文章编号:1671-4229(2024)02-0037-11

基于增量学习的冠状病毒人际感染预测研究

杨晓宇[a]，沈　骜[a]，沈嘉豪[a]，廖玉琼[a]，强小利[b]，寇　铮[a*]
(广州大学 a. 计算科技研究院，b. 计算机科学与网络工程学院，广东 广州　510006)

摘　要：2019 年底至今，新型冠状病毒大流行严重影响了公众卫生和社会秩序，而基于机器学习的预测方法可以判别冠状病毒的可感染性表型和大流行风险。目前，已发现 6 类感染人的冠状病毒，病毒基因组序列差异显著，病毒持续遗传变异导致机器学习模型性能下降并引发潜在的学习遗忘现象。文章基于增量学习的模型框架，使用 One-class SVM 算法对冠状病毒新类群进行持续鉴别，并进一步使用参数共享和知识蒸馏的联合策略改造 BP 神经网络，对冠状病毒人际感染表型进行持续学习和预测。结果显示，One-class SVM 对 6 类病毒区分的权衡参数 v 组合在 0.92、0.81、0.24、0.11、0.55、0.20 下达到最优的病毒类群分类效果；当隐藏层节点批次增加为 6 时，预测模型取得最好性能表现，IAC 取得最大值 0.903 5，BT 取得最大值 −0.039 9，有效地抑制了神经网络模型的学习遗忘趋势，模型的预测性能接近联合训练的性能表现(*IAC*：0.923 6)，明显优于未使用知识蒸馏的神经网络(*IAC*：0.776 4)，进一步与其他增量方法比较，优于基于样本的 ESRIL 方法(*IAC*：0.866 2)和基于模型参数的 CCLL 方法(*IAC*：0.885 3)，具有重要的公共卫生应用价值。

关键词：增量学习；冠状病毒；刺突蛋白；人际感染
中图分类号：TP391　　**文献标志码**：A

Using incremental learning to identify human infection of zoonotic coronavirus

YANG Xiao-yu[a]，*SHEN Ao*[a]，*SHEN Jia-hao*[a]，*LIAO Yu-qiong*[a]，*QIANG Xiao-li*[b]，*KOU Zheng*[a*]
(a. Institute of Computing Science and Technology，b. School of Computer Science and Cyber Engineering，Guangzhou University，Guangzhou 510006，China)

Abstract：Since late 2019，the widespread outbreak of the novel coronavirus has had a severe impact on public health and social order. Machine learning-based prediction methods have the capability to determine the infectivity phenotype and pandemic risk of coronaviruses. Presently，six classes of coronaviruses that infect humans have been identified. These viruses exhibit significant differences in their genomic sequences，and the continuous genetic variation in these viruses has resulted in a decline in the performance of machine learning models，potentially causing issues related to learned forgetting. This study，based on an incremental learning model framework，employed a One-class SVM algorithm

收稿日期：2023 − 10 − 29；**修回日期**：2023 − 11 − 05
基金项目：国家自然科学基金资助项目(61972109，62172114)；广东省基础与应用基础研究基金资助项目(2022A1515011468)；广州市科技计划资助项目(202201020237，SL2022A03J01035)
作者简介：杨晓宇(1999—)，男，硕士研究生. E-mail：2112106065@ e. gzhu. edu. cn
* 通信作者. E-mail：kouzheng@ gzhu. edu. cn
引文格式：杨晓宇，沈骜，沈嘉豪，等. 基于增量学习的冠状病毒人际感染预测研究[J]. 广州大学学报(自然科学版)，2024，23(2)：37-47.

for continuous discrimination of novel coronavirus subgroups. Furthermore, a combined strategy of parameter sharing and knowledge distillation to adapt a backpropagation (BP) neural network for continuous learning and prediction of the human-infecting phenotype of coronaviruses was employed. The results indicate that the One-class SVM, with a combination of balancing parameters v at 0.92, 0.81, 0.24, 0.11, 0.55, and 0.2, achieved the optimal classification performance for the six virus classes. It was found that the prediction model achieved the best performance when the number of hidden layer nodes was increased to 6, with a maximum Index of Agreement (*IAC*) value of 0.903 5 and a maximum Bias Total (BT) value of −0.039 9. This effectively suppressed the learning amnesia trend in the network model, with the model's predictive performance being close to that of joint data training (*IAC*: 0.923 6). This performance was significantly better than that of neural networks without knowledge distillation (*IAC*: 0.776 4). Moreover, in comparison to other incremental methods, our approach outperformed sample-based methods such as ESRIL (*IAC*: 0.866 2) and model parameter-based methods like CCLL (*IAC*: 0.885 3). This research holds important implications for public health applications.

Key words: incremental learning; coronavirus; spike protein; human infection

目前已发现6类可以感染人的冠状病毒(HCoV-229E、HCoV-OC43、SARS-CoV、HCoV-NL63、HCoV-HKU1和MERS-CoV)[1]。在新冠疫情之前,全球10%~30%的上呼吸道感染由HCoV-229E、HCoV-OC43、HCoV-NL63和HCoV-HKU1 4类冠状病毒引起[2-3]。2002年11月至2003年7月SARS流行期间,全球共报告临床诊断病例8 096例,死亡774例,病死率9.6%[4]。2012年出现的MERS是一种由MERS-CoV引起的病毒性呼吸道疾病,波及中东、亚洲、欧洲等27个国家和地区,病死率约35%,潜伏期最长为14天,人群普遍易感。自2019年底,新型冠状病毒SARS-CoV-2开始在世界范围内传播,极大地影响了全球公众卫生秩序,人类健康面临极大挑战,截止2021年底已确诊3亿人,造成逾500万死亡病例[5]。目前,已发现的6类冠状病毒基因组序列差异显著,病毒持续遗传变异导致机器学习模型性能下降并引发潜在的学习遗忘现象。

