DOI:10.11737/j.issn.1003-0239.2023.05.001

基于3种机器学习算法的台风频数预测

荣新[1]，覃卫坚[2*]，韦文山[1]

（1.广西民族大学 电子信息学院,广西 南宁 530000;2.广西气候中心,广西 南宁 530022）

摘 要： 为了提高影响广西台风频数的年度预测准确率,利用中国气象局上海台风研究所提供的1951—2020年影响广西的台风样本数据、国家气候中心提供的88项大气环流特征量和26项海温指数资料,使用相关方法找出高影响因子。针对影响台风物理因素的复杂性,为了获取更综合的预测因子信息,使用随机森林对影响因子进行二次筛选,建立基于随机森林、支持向量回归和循环门单元(GRU)3种机器学习算法的影响广西台风频数气候预测模型。实验结果表明:使用随机森林二次筛选得到的因子的建模预测效果明显提高,机器学习算法预测效果整体高于岭回归方法,其中GRU预测效果最好,绝对误差较岭回归方法减少10.30%,其次为随机森林和支持向量回归,误差分别减少9.44%和7.47%。

关键词： 影响广西台风频数;特征选择;随机森林;支持向量回归;循环门单元

中图分类号：P444 **文献标识码：**A **文章编号：**1003-0239(2023)05-0001-09

0 引言

台风是一种破坏性极大的气象灾害,台风期间常伴随狂风、暴雨、风暴潮等现象。广西位于中国南部沿海地区,每年平均受5个台风的影响。台风常给广西造成严重的经济损失和人员伤亡,如2001年7月2—9日,台风“榴莲”和台风“尤特”给广西带来强降雨,导致左江、右江、邕江、郁江、浔江洪水泛滥,百色市遭遇了百年不遇的洪涝灾害,1 650万人受灾,24人死亡,直接经济损失达159亿元。因此,提高预测影响台风频数的气候要素的能力,对提前做好台风防范工作、减少灾害损失具有重要意义。

台风预报方法研究一直受到人们的关注,传统的线性回归、广义加性模式、动态统计混合模式等统计预报方法在热带气旋活动预测中取得了巨大的成功[1-8]。近年来,基于机器学习和人工智能算法在处理非线性问题上有较好的自适应学习能力,被广泛应用于天气预报中[9-15],例如:在探索台风生成、路径以及强度时,CHEN等[16]关注了大气和海洋变量的时空相关性,将台风的形成和强度预报分别定义为时空序列预报的分类和回归问题,建立了卷积神经网络-长短期记忆网络(Convolutional Neural Networks-Long Short-Term Memory,CNN-LSTM)混合预测模型;高珊等[17]、徐光宁[18]分别运用LSTM和深度学习建立台风强度预测模型;HAGHROOSTA等[19]在台风强度预测上证明了运用自适应模糊神经网络(Adaptive-Network-based Fuzzy Inference Systems,ANFIS)方法优于单独的人工神经网络方法;GAO等[20]建立了基于LSTM的台风路径预报模型,得到了理想的6～24 h的台风路径预报结果;SONG等[21]结合两次数据降维,建立了基于支持向量回归(Support Vector Regression,SVR)方法的台风路径预报;LIU等[22]通过粒子群投影寻踪和模糊数学计算权重来优化预测因子,建立了基于自然正交展开和组合权值的非线性小波神经网络模型,同样TAN等[23]使用最小绝对收缩和选择算子方法

收稿日期：2022-05-11。
基金项目：广西科技计划项目(桂科AB21075005);广西自然科学基金(2019GXNSFAA245048)。
作者简介：荣新(1997-),女,硕士在读,主要从事数据处理和机器学习应用研究。E-mail:rxsdhzcx@163.com
*通信作者：覃卫坚(1971-),男,教授级高级工程师,博士,主要从事气候预测技术研究。E-mail:qinweijian2008@126.com

获取预测因子并结合随机森林(Random Forest,RF)建立了预测方案,两者在热带气旋频数预测上都取得了较好的预测结果。综上可见机器学习和人工智能算法多应用于台风路径、强度等天气预报中,而在台风频数预测的应用中还不多见。本文以影响广西的台风年频数作为研究对象,针对台风频数预测的非线性特点,汲取当前人工智能的研究成果筛选最优的预测因子,在数据处理上运用具有优越选择特征的随机森林方法进行因子二次筛选来得到最优预测因子,使用SVR、RF以及循环门单元(Gated Recurrent Unit,GRU)3种机器学习算法建立台风个数预测模型,综合对比分析得出最优算法,为年度台风频数预测提供新的可行性方法。

