模型的研究重点是提供一系列不同类型机器学习模型，而不是优化一个单一模型。本研究提出并优化了一个带有梯度优化器的线性机器学习模型，用于预测美国未来COVID-19病例和死亡情况。本研究建议，将用于较短范围预测的机器学习模型与用于较长范围预测的高斯曲线拟合或流行病学模型相结合，可以大大提高COVID-19预测的准确性。

关键字

SARS-CoV-2 COVID-19 Coronavirus 机器学习

介绍

COVID-19是由SARS-CoV-2病毒引起的疾病，在过去六个月里，它已经成为美国超过13万人死亡的原因。在世界卫生组织(WHO)宣布COVID-19为全球卫生紧急事件仅3天之后，2020年2月3日美国政府宣布COVID-19为公共卫生紧急事件。一个多月后，即2020年3月6日，美国宣布新冠肺炎为国家紧急状态，并为新冠肺炎研究提供了额外资金。随着美国和国际上COVID-19病例和死亡总数快速增加，预测冠状病毒增长和下降的模型变得越来越重要。

在对大多数流行病和暴发建模时，预测未来病例数和死亡数的准确性很大程度上取

决于相关的、可获得的数据数量。相关数据数量可用两个度量标准来度量：数据长度(以天为单位)和相关类型数据的宽度。流行与暴发的传统预测模型往往面临同样的挑战：起始数据的小“长度”及包含几个无关类型的数据宽度。即使在疫情持续期间，仍可能有一些对疫情未来病例和死亡的预测特别相关的输入，但这些输入没有得到跟踪或无法向公众提供。

传统流行病学模型多数有一个共同特征，即用于复制疾病在人间传播的方程组。然而,华盛顿大学健康度量和评估研究所(IHME)在2020年3月27日medRxiv预印本上描述他们预测每个州COVID-19死亡和医疗需求时,更多的关注是在不同类型COVID-19预测模型，一个符合COVID-19死亡的假定高斯分布模型。虽然这些类型模型目前都在开发和改进，但对COVID-19预测的简单自回归模型的关注却少得多，甚至是基于这些模型的机器学习也如此。我们研究的目标是检测一个主要是自回归的机器学习模型，与传统流行病学模型比较，这些模型的现状及机器学习模型可否优化2019冠状病毒病预测。

方法

本文侧重于预测美国未来COVID-19病例和死亡，因为根据之前对COVID-19预测的研究，这些指标似乎最相关，也常在医疗保健和政策制定中。我们选用纽约时报Github的covid-19数据库中美国全国数据，因为该数据每天更新，且没有遗漏值。该数据集包含三列：日期、美国COVID-19总病例数和自2020年1月21日美国发现首例COVID-19确诊病例以来美国因COVID-19死亡的总人数。

为预测未来美国COVID-19病例或死亡数，我们假设存在函数f和函数g，分别对应过去某一常数天的COVID-19病例和死亡与第二天或i + 1天的COVID-19病例和死亡。从下面的方程可以看出， c_k代表k索引下病例数，d_k 代表k索引下死亡数，t代表当前日期索引，n和k是为预测未来的病例和死亡数的几天前输入的数。

$$f(c_t, d_t, c_{t-1}, d_{t-1} \ldots c_{t-n+1}, d_{t-n+1}) = \hat{c} \cong c_{t+1} \quad (1)$$

$$g(c_t, d_t, c_{t-1}, d_{t-1} \ldots c_{t-k+1}, d_{t-k+1}) = \hat{d} \cong d_{t+1} \quad (2)$$

为创建这些模型，我们下载并预处理了《纽约时报》github数据，用几个常见的Python库创建了上面描述的输入和输出数组。我们先分出第一个80%按时候排序的数据作为训练数据，供模型“学习”从输入到输出的过程的功能，我们再用剩下的20%数据

测试模型对于它从未见过的数据如何表现。我们发现：当我们将n设为7和k设为5后，有2个交叉验证和12个抑制(Ridge回归)的简单线性回归模型表现最佳。模型可用下列公式描述，其中b_0和b^j_0是偏差，w_1-w_{2n}、 w^j_1-w^j_{2k}是权重。

$$\hat{c} = b_0 + w_1 c_t + w_2 d_t + w_3 c_{t-1} + w_4 d_{t-1} + \ldots + w_{2n-1} c_{t-n+1} + w_{2n} d_{t-n+1} \quad (3)$$

$$\hat{d} = b^j_0 + w^j_1 c_t + w^j_2 d_t + w^j_3 c_{t-1} + w^j_4 d_{t-1} + \ldots + w^j_{2k-1} c_{t-k+1} + w^j_{2k} d_{t-k+1} \quad (4)$$