虽然神经网络模型在科学领域得到广泛应用,但神经网络模型依然有很多不足。给定训练数据集,神经网络训练后生成特定预测分类模型,当有新训练集加入训练后,对于旧数据的预测准确率会有所降低。遗忘问题在二十世纪八九十年代被相关学者发现并引发了广泛讨论[6],部分学者采用正交编码来规避存储在神经网络内部表征重叠,也有学者使用双网络的方法来扩充参数数量层次来解决,该问题还有部分学者利用伪数据预训练的方法来减弱学习遗忘的程度[7-8]。近年来,针对神经网络中的遗忘问题,出现了增量学习(Incremental learning)或终身学习(Lifelong learning)研究范式,这类学习范式的网络模型通过不断学习,从过去的任务当中不断地积累旧知识,同时使用这些旧知识帮助新任务的学习,Rebuffi等[9]提出iCaRL方法,该方法选择性地存储之前任务的样本[9],Long等[10]提出使用正交权重修改结合情景模块依赖的方法。

目前已经发现6类可感染人的冠状病毒。在病毒风险预测研究中,遗传差异显著的病毒数据[11]会引发潜在的学习遗忘问题,有必要构建具有增量学习特点的神经网络模型来解决病毒感染预测问题。本文使用One-class SVM进行类群划分,辨别潜在的病毒新类群,用于持续划分预测新任务,同时预测模型能够达到最佳性能,有效地抑制了神经网络模型产生的学习遗忘趋势,进一步与其他增量方法比较,优于基于样本的方法ESRIL和基于模型参数的方法CCLL,本文提出的预测模型具有重要的公共卫生应用价值。

1　冠状病毒类群划分

1.1　刺突蛋白序列特征

本文采用 G-gap[12] 的方法来实现 k-mer 氨基酸序列特征。其公式见式(1):

$$GGAP(g)=(fv_1^g, fv_2^g, \cdots, fv_{400}^g), \tag{1}$$

其中,fv_i^g 是第 i 个 G-gap 二肽出现的频率,其计算公式见式(2):

$$fv_i^g=\frac{O_i^g}{\sum_{i=1}^{400} O_i^g}, \tag{2}$$

其中,O_i^g 代表蛋白序列中第 i 个 G-gap 二肽出现的频率。使用氨基酸特征 G-gap($g=3$)作为氨基酸序列的编码方法处理冠状病毒刺突蛋白数据。

取人源冠状病毒为阳性样本(Positive),动物源冠状病毒为阴性样本(Negative)[13]。利用 K 最近邻算法[14](K-NearestNeighbor,KNN),以 6 类阳性样本为训练数据,计算 6 类阳性样本的空间中心点,进一步在保证数据基本平衡的前提下,划分 6 类阳性样本对应的阴性样本,各自聚合形成 6 个冠状病毒类群。目前,已经发现 6 类可以感染人类的冠状病毒(HCoV-229E、HCoV-OC43、SARS-CoV、HCoV-NL63、HCoV-HKU1、MERS-CoV),使用 KNN 算法保持生物学的遗传差异,考虑了阳性样本与阴性样本之间生物近缘性[15],结果见表 1。

表 1　冠状病毒 6 个类群数据集

Table 1　Six cluster datasets of coronavirus

Virus	HCoV-NL63	HCoV-229E	HCoV-OC43	HCoV-HKU1	SARS-CoV	MERS-CoV
Human	728	215	140	18	342	253
Animal	924	202	201	24	474	313

1.2　One-class SVM 分类模型

由于自然界中冠状病毒具有持续进化的能力,病毒类群判别是一个需要解决的问题,便于增量学习批次任务的构建。基于刺突蛋白序列特征,本文采用 One-class SVM 的方法进行划分冠状病毒类群。如图 1 所示,One-class SVM 是一种基于支持向量机的改进算法,它运用的是一种无监督的训练思想[16],该算法在非线性变换的特征空间中拟合一个紧超球体,以包含大多数基于正例子的目标对象。核心思想就是寻找一个超平面将样本中的正样本圈出来,通过超平面做出单分类决策[17],在圈内的样本被认为是正样本。

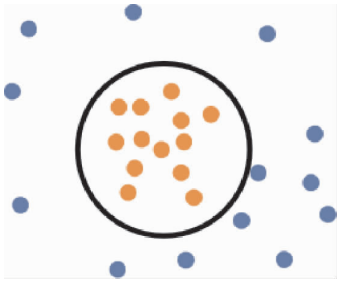

图 1　One-class SVM 示意图

Fig. 1　One class SVM schematic diagram

基于刺突蛋白序列,本文所用到的病毒新类群划分策略是将每一类冠状病毒映射到特征空间,然后尝试找到特定的超球体圈,优化其半径,以尽量多地包含特定类群的冠状病毒样本。超球体的半径与其可容纳的冠状病毒样本数之间的权衡由参数 v 设定。当 v 很小时,可以把更多的冠状病毒类群样本放入超球体内;而当 v 较大时,需把超球面缩小,“挤压”超球面的空间使其更为紧凑,将这个问题转化成了一个拉格朗日乘子优化问题[18]。由上述算法特性可以得知,基于刺突蛋白序列逐次训练特定类群冠状病毒,分别得到各自类群的决策边界,从而区分特定类群冠状病毒和其他冠状病毒类群,将不属于现有决策边界内的病毒类群视为自然界进化出的新型冠状病毒类群,从而形成新批次学习任务。