1 资料与方法

1.1 资料来源

影响广西台风观测数据(1951—2020年)来源于中国气象局上海台风研究所提供的台风年鉴和热带气旋年鉴。台风等级包括台风、热带风暴及热带低压,影响广西的台风定义为进入19°N以北、112°E以西的台风[24]。

国家气候中心提供了1951—2020年88项大气环流特征向量和26项海温指数资料(获取地址:http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php)。对上述资料进行归一化预处理,归一化公式为:

$$f'_i = \frac{f_i - \min(k)}{\max(k) - \min(k)} \tag{1}$$

式中:i表示第i年;k表示第k个特征因子。

1.2 岭回归模型方法

线性回归分析是探索因变量和自变量关系程度的统计方法,通过将真实值与预测值的平方误差最小化,可建立反应变数(Y)和解释变数(X)之间的关系模型。最小二乘法代价函数为:

$$h_\alpha(\alpha) = \arg\min(Y - X\alpha)^{\mathrm{T}}(Y - X\alpha) \tag{2}$$

式中:α是线性回归系数。

解出系数α为:

$$\alpha = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

式中:$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$为满秩矩阵。

为了解决线性回归分析中过拟合的问题,岭回归方法(Ridge Regression,RR)在建模时加入正则化项,在矩阵$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$的对角线元素上加入岭系数$\sigma$,代价函数$h_\alpha(\alpha)$转变为:

$$h_\alpha(\alpha) = \sum_{k=1}^{m}(Y_k - X_k\alpha) + \sigma\|\alpha\|_2^2 \tag{4}$$

得到系数α的解:

$$\alpha = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \sigma\boldsymbol{I}_m)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{5}$$

式中:σ是超参数,可通过调节σ的值来改变对α的惩罚强度。

1.3 SVR模型方法

SVR模型方法是一种用于分类和回归、有监督的机器学习算法,在处理高维问题方面具有较强的鲁棒性。SVR的主要思想是利用支持向量机找到可能的最佳预测模型,并容忍一些预测误差[25]。首先需要构建一个样本标签,选择最有影响力的样本集构造超平面,方程表示为:

$$g(x) = \boldsymbol{w}^{\mathrm{T}}x + b \tag{6}$$

式中:$\boldsymbol{w}$表示加权矩阵;b为偏置项。当且仅当训练样本落入划分的超平面外时计算损失,将回归风险最小化为:

$$R_{\min}(g) = \min\frac{1}{2}\|\boldsymbol{w}\|^2 + B\sum_{k=1}^{n}l_\theta(g(x_k) - y_k) \tag{7}$$

式中:B为正则化常数;$g(x_k)$为第k个样本的预测值;y_k为第k个真实值;l_θ为不敏感损失函数,其中θ为容忍偏差。

$$l_\theta(m) = \begin{cases} 0, & |m| \leqslant \theta \\ |m| - \theta, & |m| > \theta \end{cases} \tag{8}$$

本模型引入高斯核函数$G(x,x_k)$,可将样本从原始空间映射到更高维的特征空间以获得更高的预测精度,超平面所对应的模型变为:

$$g(x) = \sum_{k=1}^{n}(\hat{\alpha}_k - \alpha_k)G(x,x_k) + b \tag{9}$$

式中:$\hat{\alpha}_k$和α_k为拉格朗日乘子。

1.4 RF模型方法

RF模型方法是决策树方法的改进[26]。RF算法由许多决策树组成,把多个决策树的计算结果进行平均作为最后的输出结果。基本流程见图1。对于给定的原始数据$D(x_k, k \in 1,2,3,\cdots,n)$:

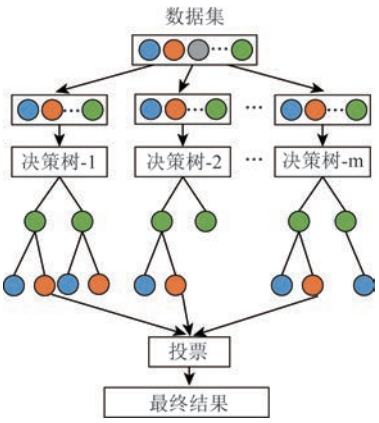

图1　RF方法流程图

Fig.1　RF method flow chart

①首先在原始数据D中有放回的随机抽样,生成m个子元组,保证每个子元组数量等于总数据集数。

②在建立决策树的过程中随机抽取备用特征,把最优特征子集划分为局部训练集,构造m棵决策树,剩余样本形成袋外数据(OOB)用于估计随机森林的拟合度。构造决策树使用基尼指数($Gini$)[27]最小化准则进行分裂,$Gini$值越小,数据集的纯度就越高。$Gini$指数可表示为:

$$Gini(D)=1-\sum_{k=1}^{n}a_k^2 \tag{10}$$

式中:a_k为训练集中样本属于某一类的概率。

③特征选择。

a.计算特征p_i在节点j中的基尼指数变化值,公式为:

$$VIM_{p_i j}=Gini(j)-Gini(l)-Gini(r) \tag{11}$$

式中:$Gini(j)$表示分枝前的基尼指数;$Gini(l)$和$Gini(r)$则为节点j分枝后产生的两个新节点的基尼指数。

b.计算特征p_i在第z棵决策树上的基尼指数变化量:

$$VIM_{zp_i}=\sum_{j\in N}VIM_{p_i j} \tag{12}$$

式中:N为节点集合。

c.求每个特征对随机森林每棵树的贡献值,即重要程度:

$$VIM_{p_i}=\frac{\sum_{z=1}^{m}VIM_{zp_i}}{\sum_{z=1}^{m}VIM_z} \tag{13}$$

d.对每一个节点求得贡献值后进行比较和排序[28]。

④将m棵树组成随机森林,求平均值并作为最后输出的预测结果。

1.5　GRU模型方法

GRU模型方法是一种高级的长短期记忆技术[29],是LSTM算法的一个变体,GRU简化了LSTM算法的3层门循环,将单元状态与输出状态合二为一,仅保留了两层的门循环即重置门和更新门。基本流程见图2。

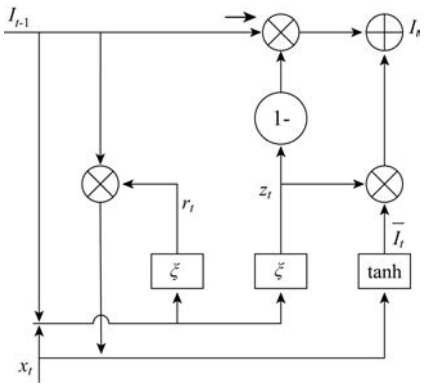

图2　GRU方法基本流程图

Fig.2　GRU method flow chart

GRU方法的公式表示为:

$$z_t=\xi(w_z\cdot[I_{t-1},x_t]) \tag{14}$$

$$r_t=\xi(w_r\cdot[I_{t-1},x_t]) \tag{15}$$

$$\bar{I}_t=\tanh(w\cdot[r_t\cdot I_{t-1},x_t]) \tag{16}$$

$$I=(1-z_t)\cdot I_{t-1}+z_t\cdot\bar{I}_t \tag{17}$$

式中:z_t表示更新门;r_t表示重置门;w表示循环层权重;x_t为t时刻的输入;I_t表示t时刻的输出状态。使用GRU算法可以有效解决数据序列在训练过程中出现的梯度消失和爆炸问题。

2　预测模型及效果检验

2.1　筛选预报因子

2.1.1　初次筛选因子

给定一组数据$D=\{(X_k,Y_k)\}$,$k\in n$,设相关系数为r,r可表示为:

$$r=\frac{\sum_{k=1}^{n}(X_k-\bar{X})(Y_k-\bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k-\bar{X})^2}\sqrt{\sum_{k=1}^{n}(Y_k-\bar{Y})^2}} \quad (18)$$

式中：$\bar{X}$为X的平均值；$\bar{Y}$为Y的平均值。

计算1951—2015年影响广西的台风年频数时间序列与同一年、前一年各月各种大气环流和海温指数的相关系数，初步筛选出相关系数绝对值达到0.4、通过水平为0.01的显著性检验的环流特征量和海温指数作为预报因子，共得到24个预报因子（见表1）。

2.1.2 二次筛选因子

通常预报因子之间的多重共线性会导致解释变量之间出现相当大的冗余，为了能够更好地反映预报因子场的综合信息，需要在已选定的一批因子中得到最优因子，进一步提高预测精度，降低计算的复杂度。使用RF方法对上节得到的24个预报因子进行二次筛选，使用经过训练的RF模型，计算每个因子的重要程度，按照从大到小的顺序逐个输出，筛选出重要性值相对较高的因子。由于前3位因子的重要性值最高，之后因子的重要性值有较大幅度的减小（如第四位重要性值仅为0.030 33），因此最后得到3个预报因子（见表2），分别为前一年9月欧亚纬向环流指数、同一年2月NINO 1+2区海表温度距平指数、前一年6月大西洋经向模式风指数

