文章编号： 1673-3193（2024）04-0464-09

基于优化LSTM网络的多区域协同流感预测方法

张玲玲[1,2,3]，杨晓文[1,2,3]，薛红新[1,2,3]，孟罗春子[1,2,3]，韩慧妍[1,2,3]

（1. 中北大学 计算机科学与技术学院， 山西 太原 030051；
2. 机器视觉与虚拟现实山西省重点实验室， 山西 太原 030051；
3. 山西省视觉信息处理及智能机器人工程研究中心， 山西 太原 030051）

摘　要： 流感通常表现出季节性、急性起病和传播速度快的特点，因此对流感的准确预测至关重要。针对流感预测精度不佳和长短期记忆网络参数寻优困难导致耗时耗力的问题，提出了一种基于皮尔逊相关系数和采用蜣螂优化算法（DBO）优化长短期记忆网络（LSTM）的多区域协同流感预测方法（MRC-DBO-LSTM）。该模型不仅学习本地的历史数据，还学习与其强相关的区域的历史数据。首先，通过皮尔逊相关系数筛选与预测地强相关的区域，以得到更高维度的输入特征；其次，通过LSTM的门机制衡量这些区域数据的权重来进行特征融合；最后，引入蜣螂优化算法对长短期记忆网络的超参数（如隐藏层数、隐藏层节点数和迭代次数等）寻优，进而生成预测结果。对山西省流感发病率预测的实验结果表明，学习多区域历史数据的DBO-LSTM模型的均方误差（MSE）仅为0.003 8，与差分整合移动平均自回归（ARIMA）模型相比，MSE降低了99.6%；与季节性差分自回归滑动平均（SARIMA）模型相比，MSE降低了98.7%；与LSTM模型相比，MSE降低了71.0%，与仅使用本地历史数据的DBO-LSTM模型相比，MSE降低了48.6%。结果证明所提模型能够有效提高流感的预测精度。

关键词： 流感预测；蜣螂优化算法；长短期记忆网络；深度学习；时间序列
中图分类号： TP183　　**文献标识码：** A　　**doi:** 10.3969/j.issn.1673-3193.2024.04.007
引用格式： 张玲玲，杨晓文，薛红新，等. 基于优化LSTM网络的多区域协同流感预测方法[J]. 中北大学学报（自然科学版），2024，45(4)：464-472.
ZHANG Lingling，YANG Xiaowen，XUE Hongxin，et al. Multi-regional collaborative influenza prediction method based on optimized LSTM network[J]. Journal of North University of China(Natural Science Edition)，2024，45(4)：464-472.

Multi-Regional Collaborative Influenza Prediction Method Based on Optimized LSTM Network

ZHANG Lingling[1,2,3]，YANG Xiaowen[1,2,3]，XUE Hongxin[1,2,3]，MENG-LUO Chuenzi[1,2,3]，HAN Huiyan[1,2,3]

（1. School of Computer Science and Technology，North University of China，Taiyuan 030051，China；
2. Shanxi Key Laboratory of Machine Vision and Virtual Reality，Taiyuan 030051，China；
3. Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center，Taiyuan 030051，China）

Abstract： Influenza usually shows the characteristics of seasonal，acute onset and rapid transmission，so

收稿日期： 2023-07-27
基金项目： 国家自然科学基金资助项目（62106238）；山西省高等学校科技创新项目（2020L0283）；山西省自然科学基金资助项目（202203021212138）
作者简介： 张玲玲（1996—），女，硕士生，主要从事人工智能与计算机视觉方面的研究。
通信作者： 杨晓文（1980—），女，副教授，博士，主要从事计算机视觉与虚拟仿真与可视化方面的研究。E-mail：wenyang1314@nuc.edu.cn。

the accurate prediction of influenza is very important. Aiming at the problems of poor accuracy of influenza prediction and the difficulty of optimizing parameters of long short-term memory（LSTM）, a multi-region collaborative influenza prediction method （MRC-DBO-LSTM） based on Pearson correlation coefficient and dung beetle optimization algorithm（DBO） was proposed. The model learns not only the historical data of the local area, but also the historical data of the region with which it is strongly related. Firstly, Pearson correlation coefficient was used to select the regions strongly correlated with the prediction place, so as to obtain the input features of higher dimensions. Secondly, the LSTM gate mechanism was used to measure the weight of these regional data for feature fusion. Finally, dung beetle optimization algorithm was introduced to optimize the super parameters（such as the number of hidden layers, the number of hidden layer nodes and the number of iterations, etc.） of the LSTM, so as to generate prediction results. The experimental results of predicting influenza incidence in Shanxi Province show that the R-Squared of the MRC-DBO-LSTM model based on multi-regional historical data is 0.988, and the mean square error（MSE） is only 0.003 8. Compared with the differential integrated moving average autoregression（ARIMA） model, MSE is decreased by 99.6%, MSE is decreased by 98.7% compared to the seasonal differential autoregressive moving average（SARIMA） model, MSE is decreased by 71.0% compared to the LSTM model, and MSE is decreased by 48.6% compared to the DBO-LSTM model using only local historical data. It is proved that the proposed model can effectively improve the prediction accuracy of influenza.