采用贪婪的优化思想,函数的原始形式如式(3):

$$\min_{R\in\Re,\zeta\in\Re^l,c\in F} R^2+\frac{1}{vl}\sum_i \zeta_i,$$

$$\text{s.t. } \|\Phi(X_i)-c\|^2\leqslant R^2+\zeta_i,\zeta_i\geqslant 0 \text{ for } i\in[l]。 \tag{3}$$

超球体的半径与其可容纳的冠状病毒样本数之间的权衡由参数 $v \in [0,1]$ 设定。当 v 很小时,可以把更多的冠状病毒类群样本放入超球体内;而当 v 较大时,把超球面缩小,"挤压"超球面的空间使其更为紧凑。

2 增量 BP 神经网络

2.1 BP 神经网络模型

BP 神经网络模型基于梯度下降算法构建。每次迭代都在参数空间寻找潜在的权向量,利用每次损失函数向负梯度方向移动的原则,让损失函数达到局部最小值[19],从而满足所需的权向量[20]。BP 神经网络模型简单高效,应用场景广泛。

2.2 增量改造

BP 神经网络的核心在于其反向传播算法[21],本文结合参数共享和知识蒸馏的策略,进行增量改造,以适应病毒批次学习的目标。在对冠状病毒数据进行训练时,随着病毒不同批次类群的批次输入,会因类群改变而产生学习遗忘现象。单纯依靠增加训练次数对原始冠状病毒进行风险感染预测训练,会使其过拟合现象加重,不利于新冠状病毒数据风险感染的训练。为了增加预测模型的鲁棒性,本文批次输入数据训练网络,在记录旧网络结构相关参数的同时,增加知识蒸馏的损失。

知识蒸馏方法[22]实质上是一种模型压缩[23]的思想方式,该方法会用软目标辅助硬目标进行小模型训练,软目标即将样本输入到预训练大模型中得到的输出,硬目标即样本的真实标签。软目标中包含的信息量巨大[24],特别是非正确类别概率的相对大小,而硬目标包含的信息量较低。同时,知识蒸馏方法迁移学习的性质也能解决神经网络模型中出现的学习遗忘问题。传统的 BP 神经网络不具有增量学习的能力。为了增加预测模型的鲁棒性,本文批次输入数据训练网络,在记录旧网络结构相关参数的同时,增加知识蒸馏的损失。为保证对新知识的学习,每次更新网络时增加隐含层节点数目。增量改造的目的是保证在每一次网络更新的时候,都在旧网络基础上进行新一轮的增量学习,详细算法如表2。

表2 增量改造的训练过程

Table 2 The training process of incremental transformation

Algorithm 1: Training process of incremental transformation

θ_0;//对于旧刺突蛋白序列冠状病毒数据的任务特定参数

X_n, Y_n;//对新刺突蛋白序列冠状病毒数据的训练数据和真实标签

$Y_0 \xleftarrow{train} BP(X_1, \theta_0)$;//使用 HCov-N163 来训练初始网络

$\theta_n \leftarrow RandInit(|\theta_n|)$;//随机初始化新参数

for $i \leftarrow 2$ to 6 do

  $Hiden\_node\_num \leftarrow Hiden\_node\_num + 2$;

  while *train* do

    $\hat{Y}_0 \leftarrow BP(X_i, \hat{\theta}_0)$;//旧冠状病毒数据输出

    $\hat{Y}_n \leftarrow BP(X_i, \hat{\theta}_n)$;//新冠状病毒数据输出

    $\theta_s^*, \theta_0^*, \theta_n^* \leftarrow L_{old}(x,y) + L_{new}(Y_n, \hat{Y}_n)$

  end while

end for

2.3 损失函数

对于新类群冠状病毒数据的蒸馏损失函数为式(4):

$$\mathcal{L}_{new}(y_n, \hat{y}_n) = -y_n \cdot \log\hat{y}_n, \tag{4}$$

其中,$\hat{y}_n$ 为网络的软标签输出,y_n 是 one-hot 编码过的正确标签向量。

对于旧冠状病毒数据任务的损失函数为式(5):

$$\mathcal{L}_{old}(y_0, \hat{y}_0) = -H(y_0', \hat{y}_0') = -\sum_{i=1}^{l} {y'}^{(i)}_0 \log\hat{y}_0'^{(i)}, \tag{5}$$

其中,l 是 $y_0'^{(i)}$ 和标签的数量,$y_0'^{(i)}$ 为记录和当前概率的修改版本,$y_0^{(i)}$ 见式(6):

$$y_0'^{(i)} = \frac{(y_0^{(i)})^{1/T}}{\sum_j (y_0^{(j)})^{1/T}}, \quad \hat{y}_0'^{(i)} = \frac{(\hat{y}_0^{(i)})^{1/T}}{\sum_j (\hat{y}_0^{(j)})^{1/T}}, \tag{6}$$

其中,T 为蒸馏损失温度,由上述得到最终损失公式,见式(7):

$$L = \lambda_0 \mathcal{L}_{new}(y_n, \hat{y}_n) + \mathcal{L}_{old}(y_0, \hat{y}_0) + R(\hat{\theta}_s, \hat{\theta}_0, \hat{\theta}_n), \tag{7}$$