表1 初选得到的预报因子

Tab.1 Predictors obtained from the primary selection

因子序号	因子名称	相关系数绝对值
1	前一年9月欧亚纬向环流指数	0.41
2	前一年10月登陆台风	0.41
3	前一年6月NINO 1+2区海表温度距平指数	0.45
4	前一年8月NINO 1+2区海表温度距平指数	0.42
5	前一年9月NINO 1+2区海表温度距平指数	0.41
6	前一年10月NINO 1+2区海表温度距平指数	0.44
7	前一年12月NINO 1+2区海表温度距平指数	0.46
8	同一年1月NINO 1+2区海表温度距平指数	0.45
9	同一年2月NINO 1+2区海表温度距平指数	0.42
10	前一年12月NINO 4区海表温度距平指数	0.44
11	同一年2月NINO 4区海表温度距平指数	0.42
12	同年1月NINO Z区海表温度距平海表温度指数	0.45
13	同年2月NINO Z区海表温度距平海表温度指数	0.42
14	同一年1月北极涛动指数	0.40
15	前一年4月太平洋-北美遥相关型指数	0.44
16	前一年5月太平洋-北美遥相关型指数	0.40
17	前一年4月大西洋经向模式（AMM）风指数	0.40
18	前一年5月大西洋经向模式（AMM）风指数	0.43
19	前一年6月大西洋经向模式（AMM）风指数	0.46
20	前一年8月大西洋经向模式（AMM）风指数	0.42
21	前一年10月大西洋经向模式（AMM）风指数	0.42
22	同一年1月大西洋经向模式（AMM）风指数	0.40
23	前一年12月西半球暖池指数指数	0.41
24	同一年1月大西洋多年代际振荡指数	0.41

表2　二次筛选得到的特征因子

Tab.2　Characteristic factors obtained from the secondary screening

因子序号	因子名称	重要性值
1	前一年9月欧亚纬向环流指数	0.087 86
2	同一年2月NINO 1+2区海表温度距平指数	0.087 85
3	前一年6月大西洋经向模式(AMM)风指数	0.109 65

(Atlantic Meridional Mode, AMM),该区域多处于西太平洋台风生成的区域。

2.2　RR模型预报

设台风样本数据为$D(x_k, k \in 1,2,3,\cdots,n)$,绝对误差计算公式为:

$$\text{绝对误差} = \text{预测值} - \text{实况值} \tag{19}$$

$$\text{平均绝对误差} = \frac{1}{n}\sum_{k=1}^{n}\left|\text{绝对误差}\right| \tag{20}$$

相对误差计算公式为:

$$\text{相对误差} = \frac{\left|\text{预测值} - \text{实况值}\right|}{\text{预测值}} \times 100\% \tag{21}$$

$$\text{平均相对误差} = \frac{1}{n}\sum_{k=1}^{n}\text{相对误差} \tag{22}$$

基于初次筛选因子和二次筛选因子建立RR预报模型。利用24个因子建立预报模型,调节岭参数值为0.7时预测结果最佳,由表3可知5 a独立样本的预测平均绝对误差为1.46,平均相对误差为38.13%。利用二次筛选得到的3个因子建立RR预报模型,调节岭参数σ为0.6时,训练64 a台风样本集效果最佳(见图3),平均绝对误差为2.12;5 a独立样本预测结果见表3,预测平均值为4.31,较使用初次筛选因子的预测更接近实况平均值,平均绝对误差为1.03,平均相对误差为24.02%,较使用24个因子的RR模型预测的平均绝对误差减少0.43,平均相对误差减少14.11%,预测效果明显提高。

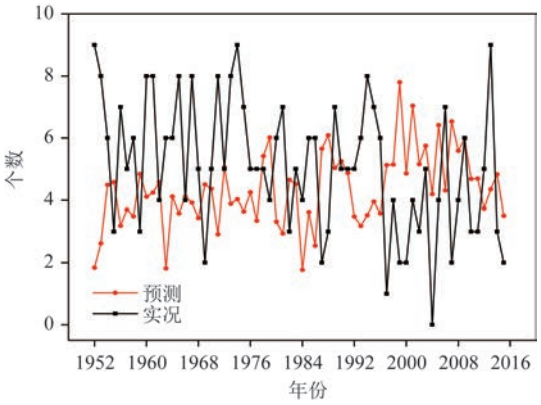

图3　RR模型训练拟合预报和实况序列

Fig.3　RR model training to fit forecasts and observations

2.3　SVR模型预报

SVR采用liblinear库来实现,选取参数惩罚函数C=0.1,损失函数explosion=2.0,最大迭代次数默认为10 000次,训练集为64 a台风个数。当预报因子为24个时,5 a独立样本预测结果见表4,预报误差绝对值平均为0.83,平均相对误差为17.28%。当