Key words: influenza prediction; dung beetle optimization algorithm; long short-term memory network; deep learning; time series

0 引 言

流感是一种通过核糖核酸病毒感染呼吸道而传播的传染性疾病[1]。季节性流感仍持续且严重地威胁着全球人类的生命健康，同时也给世界各地带来了巨大的经济负担[2]。因此，如果能对流感进行有效预测，使人们能够提前采取预防措施，为相关机构提供更多的应对时间，将会对流感大流行的预防和控制产生积极的影响[3]。

数学和统计建模一直是了解和预测流感的重要工具[4]。Thomas等[5]开发了一种通过对极值进行统计来估计超标率的条件概率的方法，并成功对法国的流感进行了实时预测。实验数据表明，他们所构建的GP预测方法相较于Logistic预测方法具有更高的准确性。郑月彬等[6]针对国家流感中心发布的2012年第1周至2018年第48周的每周流感发病病例数据，采用自回归滑动平均模型（Autoregressive Integrated Moving Average, ARIMA）和Holt-Winters指数平滑模型对数据建模，并进行了对比。结果显示ARIMA模型的平均相对误差为7.06%，Holt-Winters指数平滑模型的平均相对误差为10.73%，证明了ARIMA模型能够有效地预测流感发病人数。

随着信息技术的发展，机器学习、深度学习和人工智能等技术在数据分析和预测建模等任务中得到了广泛应用[7]。在这些技术中，长短期记忆网络（Long Short-Term Memory, LSTM）模型以其能够有效解决长期依赖问题的特点，而被广泛应用于时间序列预测[8]。Wang等[9]的实验结果表明，实际新增确诊病例数与LSTM预测曲线显著吻合，预测数据与官方数据总体上吻合良好。Tsan等[10]实验对比证明了LSTM模型相较于ARIMA模型具有更大的优势。然而，该实验仅使用了本地的历史数据对未来进行预测。

尽管LSTM在处理非线性、多变量和多步骤预测问题时具有鲁棒性[11]，但与其他神经网络一样，LSTM模型的超参数会对结果产生重大影响，不同的参数设置会导致预测性能的显著差异[12]，而超参数的优化通常需要大量资源和时间[13]。为了获得更为准确的流感预测结果，本文提出了MRC-DBO-LSTM模型，即采用蜣螂优化算法对LSTM优化，以实现小区域的流感预测。本文的主要贡献如下：

1）引入新的数据集，对更小一级的地理单位（如省级）进行流感预测，可以方便地方相关机构采取更加有针对性的预防措施；

2)通过进行皮尔逊相关性分析,选择与预测地更相关的区域的历史数据作为共同的预测依据进行特征融合,而非仅将本地的历史数据作为特征进行预测;

3)依靠蜣螂优化算法全局优化的特点,对MRC-DBO-LSTM模型的超参数进行优化,如隐藏层数、隐藏层节点数、迭代次数和周期数,以提高模型的性能。

1 MRC-DBO-LSTM流感预测方法

1.1 长短期记忆网络

1997年,Hochreiter等[14]提出了长短期记忆网络。LSTM因其可以解决梯度消失问题的优越性,已被广泛用于生物医学[15]、语音识别[16]、金融[17]和图像分类[18]等领域。

LSTM是一种由LSTM区块组成的神经网络,每一个LSTM区块由3个逻辑门和1个单元组成,即输入门、遗忘门、输出门和记忆元,3个门控制进出记忆元的完整信息流。多个LSTM区块堆叠在一起形成一个完整的LSTM,如图1所示。

对于t时刻的LSTM区块,其输入$\boldsymbol{X}_t$与前一个时刻$t-1$的隐状态$\boldsymbol{Hid}_{t-1}$作为数据送入LSTM区块中,由3个具有sigmoid激活函数的全连接层分别处理得到当前时刻的输入门、遗忘门和输出门的值,由具有tanh激活函数的全连接层处理得到当前时刻的候选记忆元,数学表达式如下

$$\boldsymbol{In}_t=sigmoid(\boldsymbol{X}_t\boldsymbol{W}_{xi}+\boldsymbol{Hid}_{t-1}\boldsymbol{W}_{hi}+\boldsymbol{b}_i),\quad(1)$$

$$\boldsymbol{Fg}_t=sigmoid(\boldsymbol{X}_t\boldsymbol{W}_{xf}+\boldsymbol{Hid}_{t-1}\boldsymbol{W}_{hf}+\boldsymbol{b}_f),\quad(2)$$

$$\boldsymbol{Out}_t=sigmoid(\boldsymbol{X}_t\boldsymbol{W}_{xo}+\boldsymbol{Hid}_{t-1}\boldsymbol{W}_{ho}+\boldsymbol{b}_o),\quad(3)$$

$$\tilde{\boldsymbol{Cm}}_t=\tanh(\boldsymbol{X}_t\boldsymbol{W}_{xc}+\boldsymbol{Hid}_{t-1}\boldsymbol{W}_{hc}+\boldsymbol{b}_c),\quad(4)$$

式中:$\boldsymbol{W}_{xi}$,$\boldsymbol{W}_{xf}$,$\boldsymbol{W}_{xo}$,$\boldsymbol{W}_{xc}$分别是当前时刻输入门、遗忘门、输出门、记忆元与输入相关的权重矩阵;$\boldsymbol{W}_{hi}$,$\boldsymbol{W}_{hf}$,$\boldsymbol{W}_{ho}$,$\boldsymbol{W}_{hc}$分别为当前时刻输入门、遗忘门、输出门、记忆元与前一个时间步的隐状态相关的权重矩阵;$\boldsymbol{b}_i$,$\boldsymbol{b}_f$,$\boldsymbol{b}_o$,$\boldsymbol{b}_c$分别为当前时刻输入门、遗忘门、输出门、记忆元的偏置。