其中,λ_0表示蒸馏损失在最终损失公式中占比数。$\hat{\theta}_s$、$\hat{\theta}_0$ 和 $\hat{\theta}_n$ 分别代表共享参数、旧网络参数和加入新数据后的新网络参数。

2.4　模型框架

在对 BP 神经网络模型进行增量改造后,通过定义合适的损失函数,训练得到预测模型,模型框架和训练过程,见图 2。

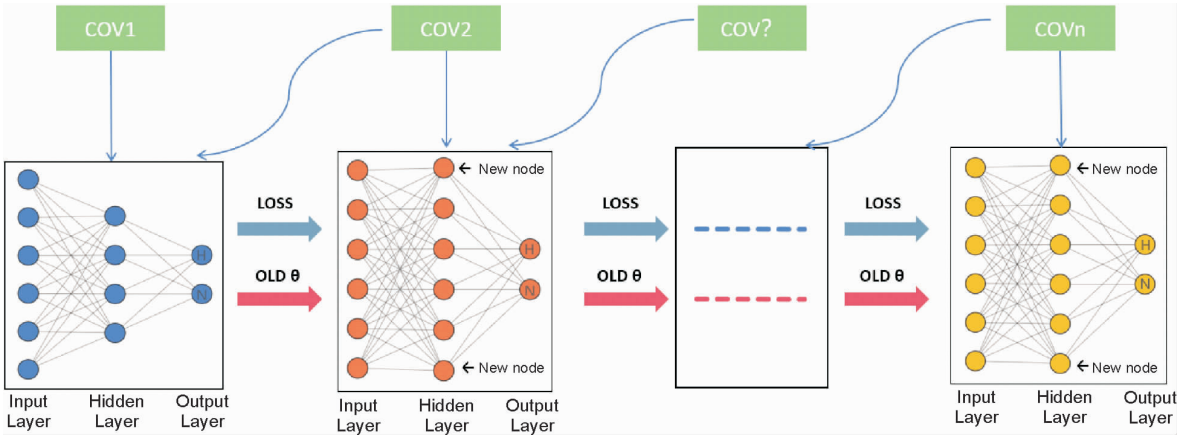

图 2　BP 神经网络增量训练

Fig. 2　Incremental training of BP neural network

3　评价指标

本文从两个角度评价预测结果,一个是网络自身的可靠性,另一个是网络增量效果的可靠性,即抗“学习遗忘”的效果。

对于网络自身可靠性来讲,以 *RAC*(Regular Accuracy)来衡量,即普通的准确率,采取训练集和测试集 4:1 的比例,以混淆矩阵准确率公式来进行验证,即式(8):

$$RAC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

其中,*TP* 为阳性样本预测为阳性样本数,*TN* 为阴性样本为阴性样本数,*FP* 为阴性样本预测为阳性样本数,*FN* 为阳性样本预测为样本数。

对于增量效果的评测相对复杂,本文按照 CoV-NL63、CoV-229E、CoV-OC43、CoV-HKU1、SARS-CoV、MERS-CoV 顺序搭建 6 个批次的学习任务,以时序输入的形式分批训练,核心是新任务增量进入后增强对旧任务的记忆能力,并在最终增量任务完成后预测模型的准确率,相应衡量指标定义为 *IAC*(Increased Accuracy)。另外,在增量的过程中,本文同时考虑任务的逆向迁移能力,相应衡量指标定义为 *BT*(Backward Transfer),对于增量效果的评测详见表 3。

表 3　增量训练评估示意流程

Table 3　Schematic process for incremental training evaluation

任务名称	增量效果的评测			
	任务 1	任务 2	…	任务 T
任务 1	$R_{1,1}$	$R_{1,2}$	…	$R_{1,T}$
任务 2	$R_{2,1}$	$R_{2,2}$	…	$R_{2,T}$
…	…	…	…	…
任务 T-1	$R_{T-1,1}$	$R_{T-1,2}$	…	$R_{T-1,T}$
任务 T	$R_{T,1}$	$R_{T,2}$	…	$R_{T,T}$

表 3 中,$R_{i,j}$代表当训练过任务 i 以后,对任务 j 的 RAC,如果 $i > j$,关注于训练完 i 以后 j 是否被遗忘,若果 $i < j$,则关注于是否很好地把 i 上的任务迁移到 j 上去。由此可得 *IAC* 和 *BT* 的公式见式(9)和式(10):

$$IAC = \frac{1}{T}\sum_{i=1}^{T} R_{T,i}, \tag{9}$$

$$BT = \frac{1}{T-1}\sum_{i=1}^{T-1} R_{T,i} - R_{i,i}, \tag{10}$$

其中,*BT* 为训练完 *T* 任务后对其的逆向判断指标,通常为负数,且越接近 0 说明其抗遗忘能力越好。

4 结果分析

4.1 One-class SVM 分类模型

基于冠状病毒的序列特征,使用无监督的 One-class SVM 方法进行病毒类群判别。影响 One-class 的重要权衡参数为 v ($v \in [0,1]$),v 的大小决定了特定冠状病毒类群与其他类群的区分尺度,本文将从 0 ~ 1 的范围找寻最优数值。依据病毒类群批次输入,搜索病毒类群的分类边界,结果显示 One-class SVM[24] 可以很好地对每种病毒类群进行划分。

