薛瑞. 基于储备池计算的传染病预测研究与优化[J]. 智能计算机与应用,2024,14(6):177-182. DOI:10.20169/j.issn.2095-2163.240625

基于储备池计算的传染病预测研究与优化

薛瑞

(南京邮电大学集成电路科学与工程学院,南京 210023)

摘 要:以新冠疫情传播为例,利用储备池计算网络建立传染病预测模型。针对算法本身与测试数据的特点,采用一种独特的同步时间差预测方式:使用多数地区的疫情发展数据来预测少数地区的发展趋势。同时使用标准 SIR 传染病模型生成的高维数据对多种层次的储备池计算网络在预测时间线上的误差进行分析,并以减小误差为目的提出一种时变权重储备池计 算网络。结果显示时变权重算法模型相比较于单一类型的储备池可以有效地提高预测精度,在测试数据集上可以将误差降低 45.6%。

关键词:机器学习;储备池计算;传染病动力学模型;组合模型 中图分类号:TP391;R18 文献标志码:A 文章编号:2095-2163(2024)06-0177-06

Research and optimization of infectious disease prediction based on reservoir computing

XUE Rui

(College of Integrated Circuit Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Taking the spread of the COVID-19 as an example, the prediction model of infectious diseases is established by using the classical reservoir computing. Based on the characteristics of the algorithm and test data, a unique synchronous time difference prediction method is adopted: using epidemic development data from most regions to predict the development trend of a few regions. Simultaneously, high-dimensional data generated by the standard SIR infectious disease model is used to analyze the errors of multiple levels of reservoir computing networks on the prediction timeline, and the time-varying weight reservoir computing network is proposed to reduce errors. The results show that the time-varying weight algorithm model can effectively improve prediction accuracy compared to a single type of reservoir computing, and can reduce the error by 45.6% on the test dataset. **Key words**: machine learning; reservoir computing; infectious disease model; composite model

0 引 言

传染病一直是威胁人类生存的一大难题,有效 地预测传染病的发展趋势对于主管部门和个人制定 防控计划有着积极意义。以新冠疫情传播为例,机 器学习在抗击新冠疫情中发挥了重要作用^[1-3]。研 究可知,基于机器学习算法的疫情预测模型,可以预 测病毒传播趋势、疫情高峰期等关键信息,为防控疫 情提供了有力支持。另外,机器学习技术还可以帮 助医疗工作者快速地对病例进行初步分类和诊断, 有助于提高病例的治疗和康复效果。此外,机器学 习还可以帮助监控和管理疫情期间的物资供应链, 及时发现疫情影响下的物资短缺问题。相较于经典 传染病模型对新冠发展的预测,其优势在于可以对 大量的数据进行分析,从而预测病毒的传播趋势,了 解患者的病情和病例,以及预测未来的病例和死亡 率。不仅如此,机器学习也可以进一步提高疾病分 类和治疗计划的准确性,并为疾病研究提供了新的 可能性^[4]。

到目前为止, Akhtar 等学者^[5]提出了一个动态 人工神经网络模型来预测 COVID-19 流行的时间跨 度,该方法同样应用于 2015 年 Zeka 病毒大流行的 预测,结果表明模型有着良好的预测效果。Yan 等 学者^[6]使用 XGBoost 预测模型用于诊断患者由冠状

作者简介: 薛 瑞(1997-),男,硕士研究生,主要研究方向:机器学习,非线性动力学。Email: 1220024701@ njupt.edu.cn 收稿日期: 2023-04-13

病毒引起的疾病,该模型可以快速预测风险最高的 患者,从而对其加以优先考虑并降低死亡率。 Google Flu Trends^[7]对其加以提出了一种使用人工 智能和社交大数据在线搜索来预测实时流感活动的 方法,这种预测模型可以有助于制定更好的策略来 应对新冠病毒的大流行趋势。总之,对于影响全球 的新冠疫情,国内外学者做了大量的工作,这也从侧 面推进了机器学习技术在传染病预测领域的发展。