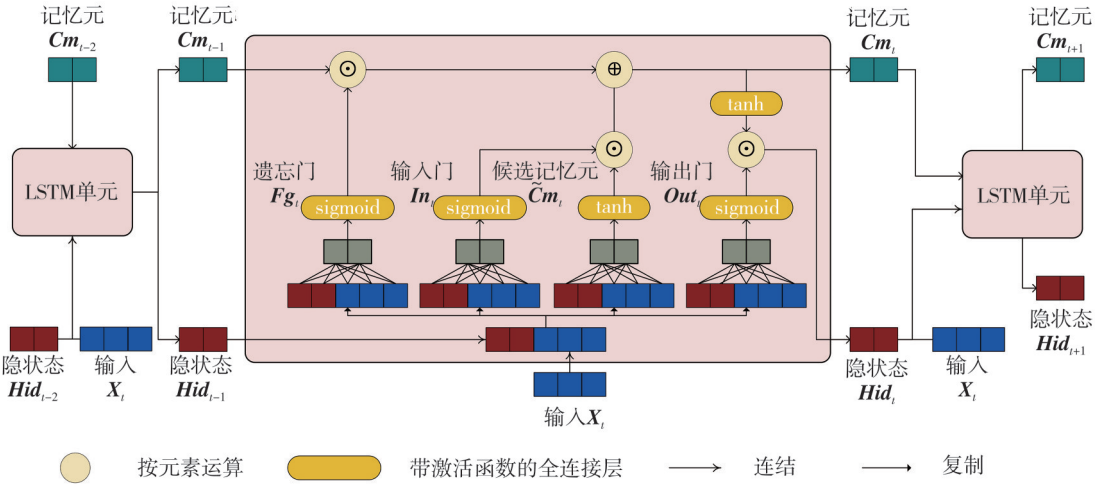

图1 长短期记忆网络架构

Fig. 1 Architecture of Long Short-Term Memory Network

当前时刻的记忆元的值C_t由输入门和遗忘门共同决定。输入门决定采用多少来自候选记忆元的数据,遗忘门决定保留多少过去的记忆元,数学表达式为

$$\boldsymbol{Cm}_t=\boldsymbol{Fg}_t\odot\boldsymbol{Cm}_{t-1}+\boldsymbol{In}_t\odot\tilde{\boldsymbol{Cm}}_t。\quad(5)$$

当前时刻的隐状态Hid_t,即当前时刻的输出门与记忆元控制得到的LSTM单元的输出。

$$\boldsymbol{Hid}_t=\boldsymbol{Out}_t\odot\tanh(\boldsymbol{C}_t)。\quad(6)$$

1.2 蜣螂优化算法

蜣螂优化算法是Xue等[19]提出的最新的元启发式算法之一,在全局搜索和局部开发中表现良好。该算法的灵感来自于大自然——通过观察蜣螂种群的行为,选取其滚球、跳舞、繁殖、觅食和偷窃行为用于模拟。

设定种群内有n只蜣螂,其中,k_1只发生滚球行为,k_2只发生觅食行为,k_3只发生繁殖行为,k_4只发生偷窃行为,对所有蜣螂进行初始化。

发生滚球行为的蜣螂,以天体作为导航,使其可以保持直线滚动。假定太阳光强度可以对其路线产生影响,在当前迭代轮次m,发生滚球行为的蜣螂z_i的位置更新方式为

$$z_i(m+1)=z_i(m)+\alpha\times k\times z_i(m-1)+b\times\Delta z,$$
$$\Delta z=\left|z_i(t)-Z^w\right|, \tag{7}$$

式中：α用于模拟使蜣螂偏离原来方向的自然因素，当$\alpha=1$时，表示不发生偏离，当$\alpha=-1$时，表示偏离原来的方向；$k\in(0,0.2]$为偏转系数常量；$b\in(0,1)$为一个自然系数；Z^w表示全局最差位置；Δz模拟光源强度的变化，其越大表示光源越弱。

当蜣螂因遇到障碍物而无法前进时，其会发生跳舞行为以完成重新定向，从而获得新的路线，此时的位置更新为

$$z_i(m+1)=z_i(m)+\tan\theta\left|z_i(m)-z_i(m-1)\right|, \tag{8}$$

式中：$\theta\in[0,\pi]$为偏转角度，由切线函数可知，当$\theta=0,\frac{\pi}{2},\pi$时，该蜣螂$z_i$位置不会改变。

当蜣螂将粪球滚动至安全的地方后，会将粪球隐藏用于产卵。作者提出了一种边界选择策略来选择产卵的区域，定义为

$$Low^*=\max\left(Z^*\times(1-R),Low\right),$$
$$Up^*=\min\left(Z^*\times(1+R),Up\right), \tag{9}$$

式中：Low^*，Up^*，Low和Up分别表示产卵区域的下、上边界和搜索空间的下、上边界；Z^*为当前种群内的最优位置；$R=1-\frac{m}{M}$为惯性权值；M为蜣螂优化算法的最大迭代次数。由定义可知，产卵区域是动态变化的且取决于R，示意图如图 2 所示。

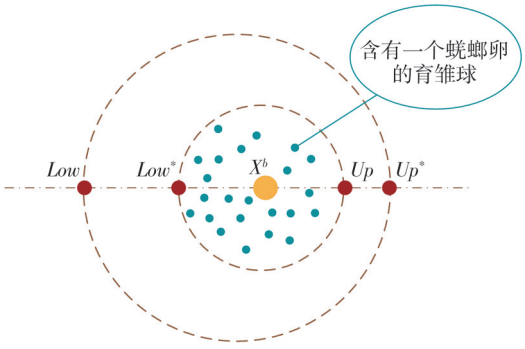

图 2　最佳边界搜索

Fig. 2　Optimal boundary search

产卵区确定后，蜣螂会发生繁殖行为。算法规定每只蜣螂每次只产一个卵，由于产卵区域是动态变化的，因此，在第m次迭代时，第i只发生繁殖行为的蜣螂的位置$Z_i(m)$更新定义为

$$Z_i(m+1)=Z^*+b_1\times\left(Z_i(m)-Low^*\right)+$$
$$b_2\times\left(Z_i(m)-Up^*\right), \tag{10}$$