基于刺突蛋白[25]序列特征的类群判断结果如图 3 所示,绿色曲线为各自冠状病毒类群的判断准确率,红色曲线为非本类群冠状病毒的判断准确率,其中,图 3(a) ~ 图 3(f)分别为 CoV-NL63、CoV-229E、CoV-OC43、CoV-HKU1、SARS-CoV、MERS-CoV,依据类内外的准确率筛选,便可以得到最佳的权衡参数 v 值组合,具体如表 4 所示:

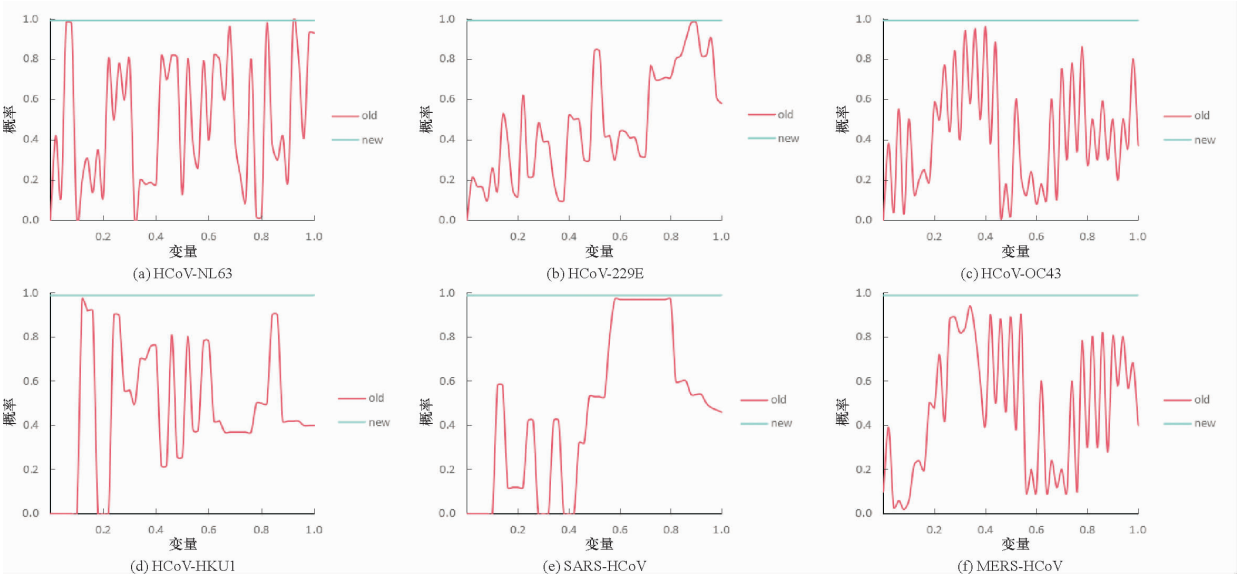

图 3 参数 v 对准确率的影响

Fig. 3 Effect of parameter v on accuracy

表 4 各人源冠状病毒数据的最佳权衡参数值

Table 4 The optimal trade-off parameter values for coronavirus data from different sources

Virus	HCoV-NL63	HCoV-229E	HCoV-OC43	HCoV-HKU1	SARS-CoV	MERS-CoV
Best v	0.92	0.81	0.24	0.11	0.55	0.20

设定最佳 v 值组合后,One-class SVM 可以很好地区分冠状病毒各类群,实验结果可用于增量学习模型的批次输入。在输入冠状病毒数据样本时,将其在 6 个训练好的 One-class SVM 分类模型中进行判断,若均判断为冠状病毒新类群,则需要对神经网络模型进行增量改造。

4.2 学习遗忘现象的对比

在不考虑增量学习的情况下,本文初始 BP 神经网络采用 3 层网络结构,其中,输入层具有 400 个神经元,隐藏层具有 10 个神经元,输出层设定为 2 个神经元,以表示冠状病毒的感染风险类别,整体网络保持全连接结构,训练过程中使用 Adagrad[26] 作为优化函数进行训练。

由表 5 可知,在不考虑增量学习的学习过程中,对当前类别冠状病毒感染风险达到较好的预测效果,即 $R_{i,j}$ 的数值较好(0.845 0 ~ 1.000),但旧病毒类群的预测表现显著变差,出现了“学习遗忘”的现象,其中,CoV-NL63 的风险预测效果从

1.000 0 下降到了 0.712 8,CoV-229E 的风险预测效果从 0.972 3 下降到 0.725 1,CoV-OC43 的风险预测效果从 0.942 3 下降到 0.718 3,CoV-HKU1 的风险预测效果从 0.910 9 下降到 0.739 8,SARS-CoV 的风险预测效果从 0.845 1 轻微下降到 0.823 7,均未表现出学习遗忘趋势,这可能与其内在的分子机制有关,学习表现与作为第一批次的 CoV-NL63 具有相似的趋势,可供病毒学进一步进行生物验证。MERS-CoV 作为训练过程中最后批次输入的数据,不评测其增量学习效果。整体而言,这种学习遗忘趋势可以从 *IAC* 值(0.776 4)和 *BT* 值(−0.190 2)很直观地看出来。

