一种基于机器学习和 SIR 模型的基本再生数标定方法

袁润嘉

The Hockaday School

【摘要】基本再生数是判断一种流行病发展趋势的首要指标，对于制定应对疾病的政策具有既根本又直观的参考意义。本文采用 SIR 模型对流行病进行建模，使用美国 COVID-19 的数据对 SIR 模型中的参数进行基于机器学习的反演，并使用反演获得的参数标定基本再生数。在基于机器学习的参数反演中，使用 SIR 模型解变量的实际数据构造了一组新的特征变量，代替模型的解变量本身用作神经网络的输入变量，获得了更为满意的反演结果。由此获得的基本再生数，可以用来对美国 COVID-19 的整体发展趋势进行基本合理的解释。

【关键词】机器学习，SIR 模型，基本再生数

An Algorithm Based on Machine Learning and SIR Model to Determine Basic Reproduction Number

Angela Yuan

The Hockaday School

【**Abstract**】Basic reproduction number is a primary criterion in determining pandemics' developmental trends, an indicator that provides fundamental and direct points of reference for creating policies in response to diseases. This article uses the SIR model for pandemic modeling, analyzes U.S. COVID-19 data to find SIR parameters through machine learning-based backpropagation, and uses those parameters to determine the basic reproduction number. In machine learning-based backpropagation, instead of using the original data, this article uses a new set of features/variables constructed through those data as the input. More accurate results are achieved with this new approach. The basic reproduction number obtained can be used to create reasonable explanations for the general trend of the COVID-19 pandemic.

【**Keywords**】Machine Learning, SIR Model, Basic Reproduction Number

1 引言

1.1 SIR 模型介绍:

SIR 模型是传染病动力学中一个常用的模型，是由 Kermack and McKendrick 提出的。[1] 此数学模型划分为 S，I，R 三个专区，分别指易感人数占总人口比例，感染人数占总人口比例，和治愈人数（包含死亡）占总人口比例。该模型作出以下假设：

(1) 总人数 N 不改变，即忽略自然出生，自然死亡，及人口迁移。

(2) 所有人在最初都对这个传染病无免疫。

(3) 唯一离开易感群组的方式是被感染，唯一离开感染群组的方式是被治愈（或死亡）。

(4) 年龄，性别等因素不影响被感染，被治愈，或死亡的概率。

在以上假设的情况下，SIR 模型使用下面的常微分方程组描述疫情的发展：

$$\frac{ds}{dt} = -bs(t)i(t),$$

$$\frac{di}{dt} = bs(t)i(t)-ki(t),$$

$$\frac{dr}{dt} = ki(t),$$

其中：b 为感染系数，即每个感染者每天传染的人数；

k 为治愈系数，即平均治愈天数的倒数。

1.2 基本再生数/基本传染数(basic reproductive number)介绍:

“基本传染数（Basic reproduction number）是在流行病学上，指在没有外力介入，同时所有人都没有免疫力的情况下，一个感染到某种传染病的人，会把疾病传染给其他多少个人的平均数。[2]基本传染数通常被写成为 R_0。R_0 的数字愈大，代表流行病的控制愈难。在没有防疫的情况下，

- 若 R_0< 1，传染病将会逐渐消失。
- 若 R_0> 1，传染病会以指数方式散布，成为流行病（epidemic）。
- 若 R_0= 1，传染病会变成人口中的地方性流行病。”

评估 R0 的文章很多，但由于估计的方法和数据等不同，得出的结果参差不一。[3]

1.3 本文再生数推导:

由于 SIR 常微分方程组中：

$\frac{di}{dt}$= bs(t)i(t)-ki(t) = i(t)[bs(t)-k]