表3　2016—2020年台风频数RR模型预报结果

Tab.3　RR model forecast results of typhoon number in 2016—2020

年份	实况值	24个因子			3个因子		
		预测值	平均绝对误差	平均相对误差/%	预测值	平均绝对误差	平均相对误差/%
2016	4.00	5.42	1.42	35.5	4.53	0.53	13.25
2017	4.00	3.93	−0.07	1.75	3.78	−0.22	5.5
2018	6.00	4.70	1.30	21.67	3.92	−2.08	34.67
2019	3.00	5.31	2.31	77.00	4.01	1.01	33.67
2020	4.00	6.19	2.19	54.75	5.32	1.32	33.00
平均值	4.20	5.11	1.46	**38.13**	4.31	1.03	**24.02**

表4 2016—2020年台风频数SVR模型预报结果

Tab.4 SVR model forecast results of typhoon number in 2016—2020

年份	实况值	24个因子			3个因子		
		预测值	平均绝对误差	平均相对误差/%	预测值	平均绝对误差	平均相对误差/%
2016	4.00	3.97	−0.03	0.75	4.29	0.29	7.25
2017	4.00	3.23	−0.77	19.25	4.08	0.08	2.00
2018	6.00	3.79	−2.21	36.83	3.99	−2.01	33.50
2019	3.00	3.13	0.13	4.33	4.14	1.14	38.00
2020	4.00	2.99	−1.01	25.25	3.92	0.08	2.00
平均值	4.20	3.42	0.83	**17.28**	4.08	0.72	**16.55**

预报因子为3个时，预测样本选取5 a的数据，从训练集的拟合曲线和实况序列来看（见图4），预测值波动幅度较实况小，对极端异常台风个数的预测能力较低，如2004年无台风影响广西时预测值为4个，2013年台风达到9个时预测值也为4个，相对误差为55.56%，平均绝对误差为1.56，较岭回归减少0.56；5 a独立样本预测结果见表4，预测平均值为4.08，总体上较实况值偏小，预测平均绝对误差为0.72，平均相对误差为16.55%，较岭回归方法分别减少了0.31和7.47%，较使用24个因子预测的平均绝对误差减少0.11，平均相对误差减少0.73%。

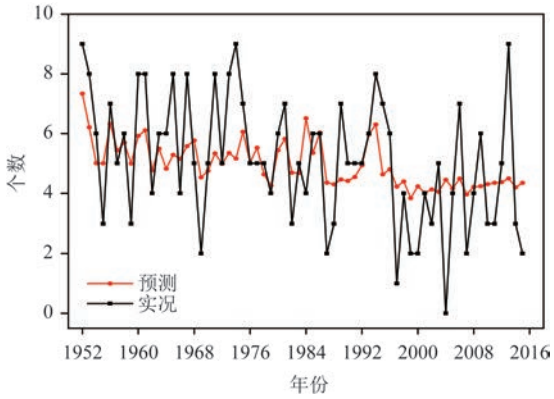

图4 SVR模型训练拟合预报和实况序列

Fig.4 SVR model training to fit forecasts and observations

2.4 RF模型预报

使用RF方法建模预报，设置n_estimators=50，n_jobs=−1，random_state=10。当预测因子为24个时，5 a独立样本预测结果见表5，平均绝对误差为0.75，平均相对误差为16.78%。利用二次筛选得到的3个预报因子，5 a独立样本预测值平均为3.65，总体上比实况值略偏小，预测平均绝对误差为0.68，平均相对误差为14.58%，分别比岭回归方法减少了0.35和9.44%，比初次选取的因子预测平均绝对误差减少0.07，平均相对误差减少2.2%；RF方法训练的拟合曲线和实况序列见图5，预测值和实况值基

表5 2016—2020年台风频数RF模型预报结果

Tab.5 RF model forecast results of typhoon number in 2016—2020

年份	实况值	24个因子			3个因子		
		预测值	平均绝对误差	平均相对误差/%	预测值	平均绝对误差	平均相对误差/%
2016	4.00	4.50	0.50	12.50	4.02	0.02	0.50
2017	4.00	3.76	−0.24	6.00	3.32	−0.68	17.00
2018	6.00	4.04	−1.96	32.67	4.31	−1.69	28.17
2019	3.00	3.81	0.81	27.00	3.30	0.30	10.00
2020	4.00	3.77	−0.23	5.75	3.31	0.69	17.25
平均值	4.20	3.98	0.75	16.78	3.65	0.68	14.58