式中：b_1和b_2是大小为$1\times Dim$的两个独立随机变量，Dim为优化问题的维数。

经过孵化，育雏球里的小蜣螂会离开育雏球发生觅食行为，因此需要建立最佳觅食区，定义为

$$Low^b=\max\left(Z^b\times(1-R),Low\right),$$
$$Up^b=\min\left(Z^b\times(1+R),Up\right), \tag{11}$$

式中：Low^b和Up^b分别为最佳觅食区的下、上边界；Z^b为全局最优位置。因此，在第i次迭代发生觅食行为的蜣螂的位置$z_i(m)$更新方式为

$$z_i(m+1)=z_i(m)+C_1\times\left(z_i(m)-Low^b\right)+$$
$$C_2\times\left(z_i(m)-Up^b\right), \tag{12}$$

式中：C_1为服从正态分布的随机数；$C_2\in(0,1)$为随机数。

在蜣螂种群内部，偷窃食物是常见的行为。文章假设Z^b是争夺食物的最佳位置，在第m次迭代时，发生偷窃行为的蜣螂的位置$z_i(m)$更新定义为

$$z_i(m+1)=Z^b+S\times g\times$$
$$\left(\left|z_i(m)-Z^*\right|+\left|z_i(m)-Z^b\right|\right), \tag{13}$$

式中：S是一个常量；g是一个大小为$1\times D$的服从正态分布的随机的常向量。

1.3　用于流感预测的MRC-DBO-LSTM模型

手动获取模型参数的最优解是非常困难和耗费时间的，因此，本研究使用蜣螂优化算法对LSTM的超参数进行微调，利用蜣螂优化算法收敛快的特性，可以有效提升模型效率。使用MRC-DBO-LSTM预测流感的流程如图 3 所示。

具体预测步骤如下：

1）对数据进行预处理。通过皮尔森相关系数筛选与预测地强相关的区域，获取强相关区域的历史数据与本地历史数据，作为特征输入LSTM网络。

2）LSTM模型参数初始化。将数据输入模型后，在LSTM模型的每次迭代内，首先利用DBO算法对每只蜣螂参数进行初始化；其次，每只蜣螂按照优化规则更新位置效果最好即评价指标最好的蜣螂的位置；最后，在DBO算法迭代结束后，将效果最优的蜣螂参数传给LSTM模型。

3）模型预测及评价。利用优化参数后的LSTM模型对数据进行拟合及预测，输出预测结果，对结果进行评价并记录，直至迭代结束，输出效果最好的结果。

DBO的每只蜣螂包含4个维度Z_i[隐藏层次，

隐藏层节点数，迭代次数，周期数]，即同时对需要优化4个超参数进行寻优，其寻优范围分别为[1, 6]，[1, 12]，[1, 500]，[1, 12]。

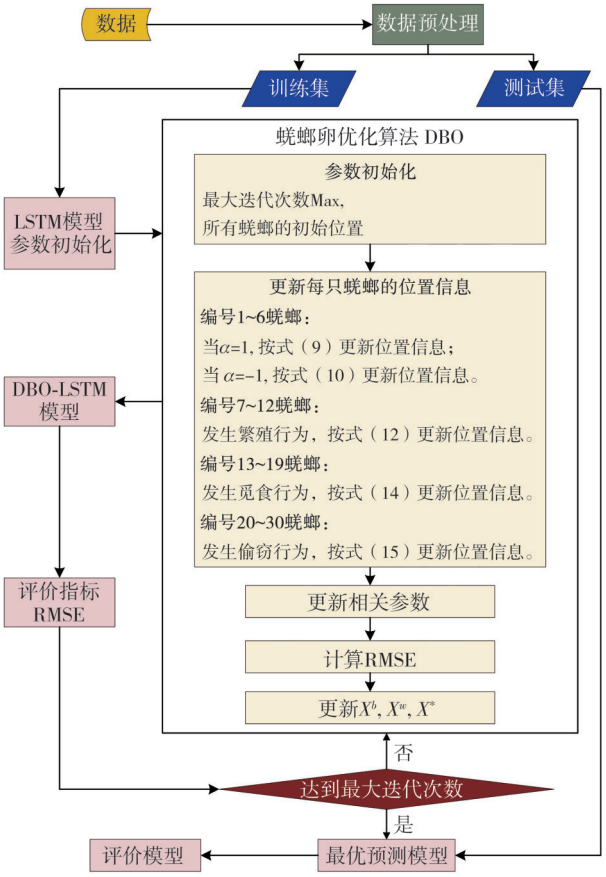

图3 MRC-DBO-LSTM模型流感发病率预测流程图

Fig. 3 Flowchart of influenza incidence prediction in the MRC-DBO-LSTM model

2 实验与分析

2.1 实验环境与数据来源

本文采用pytorch1.13+python3.8实现预测建模过程。实验环境：处理器Intel(R) Core(TM) i7-7700 CPU 3.60GHz，内存16 GB，操作系统为Windows 10。