但是储备池计算(Reservoir Computing, RC),也 被称作回声状态网络(Echo State Network, ESN),作 为神经网络(Neural Network)的一种拓展框架,还没 有很好地应用于传染病的研究。ESN 的出现,极大 地优化了机器学习对非线性动力学的预测,并且解 决了 RNN 难以训练的缺点^[8-10]。ESN 是由 Jaeger^[11]提出的一种新型递归神经网络,其特点是 具有高效的处理非线性动态系统的能力、满足易于 解释的动态神经网络模型以及具有较好的时效性 能。ESN 网络是一种单向的、具有固定权重的神经 网络,可以通过简单的随机映射实现增量式学习,避 免了反向传播算法中的训练过程^[12]。同时,ESN网 络还能够在线适应非平稳环境下的输入,具有较好 的稳定性和鲁棒性。其核心的中间层是随机稀疏连 接的储备池。ESN 只需要对输出权值进行训练,通 过求解简单的线性回归问题而得到,其他权值及阈 值根据特定的要求随机生成,训练过程中不需要再 调整^[13-14],为此受到了广泛的关注。

因此,本文使用 RC 网络建立传染病预测模型, 通过对全球部分地区的疫情数据进行描述性统计分 析,根据疫情的实际发展情况和数据特点,选取真实 疫情数据作为研究对象,从而达到预测疾病发展趋 势的效果。并且对不同层次的 RC 网络的预测结果 进行误差分析,提出一种时变权重储备池计算网络, 并将其应用到地区真实数据的预测中。结果显示, 时变权重模型相比较于单层次的储备池网络具有更 高的精确度。

1 储备池计算结构概述

1.1 经典储备池

RC 算法的功能和形状类似于传统的 RNN 和 LSTM,其在时间序列的预测上有着出众的效果。相 较于传统递归神经网络的隐藏层是一层层的全连接 神经元,RC 引入了一个储备池计算的模式来代替隐 藏层,结构如图 1 所示。



Fig. 1 Reservoir computing structure

由图 1 可以看出, RC 的结构从输入到输出,分别为输入层、具有 N 个动态节点的储备池层和输出层,所谓的储备池就是由一些随机连接的节点构成。

在 t 时刻,储备池层状态向量 r(t) 的更新方程 可表示为:

$$\boldsymbol{r}(t) = (1 - \alpha)\boldsymbol{r}(t - 1) + \alpha \tanh(\boldsymbol{A}\boldsymbol{r}(t - 1) + W_{\text{in}} \overset{\circ}{\boldsymbol{g}} \overset{b_{\text{in}}}{\boldsymbol{g}} \overset{\flat}{\boldsymbol{h}}_{\text{in}} \overset{\flat}{\boldsymbol{y}})$$
(1)

其中, A 为储备池节点之间的随机连接矩阵; u(t) 是通过输入权重矩阵 W_{in} 输入到储备池层的时间序列 标量; α 为泄露率,主要用来控制储备池层各节点之间 权重更新的速度; b_{in} 为偏差,可用常数1带入。

储备池的输出方程可由式(2)来描述:

$$Y(t+1) = W_{out} \hat{\hat{\mathbf{e}}} U(t) \hat{\mathbf{u}}$$
(2)
$$\hat{\hat{\mathbf{e}}} r(t) \hat{\mathbf{u}}$$

其中, W_{out} 表示输出权重, 可以通过岭回归的 方法得到, 此处用到的公式为:

$$\boldsymbol{W}_{\text{out}} = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{\lambda}\boldsymbol{I}\right)^{-1}$$
(3)

其中, λ 表示避免过拟合的正则化因子,本文 取值 1×10⁻⁷, *I* 表示方便运算形式的单位矩阵。

1.2 深度储备池

本文在研究中,也利用到了深度储备池算法 (Deep Reservoir Computing, DRC)。DRC 相对于 RC 的区别在于增加了中间储备池层的深度,并将多 层次结构串联排列,每一层的状态主要依赖前一时 刻的状态和前一层储备池的输入信息,具体结构如 图 2 所示。



假设 DRC 的中间层数为 n,则在 t 时刻,状态向量 **R** 的数学表达式可写为:

$$\boldsymbol{R}(t) = \left[r^{(1)}(t), r^{(2)}(t), \cdots, r^{(n)}(t) \right]$$
(4)