表 5　未进行增量改造的冠状病毒预测结果

Table 5　Prediction results of coronavirus without incremental modification

病毒类型	准确率					
	CoV-NL63	CoV-229E	CoV-OC43	CoV-HKU1	SARS-CoV	MERS-CoV
CoV-NL63	1.000 0	0.798 4	0.698 2	0.522 1	0.889 2	0.545 3
CoV-229E	0.827 1	0.972 3	0.523 5	0.546 7	0.871 2	0.558 9
CoV-OC43	0.824 1	0.782 4	0.942 3	0.508 7	0.942 1	0.602 1
CoV-HKU1	0.782 3	0.758 2	0.821 1	0.910 9	0.906 9	0.509 3
SARS-CoV	0.741 2	0.803 4	0.789 2	0.742 5	0.845 1	0.503 9
MERS-CoV	0.712 8	0.725 1	0.718 3	0.739 8	0.823 7	0.939 1
	IAC = 0.776 4			*BT* = −0.190 2		

基于增量 BP 神经网络的模型结构,本文完成了 6 个批次冠状病毒类群的训练和统计。为了寻找到合适的增量结构,本文对隐藏层批次分别增加 2、4 和 6 个神经元等 3 种方式进行了比较,统计增量 BP 网络性能相关的 IAC 和 BT 数值。每组数据训练 5 次,取 5 次的平均值得到结果如表 6 所示。

表 6　增加隐藏节点为 2 的冠状病毒预测结果

Table 6　Coronavirus prediction results with 2 hidden nodes added

病毒类型	准确率					
	CoV-NL63	CoV-229E	CoV-OC43	CoV-HKU1	SARS-CoV	MERS-CoV
CoV-NL63	1.000 0	0.809 0	0.637 6	0.555 6	0.970 0	0.565 2
CoV-229E	1.000 0	0.988 0	0.623 1	0.548 2	0.930 0	0.556 5
CoV-OC43	0.969 7	0.944 6	0.956 5	0.444 4	0.900 5	0.582 6
CoV-HKU1	0.959 7	0.943 6	0.913 0	0.888 9	0.880 0	0.591 3
SARS-CoV	0.929 3	0.863 6	0.874 1	0.778 9	0.835 3	0.573 9
MERS-CoV	0.954 6	0.830 9	0.842 4	0.768 9	0.842 9	0.850 8
	IAC = 0.848 4			*BT* = −0.085 8		

由表6 可知,在增加隐藏节点数量为 2 的情况下,对旧病毒类群的预测表现显著变好,减弱了“学习遗忘”的现象,其中,CoV-NL63 的风险预测效果从 1.000 0 轻微下降到了 0.954 6,与没有增量学习的情况相比较提升了 0.241 8;CoV-229E 的风险预测效果从 0.988 0 轻微下降到了 0.830 9,与没有增量学习的情况相比较提升了 0.105 8;CoV-OC43 的风险预测效果从 0.956 5 下降到 0.842 4,与没有增量学习的情况相比较提升了 0.124 1;CoV-HKU1 的风险预测效果从 0.888 9 下降到 0.768 9,与没有增量学习的情况相比较提升了 0.029 1。整体来说,增加隐藏节点为 2 的情况下,学习遗忘现象得到了纠正,IAC 值从 0.776 4 提升到 0.848 4 和 BT 值从 −0.190 2 提升到了 −0.085 8。

由表7可知,在增加隐藏节点数量为4的情况下,对旧病毒类群的预测表现也显著变好,其中,CoV-NL63的风险预测效果从1.000 0轻微下降到了0.989 8,与没有增量学习的情况相比较提升了0.277 0;CoV-229E的风险预测效果从0.988 0轻微下降到了0.840 9,与没有增量学习的情况相比较提升了0.115 8;CoV-OC43的风险预测效果从0.943 5下降到0.852 3,与没有增量学习的情况相比较提升了0.134 0;CoV-HKU1的风险预测效果从0.887 2略微提升到0.888 9,与没有增量学习的情况相比较提升了0.149 1。整体来说,增加隐藏节点为4的情况下,学习遗忘现象得到了纠正,*IAC*值从0.776 4提升到0.880 7,*BT*值从-0.190 2提升到了-0.047 0。

表7 增加隐藏节点为4的冠状病毒预测结果

Table 7 Coronavirus prediction results with 4 hidden nodes added

病毒类型	准确率					
	CoV-NL63	CoV-229E	CoV-OC43	CoV-HKU1	SARS-CoV	MERS-CoV
CoV-NL63	1.000 0	0.842 0	0.632 3	0.533 6	0.942 9	0.545 2
CoV-229E	1.000 0	0.988 0	0.642 1	0.565 2	0.923 6	0.553 2
CoV-OC43	0.979 8	0.977 2	0.943 5	0.478 2	0.912 6	0.542 5
CoV-HKU1	0.959 5	0.954 7	0.919 5	0.887 2	0.874 2	0.584 2
SARS-CoV	0.939 3	0.863 6	0.884 3	0.821 2	0.842 1	0.575 1
MERS-CoV	0.989 8	0.840 9	0.852 3	0.888 9	0.853 9	0.858 2
	IAC = 0.880 7			*BT* = -0.047 0		

由表8可知,在增加隐藏节点为6的情况下,对旧病毒类群的预测表现显著变好,其中,CoV-NL63的风险预测效果从1.000 0轻微下降到了0.932 2,与没有增量学习的情况相比较提升了0.219 4;CoV-229E的风险预测效果从0.988 0轻微下降到了0.853 3,与没有增量学习的情况相比较提升了0.128 2;CoV-OC43的风险预测效果从0.956 5下降到0.862 4,与没有增量学习的情况相比较提升了0.144 1;CoV-HKU1的风险预测效果从0.921 1略微下降到0.888 0,与没有增量学习的情况相比较提升了0.141 8。整体来说,增加隐藏节点为6的情况下,学习遗忘现象进一步得到了纠正,IAC值从0.776 4提升到0.903 5,BT值从-0.190 2提升到了-0.039 9。

