第 30 卷第 3 期　　安 徽 建 筑 大 学 学 报　　Vol.30 No.3
2022 年 6 月　　Journal of Anhui Jianzhu University　　Jun.2022

DOI：10.11921/j.issn.2095-8382.20220312

基于时间序列的季节性气温预测研究

赵成兵，刘丹秀，谢新平，刘　静

（安徽建筑大学　数理学院，安徽　合肥　230601）

摘　要：气温的变化受风速、湿度、日照时数等因素的影响，可以通过分析这些因素预测气温的变化情况。考虑到气温序列中存在季节特性，采用 One-Hot 编码方法提取气温序列中的季节性信息，并作为随机森林模型的输入特征，对月平均气温进行拟合与预测。由于模型构建时涉及众多超参数，文中利用随机搜索和网格搜索两种算法优化模型中的超参数。结果表明：考虑季节性的随机森林模型拟合效果优于简单随机森林模型，预测数据变化趋势与实际观测基本一致，拟合精度可以达到 96.14%。经两种方法对超参数寻优之后，模型拟合精度可以达到 96.45%。

关键词：随机森林；自回归移动平均模型；平均温度；季节性；Python

中图分类号：C812；P456.2　　**文献标识码：**A　　**文章编号：**2095-8382（2022）03-083-07

Research on Seasonal Temperature Forecasting Based on Time Series

ZHAO Chengbing，LIU Danxiu，XIE Xinping，LIU Jing

（School of Mathematics and Physics，Anhui Jianzhu University，Hefei 230601，China）

Abstract：The variation of temperature is influenced by factors including wind speed，humidity，sunshine duration，which can be analyzed to predict the temperature. Considering the seasonal feature in the temperature series，the seasonal information in the temperature series was extracted with the One-Hot encoding method and used as the input variable of the random forest model to fit and forecast the monthly average temperature. Since numerous hyperparameters are involved in the model，random search algorithm and grid search algorithm are used to optimize the hyperparameters. The results show that the fitting effect of the random forest model considering seasonality is better than that of the simple random forest model，and the predicted data is basically consistent with the actual situation，with the fitting accuracy reaching 96.14%. The fitting accuracy can be up to 96.45% after the hyperparameters search by both methods.

Keywords：random forest；autoregressive integrated moving average；average temperature；seasonality；Python

气候变化对人类活动产生重要的影响，研究气候非常有必要。随着大数据时代的到来，为气象预报提供了更加科学的技术支持。由于气象数据与时间紧密相关，因而可以采用时间序列的方法对气象数据进行处理及分析。目前对于时间序列研究主要分为三个方面[1-2]：一是传统统计模型包括线性模型、自回归移动平均模型（ARIMA）等。如 Dimri 等[3]使用季节性 ARIMA 拟合气温单变量模型，达到很好的拟合效果；谭小花[4]用随机分析法对重庆市气温数据做了趋势分析，选用季节指数和 ARMA 模型对序列拟合预测，发现采用季节指数能更好地拟合趋势并预测未来序列。二是构建机器

收稿日期：2021-07-06

基金项目：安徽省高等学校自然科学基金项目（KJ2021A0631）

作者简介：赵成兵（1970-），男，教授，硕士生导师，研究方向：多复变函数论、几何分析；
刘丹秀（1996-），女，硕士，研究方向：应用统计。

学习模型。朱晶晶等[5]依据 CMSVM2.0 函数估计和交叉验证等方法，利用月平均气温建立了 SVM 回归预报模型，发现交叉验证下的模型预测效果更好；张曼玉[6]对长三角地区的日温差进行了随机森林拟合，发现影响温度差的主要因子是地表温度；陶晔等[7]利用随机森林筛选出与气温变量高度相关的因子，将这些因子带入长短期记忆网络中，建立了预测性能更佳的 RF–LSTM 模型；王可心等[8]将输入特征进行复合，引入复合特征随机森林回归模型，并用袋外误差率调试参数，发现雨雪天气状况下的路面温度预报精确度最高。三是以各种方式将统计模型与机器学习模型结合起来的混合模型。门晓磊等[9]使用岭回归，随机森林和深度学习三种方法分别对逐日地面 2 m 处的气温进行预报，发现三种方法预测能力相差不大，甚至在小数据集上，随机森林和岭回归可能优于深度学习方法；曾静[10]将输入变量进行多项式扩充，再采用回归方法和随机森林等方法，得出最优拟合温度订正模型，再利用长短期记忆模型建模，最终建立多气象因子模式的温度预报模型；卢维学等[11]提出了基于随机森林算法的偏最小二乘回归模型，通过比较发现该回归模型的稳定性和预测精度优于其他模型。