该模型通过最小化一个损失函数来拟合训练数据，我们选择了均值平方差加上12个调整项，该项旨在引导模型使用相对较小权值来防止过度拟合。下面的损失函数中，yˆ是大小为n的预测值数组,y是大小为n的真实值数组，λ是一个常数，θ是一个权重数组。

$$Loss = (\frac{1}{n})\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \lambda\sum_{i=1}^{n}(\theta_i^2) \quad (5)$$

为预测当前时间t下一个1个未来病例或死亡数，我们递归地使用模型t + 1时间的病例和死亡数的输出作为输入，用于预测t + 2时间的病例或死亡，以此类推。在进行预测时，我们选择了1为7，因为之前关于这个主题的文献中使用了这个数字，这使得这个模型很容易进行比较。从6/27/20开始，我们对测试数据的每个1天间隔做6个1天预测。

结果

我们通过测试数据中预测的平均平方误差(RMSE)的平方根来衡量模型的误差，RMSE表示预测偏离的平均病例数或死亡人数。

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (6)$$

我们还根据它的r2或它的确定系数来测量模型的准确性，它是一种统计方法，用来测量回归与真实值的接近程度。

$$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (7)$$

根据7天预测任务的平均RMSE，预测未来美国COVID-19总死亡的边际回归递归模型的平均RMSE为570.264。基于预测任务的平均RMSE，用于预测未来美国COVID-19病例总数的边际回归递归模型的平均RMSE为4282.341。递归死亡预测因子的r2为0.996，递归病例预测因子的r2为0.992。图1和图2描述了模型预测的准确性。

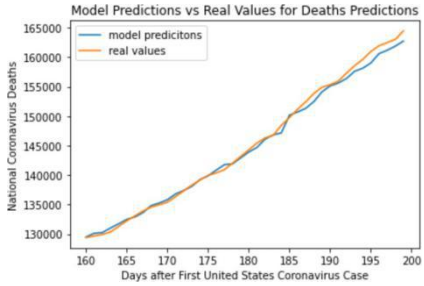

图1:边际回归对美国冠状病毒死亡总人数的预测

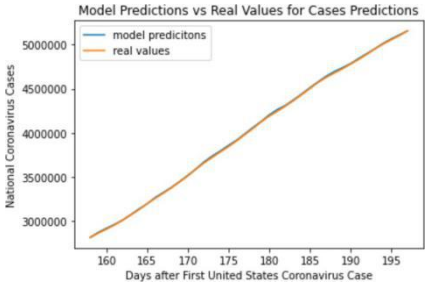

图2:边际回归对美国冠状病毒病例总数的预测

讨论

从对未来COVID-19病例和死亡数预测的r^2值可看出，该模型与两项预测任务的测试数据吻合得很好。

我们将该模型与文献中几种模型比较，即[5]的流行病学模型和[7]的高斯曲线拟合模型。从指数平滑模型的精度统计数据中可看出：在美国，案例预测最低RMSE略低于6000，而在其6天、7天预测任务结束时的平均RMSE是4282.341。边际回归预测是基于最新测试数据，这意味着更多数量的训练数据，美国COVID-19病例总数在大幅增加，这使得较小RMSE在其百分比误差方面更有意义。然而需注意的是，指数平滑模型预测是10天间隔，这是一个更困难任务。尽管如此，模型明显与传统的流行病学指数平滑模型具有竞争力。

既然我们已经知道,它与传统的流行病学模式有竞争性,我们就将它与来自IHME和德克萨斯大学奥斯汀分校COVID-19建模联盟的研究的预测美国COVID-19死亡数的高斯曲线拟合模型进行了比较。正如德克萨斯大学奥斯汀分校COVID-19建模联盟的预印本所描述的那样，他们的模型比IHME的模型表现得好得多，后者被广泛用于医疗保健和政策制定。改进后的模型对美国COVID-19死亡总人数13天的预测的州RMSE均值是13.2，或国家RMSE均值是660。边际回归模型在国家死亡数预测上有较低的绝对RMSE。应该注意的是,

训练数据规模越大的边际回归，在预测区间越小和更简单任务时表现会越好。然而，机器学习模型，如边际回归，无疑在短期预测任务上具有竞争力。

结论

在美国和其他许多国家，COVID-19已造成国家紧急状态和巨大卫生保健问题。虽然本文对现有机器学习模型进行了改进，但很明显，机器学习模型的长期预测仍有改进空间。然而，机器学习似乎是COVID-19建模的一个有竞争力选择，曲线拟合模型和机器学习模型混合似乎是预测COVID-19的希望。

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**