流感样病例数据来源于公共卫生科学数据中心网站(https://www.phsciencedata.cn)，该网站公开发布了2014年至2018年的流行性感冒数据。本研究采用的数据样本时间为2014年1月至2018年12月。该网站统计的发病率数学定义为

$$\text{发病率}=\frac{\text{该地区发病人数}}{\text{该地区总人口数}}\times 10^5,$$

2.2 数据分析与处理

如何选择适当的区域作为共同输入是本研究的一个重点。为了量化上文所述数据特点，首先采用皮尔逊相关系数(Person Correlation Coefficient，PCC)来分析各区域间历史数据的相关性。

皮尔逊相关系数是一种用于度量两个变量之间相关性的系数。本研究通过皮尔逊相关系数来筛选与待预测区域相关性更高的其他区域。皮尔逊相关系数数学定义为

$$\gamma_{ab}=\frac{\sum a_ib_i-n\bar{a}\bar{b}}{(n-1)s_as_b}=\frac{n\sum a_ib_i-\sum a_i\sum b_i}{\sqrt{n\sum a_i^2-\left(\sum a_i\right)^2}\sqrt{n\sum b_i^2-\left(\sum b_i\right)^2}},\quad(14)$$

式中：a和b分别表示两个区域；$\gamma_{ab}\in[-1,1]$，其取值与相关性表述如图4所示。

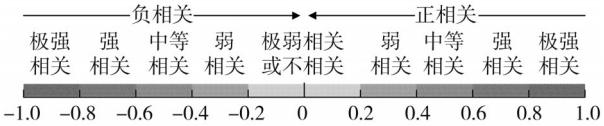

图4 皮尔逊相关系数与相关性表述

Fig. 4 Pearson correlation coefficient and correlation expression

按照相关性的排名，选取排名前5的区域作为备选输入。为了确保数据的统一性，使用MinMaxScaler函数对所有数据进行归一化处理。数据处理流程如图5所示。

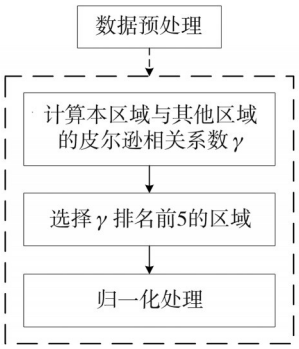

图5 数据处理流程

Fig. 5 Data processing flow

与山西省相关性排名前5的区域包括内蒙古自治区、吉林省、山东省、甘肃省和宁夏回族自治区，其皮尔逊相关系数分别为0.833，0.802，0.853，0.885和0.876，均为强相关，其历史数据如图6所示。

由图6可知，流感发病率数据呈现明显的周期性，即在每年年末至次年年初出现高峰。另外，尽管这些区域发病率数值不是完全一致的，但流感发病率变化趋势大体是相似的。

将数据分为训练集和测试集，训练集数据包含从2014年1月至2018年4月，共计52个月的数据，

本研究根据优化得到的批量数向后滚动预测，预测2018年5月至2018年12月，共计8个月的数据。

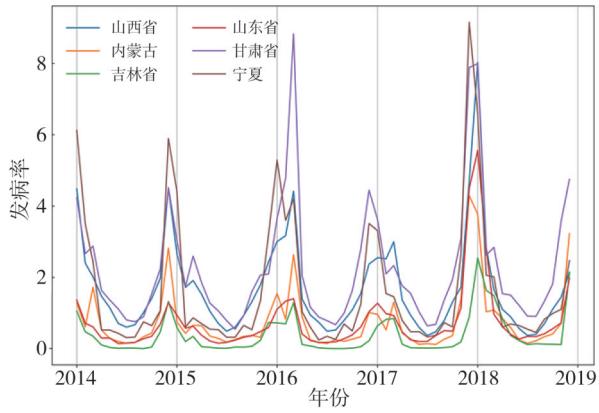

图6　2014—2018年山西省与皮尔逊相关系数排名前5区域流行性感冒的发病率

Fig. 6　The incidence of influenza in Shanxi Province and its Pearson correlation coefficient ranked among the top 5 regions during the year of 2014—2018

2.3　评价指标

均方根误差（Root Mean Squared Error，RMSE）为

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2}。\tag{15}$$

平均绝对误差（Mean Absolute Deviation，MAE）为

$$R_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|。\tag{16}$$

均方误差（Mean Squared Error，MSE）为

$$R_{\mathrm{MSE}}=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2。\tag{17}$$

上述指标越小，则预测值与真实值的误差越小，即预测越准确。

R^2即R-Squared，也称为拟合优度，如式（19）所示。R^2越接近于1，模型效果越好。

$$R^2=\frac{\sum\left(\hat{y}_i-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2},\tag{18}$$

式中：$\hat{y}_i$为预测值；y_i为观测值；$\bar{y}$为观测值的平均值。

2.4　实验和结果分析

通过多次实验，本研究中蜣螂优化算法的超参数设定为$n=30$，$k_1=6$，$k_2=6$，$k_3=7$，$k_4=11$，即种群内有30只蜣螂，其中，6只滚球，6只觅食，7只繁殖，11只偷窃。

2.4.1　各模型对比

为了评估本研究所建立的MRC-DBO-LSTM模型的性能，使用4个基线模型进行对比，分别是ARIMA模型、季节性差分自回归滑动平均模型（Seasonal Autoregressive Integrated Moving Average，SARIMA）、Holt-winters模型，以及没有使用DBO优化算法的LSTM模型，预测结果如图7所示，图中垂直虚线前为拟合值，垂直虚线后为预测值。