在疫情开始后 i(t)>0，所以要判断$\frac{di}{dt}$正负仅需比较[bs(t)-k]与 0 的大小，即 bs(t)/k 与 1 的大小。

若 R_0<1，bs(t)/k<1，[bs(t)-k]<0，$\frac{di}{dt}$<0，则感染的人口比例下降。

若 R_0=1，bs(t)/k=1，[bs(t)-k]=0，$\frac{di}{dt}$=0，则感染的人口比例不变。

若 R_0>1，bs(t)/k>1，[bs(t)-k]>0，$\frac{di}{dt}$>0，则感染的人口比例上升。

由于 s(t)近似于 1，则基本再生数可以近似表示为 R_0=b/k。

2 参数反演方法

在老师指导下，假设已知不同参数组 b、k，我们用 Runge-kutta methods 求解 SIR 模型的常微分方程组，得到 N 多组前 180 天 S、I、R 的数据矩阵。再用这些正演得到的解变量作为 features 输入，反演求参数 b、k。结构如下：

	S	I	R	b	k
第 t+1 天					
第 t+2 天					
第 t+3 天					
第 t+4 天					
	features 作为 input			output	

实际中用 keras 人工神经网络库反演的数据，与预知的 b、k 对比，差距忽大忽小，效果不好。究其原因可能是 b、k 对 input features 的依赖关系过于复杂。

此后，我们把原来的 input features 先组成新的 features，组合的方法是基于把常微分方程组离散化降维，给神经网络更多线性依据。离散化，即用有限小的差分代替微分，把微分方程转化为差分方程。

参数表示如下：

$k\approx\frac{r(t+1)-r(t)}{i(t)}$

$b\approx\frac{i(t+1)-i(t)}{s(t)i(t)}+k\frac{1}{s(t)}$

根据上面公式，我们引入了 4 个和 b，k 呈线性关系的 features。用 SIR 模型解变量的实际数据构造了一组新的特征变量，从而使得预测的难度降为线性关系。通过引入中间的多个变量/features，人工地完成一部分工作，使神经网络预测的依据更简单，从而预测结果更准确。

Feature 1: $\frac{r(t+1)-r(t)}{i(t)}$

Feature 2：$\frac{i(t+1)-i(t)}{s(t)i(t)}$

Feature 3：$\frac{1}{s(t)}$

Feature 4: $k\frac{1}{s(t)}$ 或 $\frac{r(t+1)-r(t)}{i(t)s(t)}$

举个例子解释我们这个想法：设 $z=\frac{x}{y}+xy$，当然这是一个关于 x、y 的函数，但是非线性的。如果用一个 z 对 x,y 的线性函数来逼近，效果肯定不好。但假如引入一个$\boldsymbol{\alpha}=\frac{x}{y}$，$\boldsymbol{\beta}=xy$，那么 $z=\boldsymbol{\alpha}+\boldsymbol{\beta}$，是关于$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$的线性表达。此时，把 z 用$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$的线性函数来逼近，效果就比之前的好。我们的新 features 就相当于这个中间变量。

因为我们是选取连续 4 天 SIR 模型解变量的实际数据作为一个时间段，则可以相应得到 13 个 input features，即：

$\frac{r(2)-r(1)}{i(1)}$, $\frac{r(3)-r(2)}{i(2)}$, $\frac{r(4)-r(3)}{i(3)}$,

$\frac{i(2)-i(1)}{s(1)i(1)}$, $\frac{i(3)-i(2)}{s(2)i(2)}$, $\frac{i(4)-i(3)}{s(3)i(3)}$,

$\frac{1}{s(1)}$, $\frac{1}{s(2)}$, $\frac{1}{s(3)}$, $\frac{1}{s(4)}$,

$\frac{r(2)-r(1)}{i(1)s(1)}$, $\frac{r(3)-r(2)}{i(2)s(2)}$, $\frac{r(4)-r(3)}{i(3)s(3)}$

改进后的结构如下：

	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10	feature11	feature12	feature13	k	b	
来源正演数据第一时间段														给定值	给定值	训练数据
来源实际数据第一时间段														预测值	预测值	预测数据