在已有的研究中，随机森林模型拟合气温时序数据将原始数据直接作为输入特征，或者将输入特征进行组合，作为复合特征引入模型，忽略了气温数据中存在的季节性特征。本文将月份信息分类并采用 One–Hot 编码，提取数据中的季节性，作为随机森林模型的输入特征，构建模型参数组合。在此基础上，利用随机搜索和网格搜索对季节性模型中的超参数进行进一步优化；最后计算拟合误差和准确率[12]，并和乘积季节 ARIMA 模型预测能力进行比较。

1 模型设计

1.1 ARIMA 模型

ARIMA 模型的基本思想是通过变换去除序列的趋势，使非平稳序列变成平稳序列[13]。ARIMA 模型的 AR 部分是根据研究变量自身的历史值进行回归，MA 模型则是出现在不同时间间隔的历史误差值的线性组合。

对于存在季节性的时间序列，季节性可能对建立的模型有影响，因而需要建立季节模型，该模型包括季节影响和非季节影响。季节 ARIMA 模型记为 SARIMA（p,d,q）（P,D,Q），其中，P,D,Q 表示模型季节性部分。本文采用季节性 ARIMA 模型进行建模，主要建模步骤为：首先观察数据时序图，当观测到序列具有趋势或异方差时，则对其进行变换或差分，去除趋势，稳定方差，直到变换后的数据满足平稳性检验的条件，然后根据最小信息量准则和贝叶斯信息准则，拟合预测模型。

1.2 随机森林

随机森林算法是监督学习算法的一个分支，使用集成学习方法回归，集成学习方法主要包括神经网络、SVM 和决策树。随机森林采用 Bagging Bootstrap 技术，通过随机抽样产生更多的样本。在 Bagging 技术中，每个模型都独立运行，且最终输出的是汇总后的模型。但决策树可能会出现过拟合现象，且预测值对训练数据过于依赖和敏感，因而应用随机森林回归作为大决策树的组合，以此代替决策树。随机森林中构建的树并行运行，没有任何交互，其基本思想就是结合多个决策树确定最终结果，而不是依赖单个决策树。

文中随机森林算法的过程可以总结为如下步骤：

（1）对数据进行预处理，提取并编码季节信息；

（2）划分数据集，将其分成训练集和测试集；

（3）对抽样的训练集建立回归树模型，汇总多棵回归树的结果，取其平均作为最终预测结果；

（4）采用两种搜索方法对训练模型进行超参数的优化。

1.3 预测能力评估

1.3.1 平均绝对误差（MAE）

定义如下：

$$MAE=\frac{\sum_{i=1}^{n}\left|f_i-y_i\right|}{n} \tag{1}$$

其中，n 为预测的时间点步长，f_i 为预测值，y_i 为实际观测值。实际观测值与预测值差值越接近，误差越小，说明预测模型的准确性越佳。

1.3.2 平均绝对百分误差（MAPE）

定义如下：

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{f_i-y_i}{y_i}\right| \quad (2)$$

MAPE 越接近于零,表示模型预测的精度越高;若 MAPE 大于 100%,说明预测模型为劣质模型。

2　实证分析

2.1　数据来源

本文中数据来自中国气象数据网,数据包括合肥市某站点观测到的每月平均最高温度(A_MAX_T)、平均最低温度(A_MIN_T)、日照时数(sunshine_duration)、最高温度(MAX_T)、最低温度(MIN_T)、平均温度(AT)、降水量(precipitation)和月份(Month)八个指标,涵盖了 1988 年 1 月至 2020 年 9 月的各月数据,数据无缺失值。

2.2　季节性 ARIMA 模型的应用

2.2.1　差分运算

差分运算是一种提取序列中确定性信息的方法,适当的差分便可以充分提取信息。

季节性 ARIMA 模型是处理时间序列的流行模型之一,它将数据具有的季节性特征考虑到预测中,需要观察自相关系数图(ACF)和偏自相关系数图(PACF)并依据赤池信息准则或贝叶斯信息准则选择合适的模型。因为平均气温序列中含有季节效应,故采用季节模型进行拟合,选取 1988 年 1 月到 2019 年 12 月的平均气温月度数据作为训练集,利用拟合的模型预测 2020 年 1 月至 2020 年 9 月的每月平均气温数据,并和真实值做对比,计算预测准确度,建模过程通过 R 语言实现。