ARIMA、SARIMA、Holt-Winters模型是3个传统的时间序列预测模型。在本文中，按照使用两模型预测时间序列的步骤对其进行了参数优化，得到两模型的最优解为ARIMA（2，0，0）、SARIMA（0，0，1）×（0，0，1，12）和Holt-Winters（加法，12）。第4个基线模型是没有使用DBO优化的LSTM模型，进行500次迭代，使用前12个月的数据预测下一月和每层12个节点的6层隐藏层等参数设置来训练该模型，并将其与其他模型进行对比，具体的参数设置和对比结果如表1所示。

表1　各模型与基线模型的对比

Tab. 1　Comparison of different models with baseline model

模型	评价指标			
	RMSE	MAE	MSE	R^2
ARIMA	0.990 4	0.885 3	0.980 9	−2.067 5
SARIMA	0.551 0	0.314 8	0.303 6	0.050 4
Holt-Wihters（加法，12）	0.640 7	0.505 5	0.410 5	−0.283 8
LSTM	0.114 5	0.092 7	0.013 1	0.958 9
MRC-DBO-LSTM	0.061 7	0.054 5	0.003 8	0.988 0

实验结果表明，尽管时间序列预测模型在处理相关问题时具有一定的实用性，但它们仍存在显著的局限性。其中，ARIMA模型在所有评估模型中的表现最为薄弱，预测能力不足。相比之下，Holt-Winters模型虽略优于ARIMA模型，但整体性能仍处于较低水平。具体地，Holt-Winters与ARIMA模型相比，RMSE降低了35.3%，MAE降低了43.0%，MSE降低了58.2%，R^2提升了86.3%。SARIMA模型尽管相较于上述两者有所提升，但在整体比较上仍然未能展现最佳性能。SARIMA模型与Holt-Winters模型相比，RMSE降低了14.0%，MAE降低了37.7%，MSE降低了26.0%，R_2提升了117.8%。LSTM模型与前三者相比，展现出了更大的优势。LSTM模型与SARIMA模型相比，RMSE降低了79.2%，MAE降低了70.6%，MSE降低了

95.7%，R^2提升了94.7%。本研究所构建的MRC-DBO-LSTM模型，在RMSE、MAE和MSE三个评判标准下，均表现最好，其值分别为0.062 0，0.054 5和0.003 8，与LSTM模型相比，RMSE降低了46.1%，MAE降低了41.2%，MSE降低了70.1%，R^2提升了2.9%。