在调试神经网络参数阶段，我们完全使用 SIR 模型解变量的实际数据矩阵，在相同时间段（比如都是第 68 天到第 71 天），让 30 组 input features 和 30 组已知 b、k 一一映射。抽取其中任意 29 组作为神经网络训练数据，另一组为测试数据。再通过神经网络预测的 k、b 和测试组已知的 b、k 进行对比，检验该神经网络预测的准确性。事实证明，SIR 模型解变量的实际数据构造这一组一组新的特征变量，代替模型的解变量本身，用作神经网络的 input features，获得了更为满意的反演结果。

此外，经过上百次测试，我们还发现：在本案例中，从预测精确度考虑，keras 神经网络参数选取 adam 优化器优于 sgd，rms，adadelta，和 adagrad 优化器。迭代次数 20000 次优于 5000 次，10000 次，30000 次，和 50000 次(由于存在过拟合)。学习率 0.001 优于 0.1，0.01，和 0.0001。在实际演算中，选取 5 个神经元的单层隐层兼顾了精确和效率。

3 应用案例分析

训练并检验神经网络后，在应用案例中，测试数据我们选用了美国 Johns Hopkins University 官网观测统计数据。选取的四个时间段分别为：4 月 6 日至 4 月 9 日、5 月 5 日至 5 月 8 日，6 月 24 日至 6 月 27 日，7 月 1 日至 7 月 4 日。每个时间段中，第一天 t=1，第二天 t=2，第三天 t=3，第四天 t=4。我们利用观测统计数据，分别计算了这 4 个时间段的 13 个 input features，并代入 keras 神经网络程序去预测相应阶段的 b 和 k 的数值。

然而，我们依然无法直观比较这 4 个时间段彼此孤立的 k 和 b，于是进一步利用公式 $R_0=b/k$ 求解基本再生数 R_0，并用 R_0 来对美国 COVID-19 的整体发展趋势进行基本合理的解释。