图 1 是平均温度时序图。其中,横轴表示日期,从 1988 年 1 月至 2019 年 12 月;纵轴表示实际观测的月平均气温值,单位为℃。由图可见,随着时间推移,温度值呈现上升后下降的循环,具有很强的周期性,且无明显增加或减少的趋势,序列平稳。

图 2 是气温序列延迟 60 阶的自相关系数图与偏自相关系数图,横轴表示延迟阶数。左侧图中,纵轴表示的是序列的自相关系数;右侧图中,纵轴表示偏自相关系数。由图可知,左图表明 ACF 呈现正负值交替的趋势,且延迟 60 阶后,自相关系数无衰减趋势,表现为拖尾性;右图显示在延迟 12 阶后,PACF 落入两倍标准差范围内,呈现截尾,表明平均气温序列间具有自相关性。

由于序列存在自相关性和季节性,故对原序列作 1 阶 12 步差分,即差分后新序列值为 $\nabla_{12}x_t=x_t-x_{t-12}$,其中 x_t 为序列值。12 步差分后的序列如图 3 所示,温度值均在零值温度线上下波动。为检验序列是否平稳,对差分后的序列进行单位根检验,结果显示 P=0.01,小于显著性水平 0.05,表示不接受原假设,即差分后的平均气温序列中不存在单位根,认为差分后序列基本平稳。

如图 4 所示,虽然延迟 12 阶的自相关系数显著不为 0,偏自相关系数在延迟 12 阶,24 阶显著不为 0,但自相关系数与偏自相关系数基本在 2 倍标准差范围内。