综合上述结果,随着隐含层节点数量的增加,神经网络承载信息能力相应提升,同时加入的蒸馏损失函数能记住冠状病毒旧类群的知识信息,从而克服冠状病毒风险预测中出现的学习遗忘现象,增强了预测结果的鲁棒性。

表8 增加隐藏节点为6的冠状病毒预测结果

Table 8 Coronavirus prediction results with 6 hidden nodes added

病毒类型	准确率					
	CoV-NL63	CoV-229E	CoV-OC43	CoV-HKU1	SARS-CoV	MERS-CoV
CoV-NL63	1.000 0	0.879 9	0.628 1	0.544 4	0.979 9	0.561 1
CoV-229E	1.000 0	0.988 0	0.623 3	0.555 6	0.931 1	0.556 4
CoV-OC43	0.980 0	0.977 2	0.956 5	0.445 5	0.922 2	0.582 6
CoV-HKU1	0.969 9	0.966 7	0.920 0	0.921 1	0.881 2	0.491 3
SARS-CoV	0.942 2	0.863 3	0.888 9	0.822 2	0.835 3	0.573 9
MERS-CoV	0.932 2	0.853 3	0.862 4	0.888 0	0.965 5	0.920 0
	IAC = 0.903 5			*BT* = -0.039 9		

4.3　不同蒸馏温度的比较

本文对 3 种不同的优化函数在不同蒸馏温度下做交叉验证,*IAC* 来作为衡量数值。考虑网络收敛时本身的随机误差,每组实验准确率通过 10 次独立训练后平均得到,固定初始学习率 0.001 5。

参数优化结果如图 4 所示。

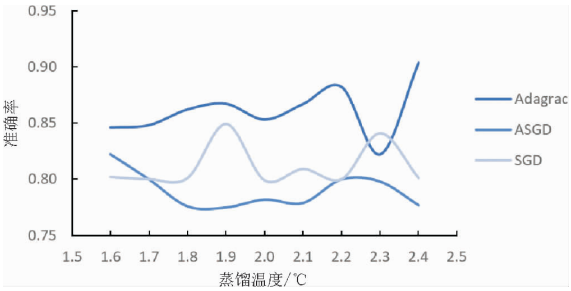

图 4　优化算法在不同蒸馏温度下准确率

Fig. 4　Optimization algorithm accuracy at different distillation temperatures

图 4 中,纵轴为增量训练的准确率,横轴为蒸馏温度 T,不同的优化器在不同的蒸馏温度下训练结果有所不同。ASGD 在温度 T = 1.6 ℃时取得最大值 0.822 2,SGD 在 T = 1.9 ℃时取得最大值 0.849 3,Adagrad 在 T = 2.4 ℃时取得最大值 0.903 5。从实验结果来看,Adagrad 相较于其他两种优化算法具有更好的整体表现。

4.4　知识蒸馏预测能力的验证

在相同的实验环境下,采用没有知识蒸馏的训练方式来验证蒸馏损失的优点,同时,与冠状病毒 6 个类群数据联合训练的结果进行比较,来验证增量 BP 神经网络的学习能力。结果如图 5 所示。

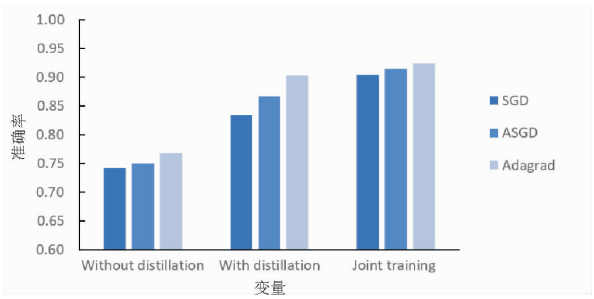

图 5　知识蒸馏预测能力的验证

Fig. 5　Verification of knowledge distillation prediction ability

图 5 中,联合训练数据集包含冠状病毒 6 个类群的所有数据,其对新冠病毒的可感染性预测是最好的,可作为增量 BP 神经网络的性能参照。当选取 Adagrad 作为优化算法时,联合训练的 IAC 达到了 0.923 6,带蒸馏损失的增量 BP 神经网络准确率达到了 0.903 5,而不使用蒸馏 BP 网络准确率仅为 0.768 1。

4.5　不同增量学习方法的比较

将本文构建的增量 BP 神经网络与两种增量学习方法进行了比较。基于样本的方法采取 ESRIL 模型结构[27],使用默认参数进行训练。基于模型参数的方法采取 CCLL 模型结构[28],使用默认参数进行训练,不同增量学习方法的比较,见图 6。

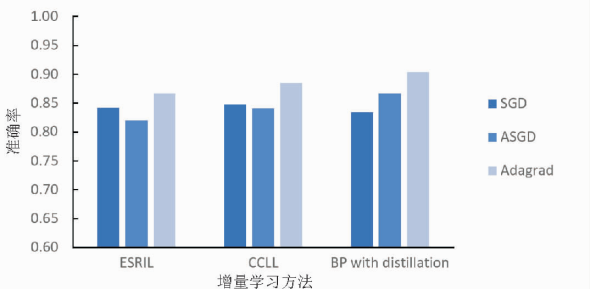

图 6　不同增量学习方法的比较

Fig. 6　Comparison of different incremental learning methods

由图 6 可知,3 种增量方法在选取 Adagrad 作为优化算法时取得最优表现,增量 BP 神经网络准确率达到了 0.903 5,优于 CCLL(*IAC*: 0.885 3)和 ESRIL(*IAC*: 0.866 2),对冠状病毒的可感染性预测工作表现更好。