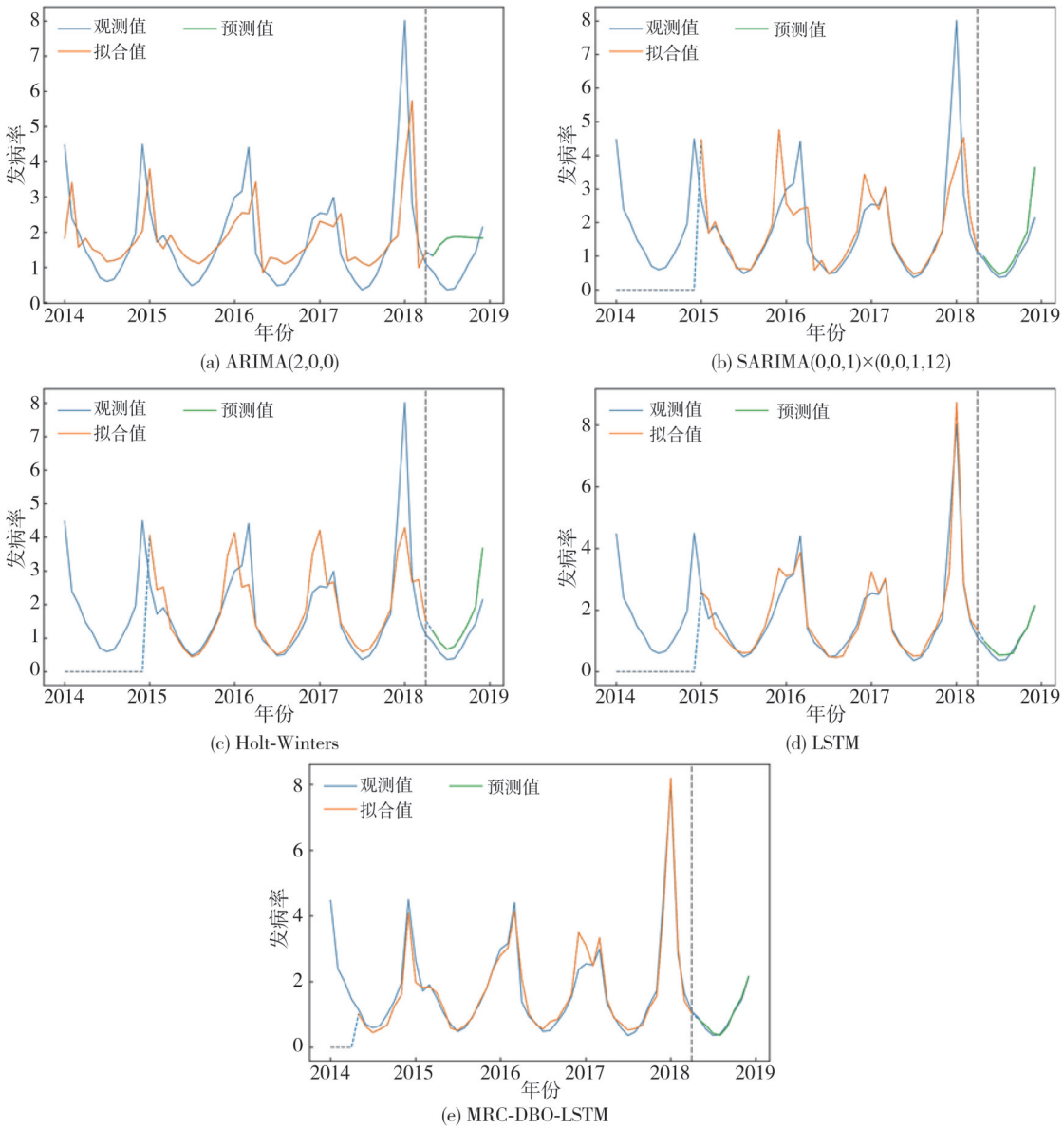

图7 基线模型与MRC-DBO-LSTM模型预测结果

Fig. 7 Prediction results of baseline model and MRC-DBO-LSTM model

2.4.2 多区域协同预测

为了验证学习相关区域的历史数据进行预测的必要性，设置3个模型进行对比。Model_1为仅学习山西省历史数据的DBO-LSTM模型，Model_2为山西省本地与$r\geqslant 0.85$区域历史数据作为共同特征输入的MRC-DBO-LSTM模型，Model_3为山西省本地与相关系数排名前5区域的历史数据作为共同特征输入的MRC-DBO-LSTM模型。由实验可知，Model_1在学习过去7周历史数据时预测效果最佳，Model_2在学习过去3周历史数据时预测效果最佳，Model_3在学习过去4周历史数据时预测效果最佳。具体预测结果和评价指标分别如图8和表2所示。

表2 实验结果评价

Tab. 2 Evaluation of experimental results

方法	评价指标			
	RMSE	MAE	MSE	R^2
Model_1	0.086 1	0.065 5	0.007 4	0.976 8
Model_2	0.079 1	0.062 9	0.006 3	0.980 4
Model_3	**0.061 7**	**0.054 5**	**0.003 8**	**0.988 0**

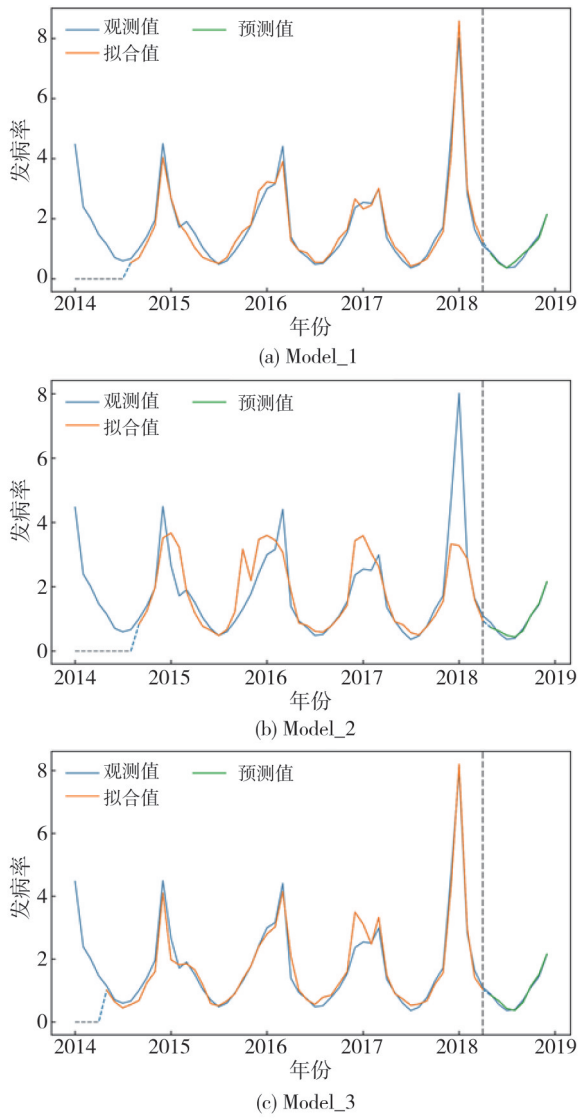

图 8　Model_1，Model_2 和 Model_3 预测结果

Fig. 8　Results of Model 1, Model 2 and Model 3

实验结果显示，LSTM 模型在流感数据预测任务上表现出了显著的优越性。无论是否将区域协同信息纳入考虑，LSTM 模型均优于先前对比的所有其他模型，在仅使用山西省流感历史数据进行训练和预测的情况下，该模型的表现相较于引入了区域协同信息的情况是最低的。具体地，Model_2 与 Model_1 相比，RMSE，MAE 和 MSE 分别降低了 8.1%，4.0% 和 14.9%，R^2 提升了 0.4%；Model_3 与 Model_1 相比，RMSE，MAE 和 MSE 分别降低了 28.5%，16.8% 和 48.6%，R^2 提升了 1.1%。

进一步的研究发现，当选择与山西省流感数据具有较高相关性的前 5 个区域的历史数据共同作为输入来训练 LSTM 模型时，其预测性能达到了最优水平。然而，值得注意是若基于 0.85 的相关性阈值来选取输入数据，即只包括那些与山西省流感数据相关性超过 0.85 的区域历史数据构建模型的情况下，模型的表现并未达到采用排名前 5 区域数据时的良好效果。具体地，Model_3 与 Model_2 相比，RMSE，MAE 和 MSE 分别降低了 22.0%，13.4% 和 39.7%，R^2 提升了 0.8%。这说明在整合多区域数据时，单纯依赖高相关性阈值可能不足以优化模型性能。