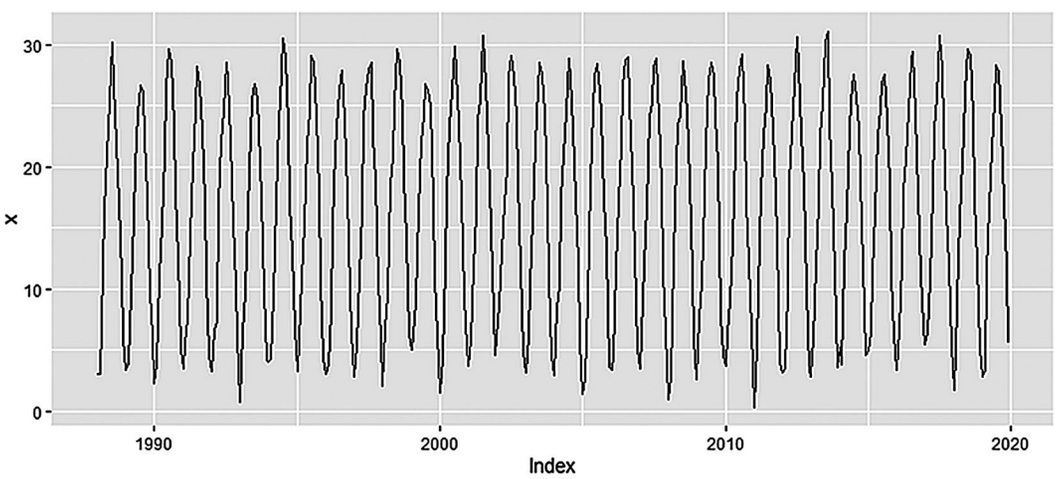

图 1　平均温度时序图

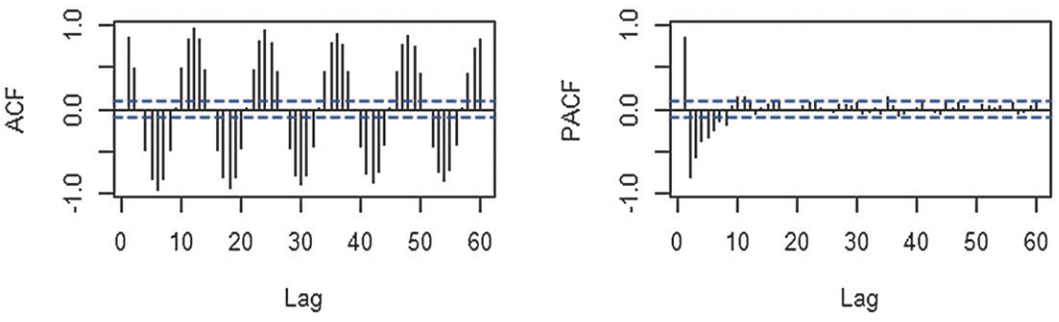

图 2 自相关系数图与偏自相关系数图

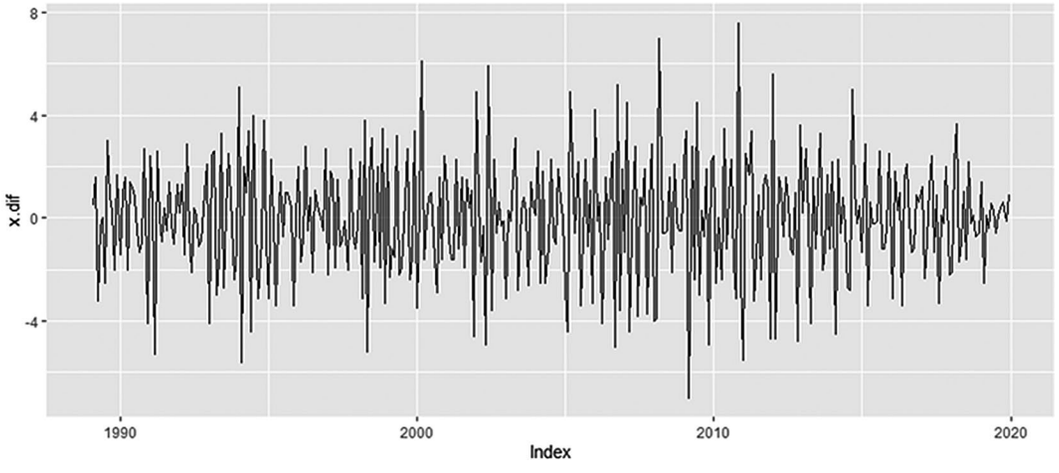

图 3 平均温度 12 步差分后时序图

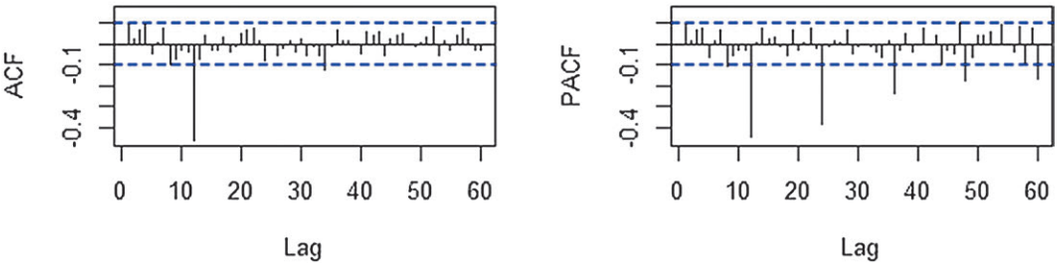

图 4 平均温度差分后自相关系数图与偏自相关系数图

2.2.2 模型建立

根据上图可得，自相关图显示延迟 12 阶自相关系数显著大于 2 倍标准差范围，偏自相关系数图也是如此，说明序列仍然蕴含显著的季节效应，尝试拟合简单 ARMA 模型，但效果并不理想，考虑该序列具有的短期相关性和季节性，尝试用乘积模型拟合序列的趋势。

首先考虑序列 12 阶以内的自相关系数和偏自相关系数均不截尾，尝试使用 ARMA（1,1）模型提取差分后序列的短期自相关信息；其次自相关图显示延迟 12 阶的自相关系数显著非零，但延迟 24 阶自相关系数落入两倍标准差范围内，偏自相关系数显示延迟 24 阶以后的偏自相关系数显著非零。故以 12 步为周期，构建 ARMA$(0,1)_{12}$，经过多次调整之后，依据 AIC，BIC 准则最终确定拟合的模型为 ARIMA$(2,0,1)(0,1,1)_{12}$，此时 BIC 值与 AIC 值达到最小。

2.2.3 模型检验

图 5 是差分后序列的残差自相关检验结果，可以发现，自相关系数呈现逐步衰减趋势，存在小幅度波动，但均在两倍标准差范围内，说明残差序列自相关性弱。同时纯随机性检验表明：12 步差分