5　结　论

病毒基因组序列会随时间发生显著变异,导致已有机器学习性能下降并引发学习遗忘现象,因此,有必要建立基于增量学习的冠状病毒人际感染预测模型,实现对病毒变种感染风险的持续监控,具有重要的公共卫生应用价值。本文主要工作包括:①基于刺突蛋白序列数据,构建冠状病毒可感染性预测模型,在传统 BP 神经网络模型基础之上,随批次数据增加隐藏层节点数目,同时使用知识蒸馏策略进行增量改造。在批次增加隐含层节点为 6 时,预测模型 *IAC* 取得最大值 0.903 5,

BT 取得最大值 -0.039 9,有效抑制了网络模型的学习遗忘趋势。Adagrad 梯度优化算法在冠状病毒可感染性风险预测中表现最好,在蒸馏温度为 2.4 ℃时准确率取得最大值 0.903 5。②面对冠状病毒的流行特点,本文基于病毒蛋白质构建增量学习预测模型。通过实验比较可知,模型的预测性能(*IAC*: 0.903 5)都接近数据联合训练的表现(*IAC*: 0.923 6),明显优于未使用知识蒸馏的神经网络性能(*IAC*: 0.776 4),进一步与其它增量方法比较,亦优于基于样本的 ESRIL 方法(*IAC*: 0.866 2)和基于模型参数的 CCLL 方法(*IAC*: 0.885 3)。

参考文献:

[1] Weiss S R, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus[J]. Microbiology and Molecular Biology Reviews, 2005, 69(4): 635-664.

[2] Cui J, Li F, Shi Z. Origin and evolution of pathogenic coronaviruses[J]. Nature Reviews Microbiology, 2019, 17(4): 181-192.

[3] 杨扬, 谭文杰. 冠状病毒载体研究进展[J]. 病毒学报, 2012, 28(3):65-70.

[4] Yin Y, Wunderin R. MERS, SARS and other coronaviruses as causes of pneumonia[J]. Respirology, 2018, 23(2): 130-137.

[5] Su S, Wong G, Shi W, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses[J]. Trends in Microbiology, 2016, 24(6): 490-502.

[6] Nembhard D A, Uzumeri M V. Experiential learning and forgetting for manual and cognitive tasks[J]. International Journal of Industrial Ergonomics, 2000, 25(4):315-326.

[7] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[8] He K M, Zhang X Y, Ren S Q, et al. IEEE Conferenceon Computer Vision Pattern Recognition (CVPR), June 27-30, 2016[C]. Piscatway: IEEE, 2016.

[9] Rebuffi S A, Kolesnikov A, Sperl G, et al. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 21-26, 2017[C]. Piscatway: IEEE, 2017.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39: 640-651.

[11] 吴琦琨,赖浪文,徐怀胜,等. 新一代数据存储介质—DNA[J]. 广州大学学报(自然科学版), 2020,19(6):35-40.

[12] Qiang X, Xu P, Fang G, et al. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus[J]. Infectious Diseases of Poverty, 2020, 9(1): 54-71.

[13] Nilashi M, Ibrahim O, Ahmadi H, et al. A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques[J]. Biocybernetics and Biomedical Engineering, 2018, 38(1): 1-15.

[14] Zhang S, Li X, Zong M,et al. Learning k for KNN classification[J]. ACM Transactions on Intelligent Systems and Technology, 2017(3):8.

[15] 钱冰,马宝山,刘玉. 蛋白质与蛋白质相互作用预测模型综述[J]. 广州大学学报(自然科学版), 2023, 22(1):25-32.

[16] Guo L, Xie G, Xu X Y, et al. Exemplar-supported representation for effective class-incremental learning[J]. IEEE Access, 2020, 8: 51276-51284.

[17] You C, Li C, Robinson D P, et al. 15th Eurepean Conference on Computer Vision (ECCV), September 08-14, 2018[C]. Bertin: Springer, 2018.

[18] Belouadah E, Popescu A. 2019 IEEE/CVF International Conference on Computer Vision, October 27-November 02, 2019 [C]. Piscatway: IEEE, 2019.

[19] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.

[20] Atkinson C, Mccane B, Szymanski L, et al. Internation Conference on Image and Vision Computing New Zealand (IVC-

NZ), December 04-06, 2017[C]. Piscatway: IEEE, 2017.

[21] Hayes T L, Kafle K, Shrestha R, et al. European Conference on Computer Vision, November 07, 2020[C]. Berlin: Springer, 2020.

[22] Li Z, Hoiem D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2935-2947.

[23] Zeng G, Chen Y, Cui B, et al. Continual learning of context-dependent processing in neural networks[J]. Nature Machine Intelligence, 2019, 1(8): 364-372.

[24] Chanmasemani F F, Singh Y P. International Conference on Bio-inspired Computing: Theories & Applications, September 27, 2011.[C]. Piscatway: IEEE,2011.

[25] Guruprasad L. Human SARS CoV-2 spike protein mutations[J]. Proteins,2020,89(5):569-576.

[26] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[27] Kong T, Sun F C, Liu H P, et al. Foveabox: Beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[28] Roy D, Panda P, Roy K. Tree-CNN: A hierarchical deep convolutional neural network for incremental learning[J]. Neural Networks, 2020, 121: 148-160.

【责任编辑: 陈　钢】