其中,当n = 1时, $r^{(1)}(t)$ 和单层储备池的状态 方程一样;当n > 1时, $r^{(n)}(t)$ 可以用式(5)来确定: $r^{(n)}(t) = (1 - \alpha)r(t - 1) + \alpha \tanh[Ar(t) + t]$

$$V(t) = (1 - \alpha)\mathbf{r}(t - 1) + \alpha \tanh[\mathbf{A}\mathbf{r}(t) + \mathbf{W}_{i}\mathbf{r}(t - 1)]$$
(5)

在输出式的表达上,同样可以使用岭回归的方法进行对输出权重 W_{out} 的训练,这里不再赘述。

2 同步时间差预测思想

本文摒弃 RC 传统的预测思想,不再采用通过 训练数据自我迭代生成预测数据的方式,尝试训练 高维数据与低维数据之间的联系,通过高维数据的 持续输入来达到预测目标数据的目的,具体结构如 图 3 所示。



图 3 同步时间差预测方式

Fig. 3 Synchronous time difference prediction method

在这种方法中,假设有 N 组输入数据,其中 M 组数据被送入预测模型(红色部分),剩下(N - M) 组数据作为目标输出(蓝色部分),二者之间通过预 测模型计算出输出权重W_{out}。其次,M 组数据点的 一部分用于训练过程(红色左上部分),作用是使训 练期间的模型拟合其余维度(N - M)的目标数据 (蓝色左下部分)。

在预测系统训练完成后,将 M 组训练数据剩余的数据点(红色右上部分)依次输入预测模型,利用输出权重 W_{out} 拟合出(N - M) 组目标预测数据(蓝色右下部分)。

值得注意的是,在训练的时候,输入数据和目标 预测数据之间存在时间差t,这部分设置也是为了满 足研究过程中时间差预测的要求。基于该预测思 想,下面以单层储备池为例展开对运算细节的进一 步解释。 在每一个时刻t,输入向量u(t)包含M个元素, 为一组(1 × M)的矩阵:[$I_1(t), I_2(t), \dots, I_M(t)$]^T, 该向量生成的储层状态矩阵如式(6)所示:

)	$W_{in}(1,M)$	À	$\acute{\mathbf{g}}^{I_1(t)}$ ỳ
$\hat{e} W_{in}(2,1)$)	$W_{in}(2,M)$	ú	$\hat{\mathbf{e}}I_2(t)$ ú
e ê :	:	:	u× ú	e u ê∶ú
$\mathbf{\hat{g}}W_{\mathrm{in}}(N_{\mathrm{res}},1)$)	$\boldsymbol{W}_{\mathrm{in}}(N_{\mathrm{res}},M)$) Ý	$\mathbf{\hat{g}}_{M}(t)\mathbf{\acute{h}}$
				(6)

在这个训练过程中,每个时刻t,储备池状态r(t)和输入数据s(t)可以组成的向量形式为:X(t) = [s(t);r(t)],此时可以将输出表示为式(7):

$$Y = W_{\rm out} X \tag{7}$$

其中, Y是一个维度为 $(N - M) \times K$ 的矩阵,这里K是输入数据的时间长度; X 是一个维度为 $(N_{res} + M) \times K$ 的矩阵; 参数 N_{res} 表示 RC 中间层的节点规模, X 的 数学公式见式(8):

é	$I_1(t_1)$	$I_1(t_2)$	•••	$I_1(t_K)$	À	
ê	$I_{2}(t_{1})$	$I_{2}(t_{2})$	•••	$I_2(t_K)$	ú	
e ê	•••	·····.		•••	u ú	
ê	$I_M(t_1)$	$I_M(t_2)$		$I_M(t_K)$	ú ú	(8)
ê	r(1,1)	r(1,2)	•••	r(1,K)	ú	(0)
ê	r(2,1)	<i>r</i> (2,2)	•••	r(2,K)	ú ú	
ê	•••	•••	•••	•••	ú	
	$(N_{\rm res}, 1)$	$r(N_{\rm res},2)$		$r(N_{\rm res},K)$	Ý	