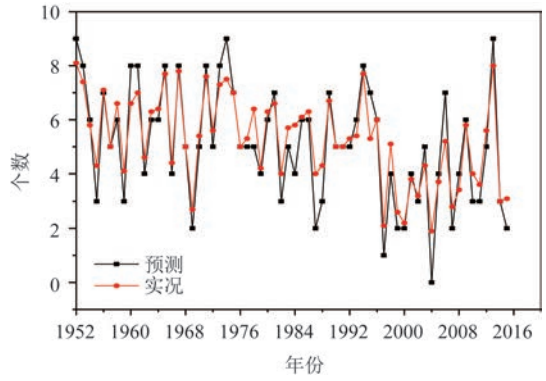

图5　RF模型训练集拟合预报和实况序列

Fig.5　RF model training to fit forecasts and observations

本吻合,尤其是对极端年份的预测能力较SVR模型和RR模型有较大提高,如2013年台风为9个,RF方法预测值为8个,非常接近;拟合预测平均绝对误差为0.64,较岭回归减少1.48。

2.5　GRU模型预报

本模型基于tensorflow搭建,样本集数量较少,只包括一个隐藏层,内含20个神经元,算法优化器选用rmsprop,使用均方误差进行误差衡量。由于特征因子之间的数值差异很小,这里不对数据进行标准化处理,而使用层处理函数进行迭代,初次迭代2 000次,步长为5,在迭代次数达到1 000次左右时拟合基本趋于平稳,因此二次实验模型迭代1 200次。当预测因子为24个时,得到5 a独立样本的预报平均绝对误差为0.98,平均相对误差为19.93%。当预测因子为3个时,5 a独立样本预测值平均为4.06,预测平均绝对误差为0.59,平均相对误差为13.72%,分别较岭回归减少了0.44和10.30%,较使用初选因子预报结果的平均绝对误差减少0.39,平均相对误差减少6.21%;GRU方法训练集的拟合曲线和实况序列见图6,拟合曲线和实况基本吻合,对极端年份的预测结果接近实况。

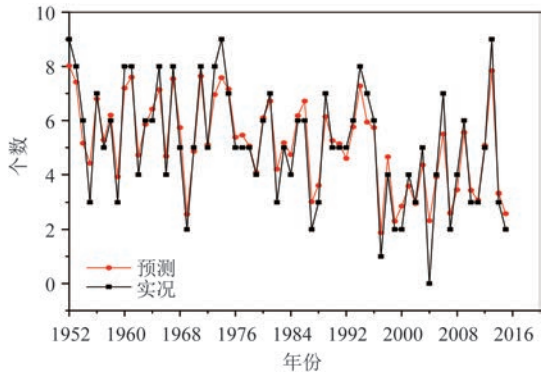

图6　GRU模型训练集拟合预报和实况序列

Fig.6　GRU model training to fit forecasts and observations

3　总结与讨论

本文计算了台风频数与88项环流特征量、26项海温指数的相关系数,初次筛选出24个高相关因子,再利用随机森林方法进行二次筛选得到3个预报因子并建立基于3种机器学习算法的预报模型,对训练样本集进行多次迭代计算,不断优化模型参数,对2016—2020年台风个数进行预测实验。岭回归方法、支持向量回归方法、随机森林方法和循环门单元方法的预测结果较使用初选因子平均相对误差分别减少14.11%、0.73%、2.2%、6.21%,可见利

表6　2016—2020年台风频数GRU模型预报结果

Tab.6　GRU model forecast results of typhoon number in 2016—2020

年份	实况值	24个因子			3个因子		
		预测值	平均绝对误差	平均相对误差/%	预测值	平均绝对误差	平均相对误差/%
2016	4.00	3.90	−0.10	2.50	4.04	0.04	1.00
2017	4.00	3.70	−0.30	7.50	3.98	−0.02	0.50
2018	6.00	3.32	−2.68	44.67	4.22	−1.78	29.67
2019	3.00	3.00	0.00	0.00	4.10	1.10	36.67
2020	4.00	2.2	−1.8	45.00	3.97	0.03	0.75
平均值	4.20	3.22	0.98	19.93	4.06	0.59	13.72

用随机森林方法对预测因子进行二次筛选是有效的，能充分发挥多信息融合的优势，在线性拟合的过程中能进一步提高数据的适应能力。由此，使用随机森林二次筛选因子建立模型，机器学习预报方法比岭回归方法的平均相对误差都有减少，其中循环门单元方法、随机森林方法、支持向量回归方法的平均相对误差分别减少10.30%，9.44%，7.47%，由此可知，机器学习方法在处理高维数据下的非线性问题上具有较大优势。在未来的工作中，我们还要考虑增加其他影响台风形成的因子，在模型分析中选择更多的预测因子，进一步优化模型参数，提高预测的精度和计算效率。

参考文献：

[1] CHAND S S, WALSH K J E. Forecasting tropical cyclone formation in the fiji region: a probit regression approach using Bayesian fitting[J]. Weather and Forecasting, 2011, 26(2): 150-165.