后的序列残差在滞后 6 期时,P=0.962 1;当滞后 12 期时,P=0.85;当滞后 24 期时,P=0.782 9,所有的 P 值均大于显著性水平 0.05,表明不拒绝原假设,即差分后的序列的残差独立,模型通过残差白噪声检验,说明拟合的乘积季节性模型 ARIMA(2,0,1)$(0,1,1)_{12}$ 有效。

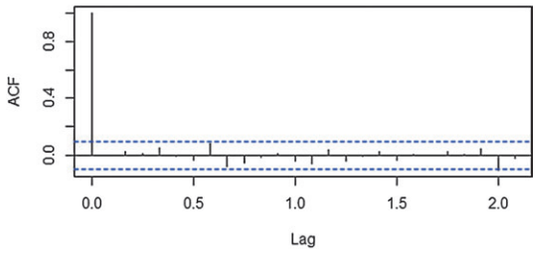

图 5　差分后序列的残差自相关图

2.2.4　模型预测

表 1 中给出了预测值与实际观察值的数据,并计算了预测误差。可以看到仅有个别温度点的预测值与实际值差异较大,最大温度预测误差为 3.97℃,最小预测误差为 0.08℃,经过计算可以得出:

$$MAE=\frac{\sum_{n=1}^{N}\left|E\right|}{N}=1.20 \tag{3}$$

$$MAPE=\frac{100\%}{N}\sum_{n=1}^{N}\left|\frac{E}{y_i}\right|=7.81\% \tag{4}$$

即使用拟合模型预测准确度可以达到 92% 以上。

表 1　预测值与真实值对比

时间	平均温度预测值 f_i/℃	平均温度真实值 y_i/℃	误差 E/℃
2020 年 1 月 1 日	3.31	3.4	−0.09
2020 年 2 月 1 日	5.67	7.4	−1.74
2020 年 3 月 1 日	10.83	11.7	−0.86
2020 年 4 月 1 日	17.10	15.4	1.70
2020 年 5 月 1 日	22.25	22.9	−0.65
2020 年 6 月 1 日	25.74	25.4	0.34
2020 年 7 月 1 日	28.97	25	3.97
2020 年 8 月 1 日	28.05	28.9	−0.85
2020 年 9 月 1 日	23.74	23.1	0.64

2.3　随机森林分析

2.3.1　数据预处理

本文采用 Python 语言编写,基于 Sklearn 环境下构建随机森林模型,样本集中包含 393 个样本,其中 70% 划分为训练集,剩余 30% 作为测试集。在对数据进行初步分析后,发现气象时序数据存在季节性特征,而冬季与夏季的温度差距很大,仅仅考虑月平均气温,精度是不充分的,所以使用文本特征提取方法,将季节性特征也纳入输入特征。春、夏、秋、冬四个分类变量是无序的、离散的,将这些特征数字化时,如果简单分类为 1、2、3、4,分类变量之间便产生了顺序,且不能直接放入机器学习算法中,故而使用 One-Hot 编码。

One-Hot 编码,又称一位有效编码,主要对 M 种状态进行编码,每个状态都有自己独立的寄存器位,并且在任意时候只有一位有效。即每个样本的 M 种属性中只能有一个为 1,表示该样本的该属性属于这个类别,其余扩展属性都为 0。具体编码过程如下:

根据季节特征,将十二月、一月、二月归类为冬季;三、四、五月归类为春季;六、七、八月归为夏季;剩余三个月份归为秋季。

即 S_{q_i}=(春季,夏季,秋季,冬季)=(0,1,0,0),若 i 为夏季,则形式如表 2:

表 2　One-Hot 编码规则

月份 / 月	春季	夏季	秋季	冬季
1	0	0	0	1
3	1	0	0	0
6	0	1	0	0
9	0	0	1	0

如果输入样本 x_i 是夏季,则以 $(x_i,0,1,0,0)$ 的形式采样。

2.3.2　对比实验

随机森林中包含大量的参数,如随机森林决策树的数目、树的最大深度,本文的数据量并不大,故将最大深度设置为 None。节点最小分裂所需样本个数是某节点样本数的最小值,当节点样本数小于该值时,不会将其划分。叶子节点最小样本数代表的是叶子节点最少的样本数目,若小于该值,则该节点会被剪枝。为了验证季节性特征是否利于提高随机森林模型预测精度,本文用简单随机森林模型和季节性随机森林模型进行比较,从平均绝对误差、均方误差和准确度三方面衡量预测能力。