此时,可以利用岭回归方法拟合出预测模型的 输出权重,数学公式具体如下:

$$\boldsymbol{W}_{\text{out}} = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{\lambda}\boldsymbol{I} \right)^{-1}$$
(9)

考虑到输入数据的维度为 *M*,共有 *N* 维数据,所 以在 *t* 时刻目标预测数据的输出向量 *y*(*t*) 的数学表 达式为:

$$\mathbf{y}(t) = \begin{cases} \mathbf{\hat{e}}^{I_{M+1}(t)} \mathbf{\hat{y}} \\ \mathbf{\hat{e}}^{I_{M+2}(t)} \mathbf{\hat{u}} \\ \mathbf{\hat{e}}^{I_{M+2}(t)} \mathbf{\hat{u}} \\ \mathbf{\hat{e}}^{I_{M+2}(t)} \mathbf{\hat{u}} \\ \mathbf{\hat{e}}^{I_{M+1}(t)} \mathbf{\hat{y}} \end{cases}$$
(10)

3 基于标准 SIR 模型数据的结果分析

3.1 数据:经典 SIR 传染病模型

在这部分工作中,选取传染病研究领域的经典 SIR 模型用来生成对模型进行前期测试的数据。该 模型在一独立时间点 t 的数据由三维数组表示: [S(t);I(t);R(t)],其动力学方程数学表达式为:

$$\hat{f} \frac{dS(t)}{dt} = -\beta SI$$

$$\hat{f} \frac{dI(t)}{dt} = \beta SI - \gamma I$$
(11)
$$\hat{f} \frac{dR(t)}{dt} = \gamma I$$

本次拟定恢复速率为 $\gamma_1 = \gamma_2 = \cdots = \gamma_N = 1/14$, 并且从均匀分布的范围(0,0.25)中随机选择传染系 数 β_j 的值,从而生成一组独立随机的合成数据序 列。在初始值方面,初始感染数据 [$I_j(0)$] 取自范 围 (10⁻⁷,10⁻⁴),初始回复人群 [$R_j(0)$] 定为0,从 而有定义 $S_i(0) = 1 - I_i(0) - R_i(0)$ 。

另一方面,传染系数β_j的参数选择是基于现有 数据对 COVID-19 基本复制数的估计,其范围从 (0,3.5)不等。对于 SIR 模型数据的采集,使用四阶 龙格-库塔方法对时间段(0,300)以时间步差 0.1 进 行采集。因此,每个合成的数据集包含 3 000 个数 据点(300 天)。以单层储备池为例,SIR 数据输入 的运算流程如图 4 所示。





Fig. 4 Overall prediction structure of the model (taking a singlelayer reserve pool as an example)

3.2 RC 超参数设定

在输入数据方面,输入维度定为 50,输出维度 定为 1,根据常例对于测试时间本文将储备池泄露 率设定为 0.5,选取一组数据作为目标预测对象,针 对其进行储备池规模参数的选取,此次采用平均绝 对误差(*MAE*)作为评价指标。具体误差分析结果 如图 5、图 6 所示。

本文在后续的实验中利用浅层回声状态网络和 深层回声状态网络相结合的方式对传染病数据进行 预测,根据以上参数设置的误差表现,对回声状态网 络的参数进行设置。其中,RC 的储备池层节点规模 设为 30,泄露率设为 0.5,这些参数将同时应用到浅 层和深层储备池计算网络中。



Fig. 6 The impact of leakage rate on errors

3.3 实验结果与分析

本次实验分别选取 1~3 层储备池结构,训练长 度为 460 个时间单位,测试长度为 30 个时间单位。 由于储备池的内部构建包含一组随机连接的稀疏矩 阵,每次预测输出的拟合结果具有波动性,所以本次 的误差分析取自 10 次模型运行的平均结果,具体预 测结果与误差分析如图 7 和图 8 所示。



图 / 3种体度储备地顶侧趋势图

Fig. 7 Trend chart for predicting of three types of RC

在图 7 结果中,可以看出几种储备池在 SIR 模型数据上拟合效果较好,3 种层次的储备池均可以较为精确地刻画出模型发展的趋势,但是 3 种模型的精确度略有差异。而在图 8 的误差分析中,可以看出在最前期 10 个时间单位的预测中单层储备池