我们对每个月份各自进行了 100 次预测，得到：

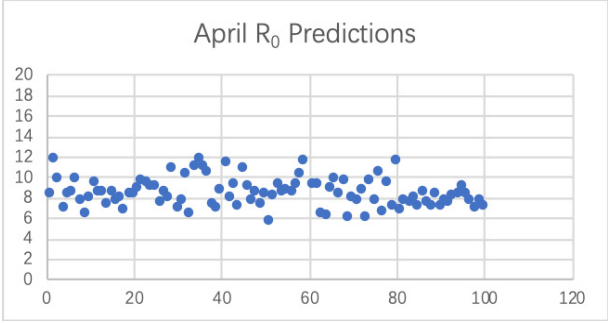
1）4 月 R_0 平均数为 9.131656431（见图一）。

2）5 月 R_0 平均数为 5.430964678（见图二）。

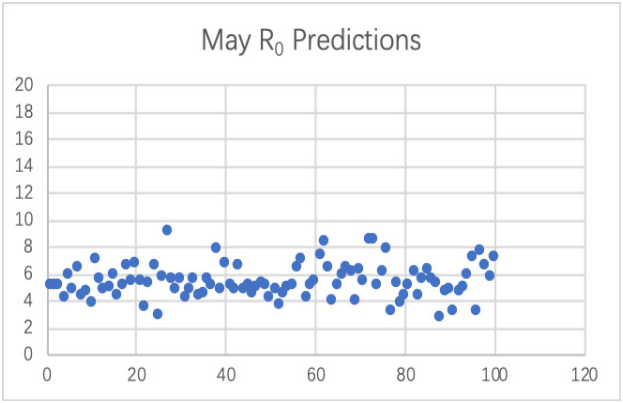
3）6 月 R_0 平均数为 2.795867213（见图三）。

4）7 月预测波动较大，无规律可循。

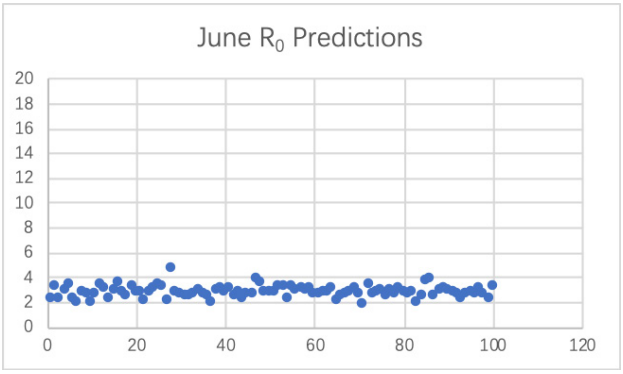
散点图如下：



图一：4 月基本再生数



图二：5 月基本再生数



图三：6 月基本再生数

由此可以看到，数据是比较贴合实际的。4 月份美国新冠疫情非常严重，之后政府实施了“居家令”、公共场所社交距离等防疫措施，5 月份 6 月份 R_0 都降低了。4 月底 5 月初各州相继经济重启（最

晚的纽约州也于 6 月 9 日重启经济）。特别是 6 月结果和帝国理工大学[4]估计的 1.9~4.2 是很接近的。相比之下，埃博拉病毒的基本再生数为 1.5~2.5，重症急性呼吸综合征（SARS）的基本再生数为 0.85~3，甲型 H1N1 流感的基本再生数为 1~2，由此可知，COVID-19 的传染能力是很强的。[5]

美国过早的经济重启、5 月底 70 多个城市发生的示威游行，以及 Memorial Day 纪念日活动等，导致 7 月美国迎来第二波疫情。七月疫情发展偏离了原有 SIR 模型，也导致神经网络预测的 R_0 无规律可循。究其原因，我们认为除了 SIR 模型固有的局限之外，更因为第一波疫情 SIR 模型的长尾效应叠加了第二波疫情的 SIR 模型的头部，使得疫情的模型在 7 月已经偏离了原有的 SIR 模型常微分方程组，基于 Runge-kutta methods 得到的那些正演数据也不再适合作为训练数据使用了。假如想进一步准确预测出疫情未来的走向，必须突破现有的 SEIR 模型（传染病有一定的潜伏期，与病人接触过的健康人并不马上患病，而是成为病原体的携带者，归入 E 类）、SIRS 模型（康复者 R 可能再次变为易感者 S），寻找新的模型。[6]

4 结论

我们发现，使用通过 SIR 模型的原始数据，构造出一些衍生的特征量，对于机器学习的效果有显著的提升，这是本文在方法上的主要贡献。相同的思路显然可以扩展应用于类似的问题，如果采用更为精细而实际的传染病模型，减少观测统计数据上的误差，可能对基本再生数给出更为可信的标定。

参考文献

[1] 马知恩，周义仓，王稳也，等. 传染病动力学的数学建模与研究，北京：科学出版社，2004：第 4-5 页；

[2] 陈嘉敏，邱增钊，钟舒怡，文思敏，舒跃龙. 基于系统综述的 COVID-19 与 2009 年 H1N1 流感大流行基本传染数研究，疾病监测（ISSN 1003-9961,CN 11-2928/R），第 2 页，[2020-11-05]；

[3] Delamater PL, Street EJ, Leslie TF, et al. Complexity of the basic reproduction number (R0) [J]. Emerg InfectDis 2019, 25(1): 1-4

[4] Natsuko Imai, Anne Cori, Ilaria Dorigatti, Marc Baguelin, Christl A. Donnelly, Steven Riley, Neil M. Ferguson (2020). Report 3: Transmissibility of 2019-nCoV. MRC Centre for Global Infectious Disease Analysis. Imperial College London,UK, 2020；

[5] 余锦芬,宋玉凯,费菲,等.基于机器学习和动力学模型的湖北省新型冠状病毒肺炎疫情分析[J/OL].生物医学工程研究:1-15[2020-06-04]；

[6] 石耀霖，SARS 传染扩散的动力学随机模型，科学通报. 2003 年 13 期，第 1373-1377 页。

收稿日期：2020 年 11 月 25 日
出刊日期：2021 年 1 月 21 日
引用本文：袁润嘉，一种基于机器学习和 SIR 模型的基本再生数标定方法[J]. 国际应用数学进展, 2021, 3(1) : 1-4.
DOI: 10.12208/j.aam.20210001
检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