如表 3 显示，在决策树的个数 M 均为 20 的前提下，简单随机森林模型的平均绝对误差为 0.28，平均绝对百分误差是 4.04%；而季节性模型所得出的平均绝对误差是 0.26，平均绝对百分误差为 3.86%，说明加入季节性特征之后，减小了模型误差，提高了预报准确度。与两种随机森林算法相比，乘积季节性 ARIMA 模型预报气温误差更大，而且需要通过 ACF 图和 PACF 图主观确定模型的阶数，预测气温的准确率相对较低。

表 3　各模型在月平均气温的预报性能对比

算法	平均绝对误差	平均绝对百分比误差	均方误差	准确度
乘积 ARIMA	1.20	7.81%	2.677 9	92.19%
简单随机森林	0.28	4.04%	0.141 5	95.96%
季节性随机森林	0.26	3.86%	0.1247	96.14%

下图是测试集包含的 118 个样本的真实标签值与预测值折线图，其中红色线表示预测值，绿色线表示标签值。可以看到，使用季节性随机森林模型进行预测，虽然有部分极值点的温度预测值与标签值存在偏差，但整体趋势一致，且准确度可以达到 96% 以上，总体预测效果较好。

2.3.3　参数优化

（1）网格搜索

超参数搜索算法一般包括目标函数、搜索范围等要素。网格搜索通过搜索上下限内的所有点确定最优值，因而极有可能找到全局最优值，但局限性在于计算量较大、耗时耗力，特别是需要调优的超参数较多时。一般先使用较广的搜索范围和较大的步长，寻找全局最优值可能的位置，然后逐渐缩小搜索范围和步长，寻找更精确的最优值。

（2）随机搜索

与网格搜索相比，随机搜索在上下限内随机选取样本点，搜索时间相对缩短，但产生的结果不一定是全局最优。当样本点集足够大时，随机采样也能找到全局最优值或其近似值。

在树的初始数目设为 20 时，季节性随机森林模型预测精度可以达到 96.14%，在此模型上进行超参数优化，并且使用三折交叉验证将数据集划分训练集和测试集，即将原始数据集进行三次划分，多次训练，取三次输出结果的均值作为算法精度的估计值，避免只将数据集一次划分而得出错误结论的情况。然后在季节性随机森林产生的最优参数

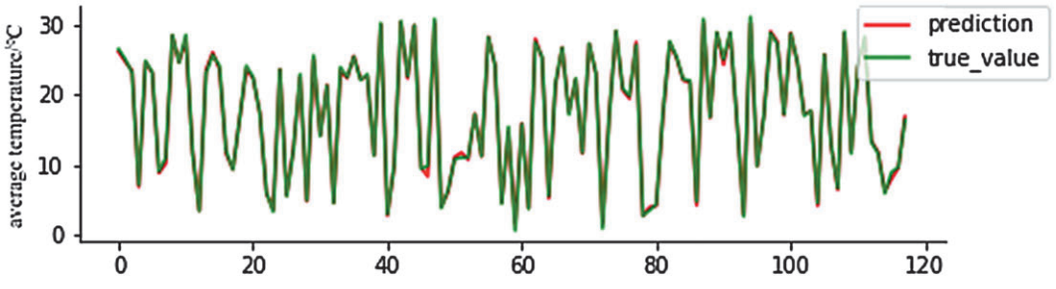

图 6　真实值与预测值对比图

表 4　随机搜索与网格搜索参数空间

	随机搜索参数空间	随机搜索最佳参数空间	网格搜索参数空间	网格搜索最佳参数组合
树的个数	20/40/…/480/500	200	50/100/150	150
叶子节点最小样本数	1/2/4	1	1/2/3	1
节点最小分裂所需样本个数	2/5/10	2	1/2/3	3
最大特征的选择方式	auto	auto	auto	auto
树的最大深度	10/20/None	10	None	None
样本采样方式（Bootstrap）	True/False	True	True	True
平均绝对误差	0.236 4		0.241 7	
准确度	96.42%		96.34%	