[2] CHU P S, ZHAO X. A Bayesian regression approach for predicting seasonal tropical cyclone activity over the central north pacific[J]. Journal of Climate, 2007, 20(15): 4002-4013.

[3] MCDONNELL K A, HOLBROOK N J. A Poisson regression model of tropical cyclogenesis for the Australian-southwest Pacific Ocean region[J]. Weather and Forecasting, 2004, 19(2): 440-455.

[4] MCDONNELL K A, HOLBROOK N J. A Poisson regression model approach to predicting tropical cyclogenesis in the Australian / southwest Pacific Ocean region using the SOI and saturated equivalent potential temperature gradient as predictors[J]. Geophysical Research Letters, 2004, 31(20): L20110.

[5] KE F. New predictors and a new prediction model for the typhoon frequency over western North Pacific[J]. Science in China Series D: Earth Sciences, 2007, 50(9): 1417-1423.

[6] MESTRE O, HALLEGATTE S. Predictors of tropical cyclone numbers and extreme hurricane intensities over the north Atlantic using generalized additive and linear models[J]. Journal of Climate, 2009, 22(3): 633-648.

[7] LI X, YANG S, WANG H, et al. A dynamical-statistical forecast model for the annual frequency of western Pacific tropical cyclones based on the NCEP Climate Forecast System version 2[J]. Journal of Geophysical Research: Atmospheres, 2013, 118(21): 12061-12074.

[8] WAHIDUZZAMAN M, CHEUNG K, LUO J J, et al. Impact assessment of Indian Ocean Dipole on the North Indian Ocean tropical cyclone prediction using a Statistical model[J]. Climate Dynamics, 2022, 58(3): 1275-1292.

[9] NONG J F, JIN L. Application of support vector machine to predict precipitation[C]//2008 7th World Congress on Intelligent Control and Automation. Chongqing: IEEE, 2008: 8975-8980.

[10] SRIVASTAVA S, ANAND N, SHARMA S, et al. Monthly rainfall prediction using various machine learning algorithms for early warning of landslide occurrence[C]//2020 International Conference for Emerging Technology. Belgaum: IEEE, 2020: 1-7.

[11] 甄亿位, 郝敏, 陆宝宏, 等. 基于随机森林的中长期降水量预测模型研究[J]. 水电能源科学, 2015, 33(6): 6-10.
ZHEN Y W, HAO M, LU B H, et al. Research of medium and long term precipitation forecasting model based on random forest [J]. Water Resources and Power, 2015, 33(6): 6-10.

[12] 覃卫坚, 陆虹, 黄志, 等. 粒子群—神经网络法在广西寒露风日数预报中的应用[J]. 气象与环境学报, 2015, 31(6): 158-162.
QIN W J, LU H, HUANG Z, et al. Application of forecasting cold dew wind day based on PSO-Fuzzy Neural Network in Guangxi province[J]. Journal of Meteorology and Environment, 2015, 31 (6): 158-162.

[13] 覃卫坚, 李耀先, 陈思蓉, 等. 粒子群-神经网络在华南夏季降水短期气候预测中应用研究[J]. 气象研究与应用, 2015, 36(2): 1-7.
QIN W J, LI Y X, CHEN S R, et al. Application on the prediction of the summer precipitation in South China basing on PSO-Artificial Neutral Network[J]. Journal of Meteorological Research and Application, 2015, 36(2): 1-7.

[14] GHAMARIADYAN M, IMTEAZ M A. Prediction of seasonal rainfall with one-year lead time using climate indices: a wavelet neural network scheme[J]. Water Resources Management, 2021, 35(15): 5347-5365.

[15] 罗芳琼, 吴建生, 金龙. 基于最小二乘支持向量机集成的降水预报模型[J]. 热带气象学报, 2011, 27(4): 577-584.
LUO F Q, WU J S, JIN L. Rainfall forecasting model based on least square support vector machine regression ensemble[J]. Journal of Tropical Meteorology, 2011, 27(4): 577-584.

[16] CHEN R, WANG X, ZHANG W M, et al. A hybrid CNN-LSTM model for typhoon formation forecasting[J]. Geoinformatica, 2019, 23(3): 375-396.