的效果最好,而预测时间在 10~20 个时间单位的时候,3 种层次的储备池预测效果相差不大,在预测发 展到后期的时候,3 层储备池的预测效果最好。由 此可以猜测,针对本文提出的预测思想,当预测时间 变得更长的时候,更深层次的储备池的预测效果可 能会更好。







4 时变权重储备池计算网络

4.1 算法结构概述

通过对第3节的结果分析,本次提出一种时变 权重储备池计算网络,旨在降低单种层次储备池计 算的预测误差,为传染疾病的防控提供更全面准确 的科学依据。

假设不同层级的 RC 模型由 f_1, f_2, \dots, f_n 表示, 那么每个层级分配的权重可以用 w_1, w_2, \dots, w_n 表 示,同时各级权重需满足条件: $\sum_{i=1}^n w_i = 1$,此时时变 权重储备池模型的输出表达式为:

$$f = \sum_{i=1}^{n} w_i f_i \tag{12}$$

本次工作以误差值作为唯一指标,在分配权重 方面,最前期10个时间单位的预测中使用1层储备 池的输出权重;预测时间在11~20个时间单位时, 使用2层储备池的输出权重;预测时间在21~30个 时间单位的时候,使用3层储备池的输出权重。以 单位t作为时间刻度表示,则预测输出f可以表达为 式(13):

$$f = \begin{cases} f_1, & w_1 = 1, w_2 = 0, w_3 = 0; 0 < t \le 10 \\ f_2, & w_1 = 0, w_2 = 1, w_3 = 0; 10 < t \le 20 \quad (13) \\ f_3, & w_1 = 0, w_2 = 0, w_3 = 1; 21 < t \le 30 \\ & \texttt{A} \texttt{K} \texttt{K} \texttt{K} \texttt{M} \texttt{M} \texttt{B} 9 \text{ } \texttt{M} \vec{\pi}_\circ \end{cases}$$

其中,输入数据在1~3 层依次进行处理,针对 之前提出的时变权重思想,对输出权重分时间段进 行选取,不同时间段使用不同层级的输出权重,下面 使用真实数据集进行测试。







4.2 地区真实数据预测结果分析

在这部分工作中,本文选取约翰霍普金斯大学统 计的 COVID-19 数据集,该数据集统计了美国各州以 及全球大部分地区的疫情数据的演变。在该数据集 中,统计了从 2020 年 1 月 23 日、即疫情刚发生的数 据,但是该数据集只统计 2 个感染状态(累计感染人 数,累计死亡人数)的时变数据,研究中初步选取累计 感染人数作为本文的处理数据。本次处理中,选取了 51 个地区的时间序列数据集,因为疫情发展前期 2020 年 5 月 1 日之前的数据存在很大的数据缺失与 误差,所以放弃了前期的一部分数据,范围覆盖 2020 年 5 月 1 日至 2021 年 10 月 14 日、共 490 天的数据 集。另外,为了利于处理数据,使用 Python 附带的 Savgol filter 函数对数据做了平滑滤波处理。

研究过程中,将数据分为2组,随机选取部分地 区数据作为训练数据。剩余的地区数据作为目标预 测数据.在时间维度上,选取前期的2020年5月1 日至2021年9月14日、共460天的数据集作为训 练的输入,剩余的2021年9月14日至2021年10 月14日、共30天数据作为预测输入。按照以上方 法,4种储备池算法的预测趋势如图10所示。同 时,将真实数据与预测数据之间的误差进行 MAE 数 值分析,仿真结果见表1。

表1 不同算法在测试集上的 MAE 误差值

Table 1 MAE error values of different algorithms on the test set

模型(RC)	测试集 MAE
1层	4 278.444 441
2 层	2 648.625 895
3 层	2 960.891 066
时变权重	1 416.834 482





图 10 真实数据与预测数据的误差趋势

Fig. 10 Error trend between real data and predicted data

由表1可以看出,本文提出的时变权重模型在 预测结果的精度中具有一定的优势,可以在一定程 度上降低不同层次储备池在不同时间段的误差,在 测试集上可以将单层相对误差降低45.6%。