总之，LSTM 模型在流感预测中展现出明显优势，无论是否结合区域协同信息，其性能均优于对比的其他时间序列模型和部分机器学习模型。值得注意的是仅使用山西省数据训练的 LSTM 模型表现最差，而当引入与山西省流感数据相关系数排名前 5 的区域历史数据进行共同预测时，模型效果最佳。然而，若选择与山西流感数据相关性超过 0.85 的所有区域作为输入，则并未达到采用排名前 5 区域时的效果，这说明在融合多区域数据时，有针对性地选择具有高影响力的特定区域数据对提升模型预测精度至关重要。

总体而言，针对具有较明显季节性周期性的时间序列，可以捕获季节性因素的模型比仅捕获周期性因素的模型效果更佳。同时，实验结果充分证明了机器学习方法优于传统时间序列预测方法，即与 ARIMA 模型、SARIMA 模型和 LSTM 模型相比，不论是使用多区域历史数据共同预测的 MRC-DBO-LSTM 模型，还是仅使用本地数据预测的 DBO-LSTM，均表现良好。通过筛选特征来提高特征维度，即利用多区域历史数据共同预测，可以进一步提升模型效果。

3　结　语

流感作为一种急性呼吸道疾病，在爆发后会快速传播，并对公共卫生造成重大威胁。因此，迫切需要可靠的预测模型来帮助预测流感的流行情况。本文针对山西省的流感发病率进行了预测研究。通过相关性分析和特征筛选，以及构建了采用蜣螂优化算法进行优化的长短期记忆网络（MRC-DBO-LSTM）模型，提升了模型的预测准确率，对山西省流感发病率做出了较为准确的预测。由于历史数据的不稳定性，本研究仅采用多区域数据共同预测的方式拓宽特征，实验结果表明，这一方法是有效的，所以在今后的研究中还可以通过纳入更多因素，如气象因素、网络信息、

人口流动信息等来更新本研究构建的模型。这样可以更加全面和准确地预测流感发病率。

参考文献:

[1] FAHLENA H, KUSDIANTARA R, NURAINI N, et al. Dynamical analysis of two-pathogen coinfection in influenza and other respiratory diseases[J]. Chaos, Solitons & Fractals, 2022, 155: 111727.

[2] LIU Y, ZHAN L, WANG Y, et al. Improved influenza diagnostics through thermal contrast amplification [J]. Diagnostics, 2021, 11(3): 462.

[3] KRISTIANI E, CHEN Y A, YANG C T, et al. Using deep ensemble for influenza-like illness consultation rate prediction[J]. Future Generation Computer Systems, 2021, 117: 369-386.

[4] KEELING M J, DYSON L, TILDESLEY M J, et al. Comparison of the 2021 COVID-19 roadmap projections against public health data in England[J]. Nature Communications, 2022, 13(1): 4924.

[5] THOMAS M, ROOTZÉN H. Real-time prediction of severe influenza epidemics using extreme value statistics [J]. Journal of the Royal Statistical Society Series C: Applied Statistics, 2022, 71(2): 376-394.

[6] 郑月彬, 朱国魂. 时间序列模型在流感预测中的应用[J]. 仪器仪表用户, 2019, 26(4): 53-56.
ZHENG Yuebin, ZHU Guohun. Application of time series model in influenza prediction[J]. Instrumentation, 2019, 26(4): 53-56. (in Chinese)

[7] PICHLER M, HARTIG F. Machine learning and deep learning—a review for ecologists[J]. Methods in Ecology and Evolution, 2023, 14(4): 994-1016.

[8] ZHANG B, WANG Q, GAO Z, et al. Temporal grafter network: Rethinking LSTM for effective video recognition[J]. Neurocomputing, 2022, 505: 276-288.

[9] WANG P, ZHENG X, AI G, et al. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran[J]. Chaos, Solitons & Fractals, 2020, 140: 110214.

[10] TSAN Y T, CHEN D Y, LIU P Y, et al. The prediction of influenza-like illness and respiratory disease using LSTM and ARIMA[J]. International Journal of Environmental Research and Public Health, 2022, 19(3): 1858.

[11] BANSAL H, BHATT G, MALHOTRA P, et al. Systematic generalization in neural networks-based multivariate time series forecasting models[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[12] KHADEM S A, BENSEBAA F, PELLETIER N. Optimized feed-forward neural networks to address CO_2-equivalent emissions data gaps-application to emissions prediction for unit processes of fuel life cycle inventories for Canadian provinces [J]. Journal of Cleaner Production, 2022, 332: 130053.

[13] WEI J, ZHANG X, JI Z, et al. DPLRS: Distributed population learning rate schedule[J]. Future Generation Computer Systems, 2022, 132: 40-50.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[15] LI C, LIU S, ZHANG Q, et al. Combining Raman spectroscopy and machine learning to assist early diagnosis of gastric cancer[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 287: 122049.

[16] YI C, WEI B, ZHU J, et al. Mordo: Silent command recognition through lightweight around-ear biosensors [J]. IEEE Internet of Things Journal, 2023, 10(1): 763-773.

[17] ZHANG Y, SONG Y, WEI G. A feature-enhanced long short-term memory network combined with residual-driven ν support vector regression for financial market prediction [J]. Engineering Applications of Artificial Intelligence, 2023, 118: 105663.

[18] PRADHAN A K, DAS K, MISHRA D, et al. Optimizing CNN-LSTM hybrid classifier using HCA for biomedical image classification[J]. Expert Systems, 2023, 40(5): e13235.

[19] XUE J, SHEN B. Dung beetle optimizer: A new meta-heuristic algorithm for global optimization [J]. The Journal of Supercomputing, 2023, 79(7): 7305-7336.