[17] 高珊, 刘峻. 基于LSTM的台风强度预测模型分析[J]. 信息与电脑, 2021, 33(11): 30-32.
GAO S, LIU J. Typhoon intensity prediction model analysis based on LSTM[J]. China Computer & Communication, 2021, 33 (11): 30-32.

[18] 徐光宁. 基于深度学习的台风路径与强度预测方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2020.
XU G N. Research on typhoon movement and intensity forecasting based on deep learning[D]. Harbin: Harbin Institute of Technology, 2020.

[19] HAGHROOSTA T, ISMAIL W R. Comparing typhoon intensity prediction with two different artificial intelligence models[J]. Evolving Systems, 2015, 6(3): 177-185.

[20] GAO S, ZHAO P, PAN B, et al. A nowcasting model for the

prediction of typhoon tracks based on a long short term memory neural network[J]. Acta Oceanologica Sinica, 2018, 37(5): 8-12.

[21] SONG H J, HUH S H, KIM J H, et al. Typhoon track prediction by a support vector machine using data reduction methods[C]// Proceedings of International Conference on Computational and Information Science. Xi'an: Springer, 2005: 503-511.

[22] LIU H X, ZHANG D L, CHEN J W, et al. Prediction of tropical cyclone frequency with a wavelet neural network model incorporating natural orthogonal expansion and combined weights [J]. Natural Hazards, 2013, 65(1): 63-78.

[23] TAN J K, LIU H X, LI M Y, et al. A prediction scheme of tropical cyclone frequency based on lasso and random forest[J]. Theoretical and Applied Climatology, 2018, 133(3): 973-983.

[24] 覃卫坚, 黄志, 李耀先. 基于海温、雪盖的影响广西热带气旋频数的气候预测模型研究[J]. 气象研究与应用, 2013, 34(3): 1-5, 32.
QIN W J, HUANG Z, LI Y X. A short-term climatic forecast model for the frequency of tropical cyclone affecting Guangxi based on SST and snow data[J]. Journal of Meteorological Research and Application, 2013, 34(3): 1-5, 32.

[25] ÜLKER E D, ÜLKER S. Modelling the currency exchange rates using support vector regression[C]//Science and Information Conference. London: Springer, 2020: 326-333.

[26] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[27] MEENA L, CHAURASIYA V K, PUROHIT N, et al. Comparison of SVM and random forest methods for online signature verification[C]//Proceedings of the 12th International Conference on Intelligent Human Computer Interaction. Daegu: Springer, 2021: 288-299.

[28] 李光华, 李俊清, 张亮, 等. 一种融合蚁群算法和随机森林的特征选择方法[J]. 计算机科学, 2019, 46(S2): 212-215.
LI G H, LI J Q, ZHANG L, et al. Feature selection method based on ant colony optimization and random forest[J]. Computer Science, 2019, 46(S2): 212-215.

[29] SAHA S, SINGH N, MOHAN B R, et al. A combined model of ARIMA-GRU to forecast stock price[C]//Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences. Singapore: Springer, 2021: 987-998.

Typhoon number prediction based on three machine learning algorithms

RONG Xin[1], QIN Weijian[2*], WEI Wenshan[1]

(1. School of Electronic Information,Guangxi Minzu University, Nanning 530006, China; 2.Guangxi Climate Center, Nanning 530022, China)

Abstract: In order to improve the prediction accuracy of annual number of typhoons affecting Guangxi, this paper uses related methods to find high impact factor based on the sample data of typhoons affecting Guangxi from 1951 to 2020 provided by Shanghai Typhoon Institute of China Meteorological Administration, the 88 atmospheric circulation feature quantities and 26 SST index data provided by the National Climate Center. In view of the complexity of physical factors in typhoon number forecasting, in order to obtain more comprehensive factor information, the random forest is used to screen the factors, and a prediction model for annual number of typhoons affecting Guangxi utilizing three machine learning algorithms, i. e. Random Forest, Support Vector Regression and Gate Recurrent Unit(GRU), is established. The results show that the prediction ability of using factors selected by Random Forest screening is significantly improved, and the prediction ability of using machine learning algorithms is higher than that of Ridge Regression method. Among them, GRU has the best prediction, and the absolute error is reduced by 10.30% compared with Ridge Regression method, followed by Random Forest and Support Vector Regression, with errors reduced by 9.44% and 7.47%, respectively.

Key words: the number of typhoons affecting Guangxi; feature selection; Random Forest; Support Vector Regression; Gated Recurrent Unit