空间基础上进行随机搜索。

从表 4 中可以看到,在 Bootstrap 方法下,随机搜索最佳模型在树的个数为 200、节点最小分裂所需样本数为 2 时得到,此时准确率已经达到 96.42%,优于树的个数 N 为 20 时的季节性随机森林模型。继续根据随机搜索产生的最佳参数空间,分别向最佳组合的左、右进行网格搜索,搜索的参数空间设定为 N 取 50,100 或 150 时,节点最小分裂所需样本数取 1,2 或 3,同样使用 Bootstrap 采样。结果显示,在树的个数为 150 时,搜索到最佳组合,准确度为 96.34%。结果未寻找到更优的参数组合,继续向右搜索。

表 5 结果表明,在三折交叉验证下,当树的个数为 100、节点最小分裂所需样本个数为 2、叶子节点最小样本数也为 2 时取得最优,此时预测准确率是 96.45%,准确度进一步提升。

表 5　向右网格搜索参数空间

	网格搜索的参数空间	网格搜索的最佳参数组合
树的个数	50/100/200/300/400	100
叶子节点最小样本数	2/3/4/5	2
节点最小分裂所需样本个数	2/3/4/5	2
最大特征的选择方式	auto	auto
树的最大深度	None	None
样本采样方式(Bootstrap)	True	True
平均绝对误差	0.253 0	
准确度	96.45%	

3　结论

气候对人类的生活会产生巨大的影响。本文使用了基于 R 语言的季节性 ARIMA 模型和基于 Python 语言的季节性随机森林两种模型对气象时间序列数据进行分析与建模,并对未来时刻进行了预测,得到如下结论:

(1)季节性 ARIMA 模型可以很好地拟合时序数据中的季节性,预测精度超 92%。虽然夏季高温极端值预报偏高,但偏差绝对值基本在 3℃以内,认为预测效果有效。

(2)文中建立的随机森林模型引入季节特征作为输入特征时,模型对于温度极值的预测值偏小,整体拟合趋势符合实际趋势,且预测效果优于季节 ARIMA 模型。

(3)在季节性随机森林模型基础上,利用随机搜索找出优化组合,然后根据该参数空间,在该组合附近进一步使用网格搜索,搜索该区域内所有可能值确定最优参数组合,此时模型的预测精度最高。

引入季节性特征的随机森林模型可用于气温预测,且预测误差较小。但由于资料限制,实验中数据仅是单个气象站的数据,输入变量较少,未能考虑到将周围地理气象数据可能存在的影响,这也是下一步研究的方向。

参考文献:

[1] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model[J].Neurocomputing,2003,50:159-175.

[2] 马廷淮,穆强,田伟,等 . 气象数据挖掘研究 [J]. 武汉理工大学学报,2010,32(16):110-114.

[3] Dimri T,Ahmad S,Sharif M. Time series analysis of climate variables using seasonal ARIMA approach[J]. Journal of Earth System Science,2020,129(1):1-16.

[4] 谭小花 . 时间序列分析方法在重庆气温研究中的运用 [D]. 重庆:重庆大学,2016.

[5] 朱晶晶,赵小平,吴胜安,等 . 基于支持向量机的海南气温预测模型研究 [J]. 海南大学学报(自然科学版),2016,34(1):40-44.

[6] 张曼玉 . 长三角夏季地表热环境的昼夜和季节内变化及其对气温日较差的影响研究 [D]. 南京:南京信息工程大学,2020.

[7] 陶晔,杜景林 . 基于随机森林的长短期记忆网络气温预测 [J]. 计算机工程与设计,2019,40(3):737-743.

[8] 王可心,包云轩,朱承瑛,等 . 随机森林回归法在冬季路面温度预报中的应用 [J]. 气象,2021,47(1):82-93.

[9] 门晓磊,焦瑞莉,王鼎,等 . 基于机器学习的华北气温多模式集合预报的订正方法 [J]. 气候与环境研究,2019,24(1):116-124.

[10] 曾静 . 基于机器学习的多气象因子模式预报温度订正模型 [D]. 金华:浙江师范大学,2020.

[11] 卢维学,吴和成,万里洋 . 基于融合随机森林算法的 PLS 对降水量的预测 [J]. 统计与决策,2020,36(18):27-31.

[12] 徐映梅,陈尧 . 季节 ARIMA 模型与 LSTM 神经网络预测的比较 [J]. 统计与决策,2021,37(2):46-50.

[13] 易丹辉 . 数据分析与 EViews 应用 [M]. 北京:中国人民大学出版社,2008.