5 结束语

本文主要研究了 RC 在传染病发展预测上的效 果,并在新冠感染数据集上进行了性能分析。首先 介绍了 RC 网络的基本结构与各种参数含义,针对 目标处理数据的格式与模型本身的特点提出独特的 同步时间差预测思想。其次,通过标准 SIR 传染病 模型生成的数据来检验预测思想逻辑的正确性,并 对这种方式进行参数数值推演,求得理想的模型参 数,此时 RC 储备池的规模为 30,泄露率为 0.5,预测 效果最为理想。然后,再次使用 SIR 模型生成的规 范数据对不同层次的储备池计算网络进行误差值分 析,得出不同层次储备池在时间线上预测效果不尽 相同的结论。最后,对于时间线上预测值的误差提 出一种时变权重储备池计算网络,同时使用部分真 实地区的疫情发展数据集来验证该方法在预测效果 上的优越性。从最后的实验结果可以发现,时变权 重网络相比较于单种层次的储备池可以有效地提高 预测精度,在测试集上可以将误差降低 45.6%。

仍需指出,本文虽在结果上做得比较详细,但归 根结底还是局限于 RC 这一单网络。目前,许多研 究表明混合网络在特定条件下有着更好的预测效 果,所以储备池网络和其他算法模型的结合还有着 很大的研究空间,后续工作有待进一步开展。

参考文献

- [1] 王灵钰,刘子昂,朱兴杰,等. 基于深度学习的新型冠状病毒肺炎趋势预测研究综述[J]. 电脑知识与技术,2021,17(23):106-109.
- [2] ACOSTA M, GARCIA ZAPIRAIN B. Machine learning algorithms for forecasting COVID 19 confirmed cases in America
 [C]// 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). Louisville, USA: IEEE, 2020:1-6.
- [3] ALDHYANI T, ALRASHEED M, AL-ADAILEH H, et al. Deep learning and holt – trend algorithms for predicting Covid – 19 pandemic[J].Computers, Materials & Continua, 2021(5): 2141– 2160.
- [4] RAHMAN M M , KHATUN F , UZZAMAN A , et al. A comprehensive study of artificial intelligence and machine learning approaches in confronting the coronavirus (COVID-19) pandemic [J]. International Journal of Health Services, 2021, 2021:1-16.
- [5] AKHTAR M, KRAEMER M U G, GARDNER L M. A dynamic neural network model for predicting risk of Zika in real time [J]. BMC Medicine, 2019, 17(1):1-16.
- [6] YAN L, ZHANG H T, GONCALVES J, et al. An interpretable mortality prediction model for COVID – 19 patients [J]. Nature Machine Intelligence, 2020, 2: 283–288.
- [7] FONG S J, LI G, DEY N, et al. Finding an accurate early forecasting model from small dataset: a case of 2019–nCoV novel coronavirus outbreak [J]. International Journal of Interactive Mulitmedia and Aritificial Intelligence, 2020, 6(1):132–140.
- [8] LUKOSEVICIUS M , JAEGER H. Reservoir computing approaches to recurrent neural network training [J]. Computer Science Review, 2009, 3(3):127-149.
- [9] WENG Tongfei, YANG Huijie, GU Changgui, et al. Synchronization of chaotic systems and their machine – learning models[J]. Physical Review A, Atomic, Molecular, and Optical Physics, 2019, 99(4-1):042203.
- [10] SCHRAUWEN B, VERSTAETEN D, CAMPENHOUT J. An overview of reservoir computing: theory, applications and implementations[C]// Proceedings of 15th European Symposium on Artificial Neural Networks(ESANN 2007). Bruges, Belgium: dblp,2007: 471-482.
- [11] JAEGER H. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach [R].Bremen, Germany: National Research Center for Information Technology, 2002.
- [12] 王磊. 回声状态网络优化设计及应用研究[D]. 北京:北京工业 大学,2019.
- [13] 彭宇,王建民,彭喜元. 基于回声状态网络的时间序列预测方法 研究[J]. 电子学报,2010,38(S1):148-154.
- [14] RODAN A, TINO P. Minimum complexity echo state network
 [J]. IEEE Transactions on Neural Networks, 2011, 22(1):131–144.